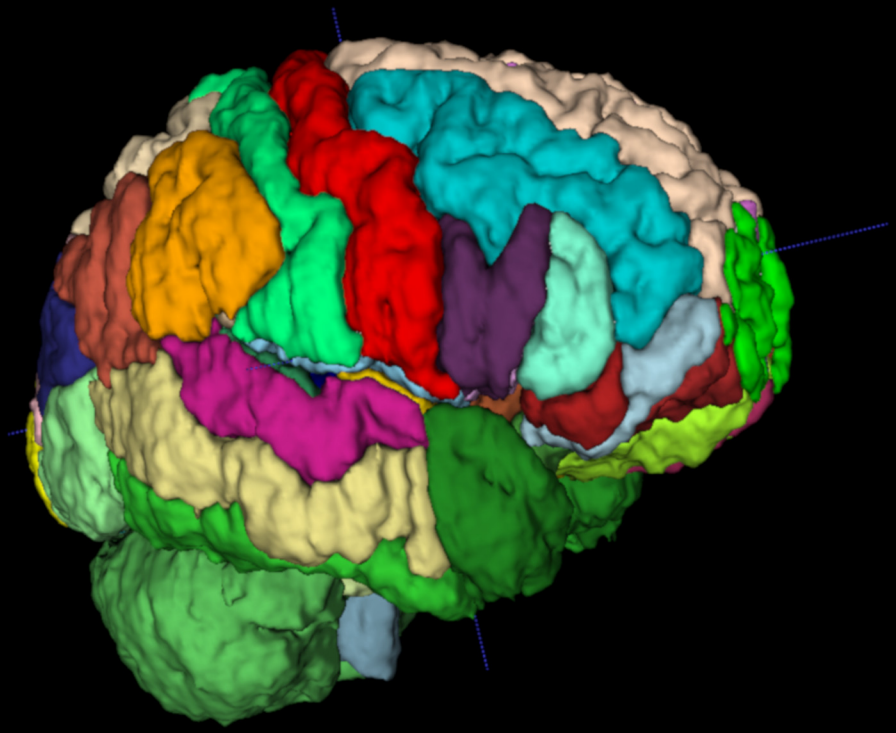
AUTOMATED BRAIN STRUCTURE SEGMENTATION IN MAGNETIC RESONANCE IMAGES OF MULTIPLE SCLEROSIS PATIENTS

Sandra González-Villà

Universitat de Girona



PhD Thesis:

Automated brain structure segmentation in magnetic resonance images of multiple sclerosis patients

Sandra González-Villà

2019

DOCTORAL THESIS

Automated brain structure segmentation in magnetic resonance images of multiple sclerosis patients

Sandra González-Villà

2019

DOCTORAL PROGRAM in TECHNOLOGY

Supervised by:
Dr. Xavier Lladó
Dr. Arnau Oliver

Presented to obtain the degree of Doctor of Philosophy at the University of Girona

A mis padres y a Joan

Acknowledgments

Aquesta tesi no hagués estat possible sense l'ajuda i suport de moltes persones. En primer lloc, voldria començar donant les gràcies al meus directors de tesi, en Xavi Lladó i l'Arnau Oliver, que han fet d'aquesta etapa que és el doctorat una molt bona experiència que recordaré de forma especial. Gràcies per ensenyar-me, per la vostra dedicació, entusiasme i inesgotable bon humor.

També voldria donar les gràcies a tots els companys que han passat pel laboratori, i a la resta de membres de VICOROB, en especial als meus companys del grup d'imatge mèdica: Kaisar, Jose, Mostafa, Albert, Robert, Óliver, Richa i Eloy G. per les xerrades inspiradores i discussions constructives d'aquests últims anys. I molt especialment, a en Sergi i l'Eloy R., per la seva paciència i predisposició per ajudar-me, sobretot quan vaig aterrar en el món de la neuro-imatge.

Moltes gràcies també a tots els amics del P-IV: Ricard, Pablo, Ferran T., Mariano, Albert, Xesca, Pepe, Quim i Ferran R., que de seguida em van acollir, fent molt més entretingut el dia a dia i amb qui he compartit totes les penes i alegries d'aquest doctorat. Però també, a tots els nous amics que he fet durant aquest camí: Aurora, Sònia, Aïda, David, Mireia, Sandra, Marc, Gerard i Kara, que han sigut els meus companys d'oci i esbarjo i que han fet de Girona la meva casa.

Òbviament, gràcies a les institucions que han fet que aquesta recerca hagi sigut possible. En especial, moltes gràcies a "La Fundació la Marató de TV3" per subvencionar el meu doctorat i a les beques de mobilitat MOB17 de la Universitat de Girona per fer possible la meva estada de recerca.

I would also like to thank Dr. Bennett Landman for hosting and teaching me during my research stay at Vanderbilt University. But also, thanks to all my colleagues and friends from the VISE institute, who made from my stay in Nashville an awesome experience.

Finalmente, quiero dar las gracias a toda mi familia por su apoyo y por los infinitos buenos momentos de todos estos años. Gràcies especialment a la meva

avia que, sempre orgullosa, ha estat amb mi en tots els moments importants. Pero sobretodo, muchísimas gracias a mis padres y a mi hermano Joan, por su soporte incondicional y su ilimitada confianza, sin cuyo respaldo y ayuda, nunca habría llegado hasta aquí.

I per últim, però no per això menys important, moltes gràcies a en Marc per fer cada dia més fàcil. Gràcies per fer que aquests anys hagin sigut tan divertits, per la teva positivitat, la teva paciència i el teu sentit de l'humor, però per damunt de tot, gràcies per compartir-ho amb mi.

Publications

Journals

- **Sandra González-Villà**, Arnau Oliver, Yuankai Huo, Xavier Lladó, Bennett A. Landman. A fully automated pipeline for brain structure segmentation in multiple sclerosis. *Submitted to NeuroImage: Clinical*, 2019. Quality index: [JCR N IF:3.869, Q1(3/14)].

- Kaisar Kushibar, Sergi Valverde, **Sandra González-Villà**, Jose Bernal, Mariano Cabezas, Arnau Oliver, Xavier Lladó. Domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Scientific Reports*, To appear, 2019. Quality index: [JCR MS IF 4.122, Q1(12/64)].

- **Sandra González-Villà**, Arnau Oliver, Yuankai Huo, Xavier Lladó, Bennett A. Landman. Brain structure segmentation in the presence of multiple sclerosis lesions. *NeuroImage: Clinical*, vol 22, pp. 101709, 2019. Quality index: [JCR N IF:3.869, Q1(3/14)].

- Kaisar Kushibar, Sergi Valverde, **Sandra González-Villà**, Jose Bernal, Mariano Cabezas, Arnau Oliver, Xavier Lladó. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Medical Image Analysis*, vol 48, pp. 177-186, 2018. Quality index: [JCR CSAI IF 5.356, Q1(6/105)].

- **Sandra González-Villà** , Sergi Valverde, Mariano Cabezas, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, Xavier Lladó. Evaluating the effect of multiple sclerosis lesions in automatic brain structure segmentation. *NeuroImage: Clinical*, vol 15, pp. 228-238, 2017. Quality index: [JCR N IF:3.869, Q1(3/14)].

- Sergi Valverde, Mariano Cabezas, Eloy Roura, **Sandra González-Villà**, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira , Arnau Oliver, Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, vol 155, pp. 159-168, 2017. Quality index: [JCR N IF:5.426, Q1(1/14)].

- Sergi Valverde, Arnau Oliver, Eloy Roura, **Sandra González-Villà**, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, and Xavier Lladó. Automated tissue segmentation of MR brain images in the presence of white matter lesions. *Medical Image Analysis*, vol 35, pp. 446-457, 2017. Quality index: [JCR CSAI IF:5.356, Q1(6/105)].

- **Sandra González-Villà** , Arnau Oliver, Sergi Valverde, Liping Wang, Reyer Zwiggelaar, and Xavier Lladó. A review on brain structures segmentation in magnetic resonance imaging. *Artificial Intelligence in Medicine*, vol 73, pp. 45-69, 2016. Quality index: [JCR CSAI IF:2.142 Q2(34/142)].

- Eloy Roura, Arnau Oliver, Sergi Valverde, **Sandra González-Villà** , Ricard Cervera, Nuria Bargalló, and Xavier Lladó. Automated detection of lupus white matter lesions in MRI images. *Frontiers in Neuroinformatics*, 10, art 33, 2016. Quality index: [JCR MCB IF:3.047 Q1(6/56)].

Conferences

- **Sandra González-Villà**, Sergi Valverde, Mariano Cabezas, Yuankai Huo, Arnau Oliver, Lluís Ramió-Torrentà, Bennett A. Landman, Xavier Lladó. An end to end automated pipeline for brain structure segmentation in multiple sclerosis patients. *Submitted to ECTRIMS 2019. Multiple Sclerosis.* Quality index: [JCR CN IF:5.280 Q1(22/197)]

- **Sandra González-Villà**, Yuankai Huo, Arnau Oliver, Xavier Lladó, Bennett A. Landman. Multi-atlas Parcellation in the Presence of Lesions: Application to Multiple Sclerosis. *4th International Workshop on Patch-based Techniques in Medical Imaging.* MICCAI 2018. 20 September 2018, Granada, Spain.

- Jose Bernal, Mostafa Salem, Kaisar Kushibar, Albert Clèrigues, Sergi Valverde, Mariano Cabezas, **Sandra González-Villà**, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. MR Brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors. *MR Brain tissue segmentation Challenge in Medical Imaging.* MICCAI Workshop, 2018. 16 September 2018, Granada, Spain.

- Mariano Cabezas, Sergi Valverde, **Sandra González-Villà**, Albert Clèrigues, Mostafa Salem, Kaisar Kushibar, Jose Bernal, Arnau Oliver, Joaquim Salvi and Xavier Lladó. Survival prediction using ensemble tumor segmentation and transfer learning. *Multimodal Brain Tumor Segmentation Challenge 2018 (BRATS) in Medical Imaging.* MICCAI Workshop, 2018. 16 September 2018, Granada, Spain.

- **Sandra González-Villà**, Sergi Valverde, Mariano Cabezas, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, Xavier Lladó. Do multiple sclerosis lesions affect automatic brain structure segmentation?. *ECTRIMS 2017. Multiple Sclerosis.* October 2017, Paris, France. Quality index: [JCR CN IF:5.280 Q1(22/197)]

- Sergi Valverde, Mariano Cabezas, Eloy Roura, **Sandra González-Villà**, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, Xavier Lladó. A deep learning approach for multiple sclerosis lesion segmentation. *ECTRIMS 2017. Multiple Sclerosis.* October 2017, Paris, France. Quality index: [JCR CN IF:5.280 Q1(22/197)]

- Jose Bernal, Kaisar Kushibar, Sergi Valverde, Mariano Cabezas, **Sandra González-Villà**, Mostafa Salem, Joaquim Salvi, Arnau Oliver, Xavier Lladó. Six-month infant brain tissue segmentation using three dimensional fully convolutional neural networks and pseudo-labelling. *MICCAI Grand Challenge on 6-month infant brain MRI segmentation iSeg-2017.* MICCAI 2017. 14 September 2017, Quebec, Canada.

- Sergi Valverde, Mariano Cabezas, Jose Bernal, Kaisar Kushibar, **Sandra González-Villà**, Mostafa Salem, Joaquim Salvi, Arnau Oliver, Xavier Lladó. White matter hyperintensities segmentation using a cascade of three convolutional neural networks. *MICCAI Grand Challenge on White Matter Hyperintensties Segmentation.* MICCAI 2017. 14 September 2017, Quebec, Canada.

- Eloy Roura, Mariano Cabezas, Sergi Valverde, **Sandra González-Villà**, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. Unsupervised Multiple Sclerosis Lesion Detection and Segmentation using Rules and Level Sets. *Multiple Sclerosis Lesions segmentation (MSSEG) Challenge.* MICCAI 2016. 21 October 2016, Athens, Greece.

- Sergi Valverde, Mariano Cabezas, Eloy Roura, **Sandra González-Villà**, Joaquim Salvi, A. Oliver, and Xavier Lladó. Multiple Sclerosis Lesion Detection and Segmentation using a Convolutional Neural Network of 3D Patches. *Multiple Sclerosis Lesions segmentation (MSSEG) Challenge.* MICCAI 2016. 21 October 2016, Athens, Greece.

- Sergi Valverde, Arnau Oliver, **Sandra González-Villà**, Eloy Roura, Deborah Pareto, Joan Carles Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Xavier Lladó. Automated tissue segmentation of magnetic resonance images of multiple sclerosis patients. *ECTRIMS 2016. Multiple Sclerosis*. September 2016, London, United Kingdom. Quality index: [JCR CN IF:4.840 Q1(27/192)]

Acronyms

AAM Active Appearance Model
ACM Active Contour Model
ANN Artificial Neural Network
ASM Active Shape Model
BAM Bayesian Appearance Model
CIS Clinically Isolated Syndrome
CNN Convolutional Neural Network
CNS Central Nervous System
CSF Cerebrospinal Fluid
DIR Double Inversion Recovery
DSC Dice Similarity Coefficient
EM Expectation Maximization
FCNN Fully Convolutional Network
FLAIR Fluid Attenuated Inversion Recovery
GAs Genetic Algorithms
GM Gray Matter
GMIs Geometric Moment Invariants
HD Hausdorff Distance
JLF Joint Label Fusion
MAP Maximum A Posteriori
m-JLF Masked Joint Label Fusion
MLP Multi-Layer Perceptron
m-NLSS Masked Non-local Spatial STAPLE
MoG Mixture of Gaussians
MP-RAGE Magnetization-Prepared Rapid Acquisition with Gradient Echo
MRI Magnetic Resonance Image
MRF Markov Random Field
MS Multiple Sclerosis
N3 Non-parametric, Non-uniform intensity Normalization

NLSS Non-local Spatial STAPLE
PCA Principal Component Analysis
PDM Point Distribution Model
PD-w Proton Density-weighted
PPMS Primary Progressive Multiple Sclerosis
PRMS Progressive Relapsing Multiple Sclerosis
PSIR Phase-Sensitive Inversion Recovery
RIS Radiologically Isolated Syndrome
RRMS Relapsing Remitting Multiple Sclerosis
SLF SALEM Lesion Filling
SLS SALEM Lesion Segmentation
SPMS Secondary Progressive Multiple Sclerosis
STAPLE Simultaneous Truth and Performance Level Estimation
SVM Support Vector Machines
T1-w T1-weighted
T2-w T2-weighted
WM White Matter

List of Figures

List of Tables

Contents

Abstract

This thesis is focused on the automated segmentation of the brain structures in magnetic resonance images, applied to multiple sclerosis (MS) patients. This disease is characterized by the presence of demyelinating lesions in the brain, that appear as focal low signal intensity areas in the T1-weighted sequence, which is the most frequently used modality to segment the brain structures. In the first place, we exhaustively analyze the state of the art on this topic, presenting a new classification of the methods based on their segmentation strategy. We further discuss each category's strengths and weaknesses and analyze its performance in segmenting different brain structures, providing a qualitative and quantitative comparison. From this first analysis, we observe that the vast majority of the reviewed methods are not designed for brains with lesions, such as those encountered in MS patients. Consequently, we also perform a thorough analysis of the effect of MS lesions on three representative state-of-the-art methods, each relying on a different category of the classification: FreeSurfer, FIRST and majority-vote label fusion. This second analysis reveals that the three segmentation approaches are indeed affected by the presence of these lesions, demonstrating that there exists a problem when using automatic methods as a tool to measure the disease progression. Therefore, based on the conclusions of these two studies, we propose a new correspondence search model able to minimize this problem on intensity-based multi-atlas label fusion strategies. Afterwards, we extend the theory of two remarkable label fusion strategies of the literature, i.e. Non-local Spatial STAPLE and Joint Label Fusion, in order to integrate our model into their corresponding estimation algorithms. Furthermore, with the aim of providing fully automated brain structure segmentation algorithms, a whole automated pipeline including lesion segmentation, pre-processing, atlas selection, masked registration and label fusion, is presented. Finally, a second extension of the theory to enable the integration of manual and automatic edits into the segmentation estimation of both strategies is also proposed. The evaluation, carried out in a quantitative and qualitative manner, includes a comparison of the proposed approaches to the original strategies when segmenting the raw images and the lesion-

filled images, using both manual and automatically segmented lesion masks. The analysis of the results obtained with the proposed strategies points out a performance improvement on the lesion areas, which is also reflected on the whole brain segmentation performance.

Resum

Aquesta tesi se centra en la segmentació automàtica de les estructures cerebrals en imatges de ressonància magnètica, aplicada a pacients d'esclerosi múltiple (EM). Aquesta malaltia es caracteritza per la presència de lesions desmielinitzants al cervell, que apareixen com àrees focals de baixa intensitat de senyal en la seqüència T1-w, que és la modalitat més utilitzada per segmentar les estructures cerebrals. En primer lloc, analitzem exhaustivament l'estat de l'art en aquest tema, presentant una nova classificació dels mètodes basada en la seva estratègia de segmentació. A més, estudiem les fortaleses i inconvenients de cada categoria i analitzem el seu rendiment en la segmentació de diferents estructures, proporcionant una comparació qualitativa i quantitativa. En aquesta primera anàlisi, observem que la gran majoria dels mètodes revisats no estan dissenyats per a cervells amb lesions, com les que apareixen en pacients d'EM. Conseqüentment, també realitzem una anàlisi exhaustiva de l'efecte de les lesions d'EM en tres mètodes representatius de l'estat de l'art, cadascun d'ells basat en una categoria diferent de la classificació proposada: FreeSurfer, FIRST i fusió d'etiquetes mitjançant majoria de vot. Aquesta segona anàlisi, revela que els tres enfocaments de segmentació es veuen afectats per la presència d'aquestes lesions, el que demostra que hi ha un problema en els mètodes de segmentació automàtica quan s'utilitzen com a eina per mesurar la progressió de la malaltia. Per tant, en base a les conclusions d'aquests dos estudis, proposem un nou model de cerca de correspondències capaç de minimitzar aquest problema en les estratègies de fusió d'etiquetes de múltiples atles basades en intensitat. Posteriorment, estenem la teoria de dues estratègies de fusió d'etiquetes notables de la literatura, Non-local Spatial STAPLE i Joint Label Fusion, per integrar el nostre model en els seus corresponents algoritmes d'estimació. Addicionalment, amb l'objectiu de proporcionar algoritmes de segmentació d'estructures cerebrals totalment automatitzats, es presenta una línia automàtica completa que inclou la segmentació de lesions, el preprocessat, la selecció d'atles, el registre emmascarat i la fusió d'etiquetes. Finalment, també es proposa una segona extensió de la teoria per permetre la integració d'anotacions manuals i automàtiques en l'estimació de segmentació de les dues es-

tratègies. L'avaluació, realitzada de manera quantitativa i qualitativa, inclou una comparació dels enfocaments proposats amb les estratègies originals al segmentar les imatges sense processar i les imatges amb "lesion filling", utilitzant màscares de lesions tant manuals com segmentades automàticament. L'anàlisi dels resultats obtinguts amb les estratègies proposades demostra una millora en el rendiment dels algorismes de segmentació en les àrees de lesió, que també es veu reflectida en la segmentació de tot el cervell.

Resumen

Esta tesis se centra en la segmentación automática de las estructuras cerebrales en imágenes de resonancia magnética, aplicada a pacientes de esclerosis múltiple (EM). Esta enfermedad se caracteriza por la presencia de lesiones desmielinizantes en el cerebro, que aparecen como áreas focales de baja intensidad de señal en la secuencia T1-w, que es la modalidad más utilizada para segmentar las estructuras cerebrales. En primer lugar, analizamos exhaustivamente el estado del arte en este tema, presentando una nueva clasificación de los métodos basada en su estrategia de segmentación. Además, estudiamos las fortalezas y debilidades de cada categoría y analizamos su rendimiento en la segmentación de diferentes estructuras, proporcionando una comparación cualitativa y cuantitativa. En este primer análisis, observamos que la gran mayoría de los métodos revisados no están diseñados para cerebros con lesiones, como las que encontramos en pacientes de EM. Consecuentemente, también realizamos un análisis exhaustivo del efecto de las lesiones de EM en tres métodos representativos del estado del arte, cada uno basado en una categoría diferente de la clasificación propuesta: FreeSurfer, FIRST y fusión de etiquetas mediante mayoría de voto. Este segundo análisis, revela que los tres enfoques de segmentación se ven afectados por la presencia de estas lesiones, lo que demuestra que existe un problema en los métodos de segmentación automática cuando se utilizan como herramienta para medir la progresión de la enfermedad. Por lo tanto, en base a las conclusiones de estos dos estudios, proponemos un nuevo modelo de búsqueda de correspondencias capaz de minimizar este problema en las estrategias de fusión de etiquetas de múltiples atlas basadas en intensidad. Posteriormente, extendemos la teoría de dos estrategias de fusión de etiquetas notables de la literatura, Non-local Spatial STAPLE y Joint Label Fusion, para integrar nuestro modelo en sus correspondientes algoritmos de estimación. Adicionalmente, con el objetivo de proporcionar algoritmos de segmentación de estructuras cerebrales totalmente automatizados, se presenta una línea automática completa que incluye segmentación de lesiones, preprocesado, selección de atlas, registro enmascarado y fusión de etiquetas. Finalmente, también se propone una segunda extensión de la

teoría para permitir la integración de anotaciones manuales y automáticas en la estimación de segmentación de ambas estrategias. La evaluación, realizada de manera cuantitativa y cualitativa, incluye una comparación de los enfoques propuestos com las estrategias originales al segmentar las imágenes sin procesar y las imágenes con "lesion filling", utilizando máscaras de lesiones tanto manuales como segmentadas automáticamente. El análisis de los resultados obtenidos con las estrategias propuestas señala una mejora en el rendimiento en las áreas de lesión, que también se refleja en el rendimiento de la segmentación de todo el cerebro.

Chapter 1

Introduction

1.1 Multiple Sclerosis

1.1.1 What is multiple sclerosis?

Multiple sclerosis (MS) is a chronic immune-mediated demyelinating disease of the central nervous system (CNS). Most MS experts believe it to be an autoimmune disease directed against CNS myelin (fatty substance that surrounds and insulates the nerve cell axons) or oligodendrocytes (cells responsible for creating and maintaining the myelin sheath) [1]. Although the disease does not fulfill all the criteria for a definite autoimmune disease, a large variety of immune abnormalities have been seen in MS, in addition to several epidemiological observations that also support this theory. Other experts believe it to be a neurodegenerative disease, based on the observations that axonal loss and neurodegeneration responsible for the irreversible disability occur early in the course of MS and, as the disease evolves predominate the underlying pathogenetic mechanisms [2, 3].

MS is characterized by the formation of lesions in the CNS (also called plaques), inflammation, and the destruction of myelin sheaths of neurons. More specifically, the immune system causes inflammation that damages myelin − which helps the neurons carry electrical signals − as well as the cell/neuron axons themselves and the oligodendrocytes. When the myelin is lost, a neuron can no longer effectively conduct electrical signals, which is particularly relevant because it is the main underlying mechanism of permanent clinical disability. In early phases of the disease, a repair process, called remyelination, occurs. However, the oligodendrocytes are unable to completely rebuild the cell's myelin sheath, which after repeated attacks ends in

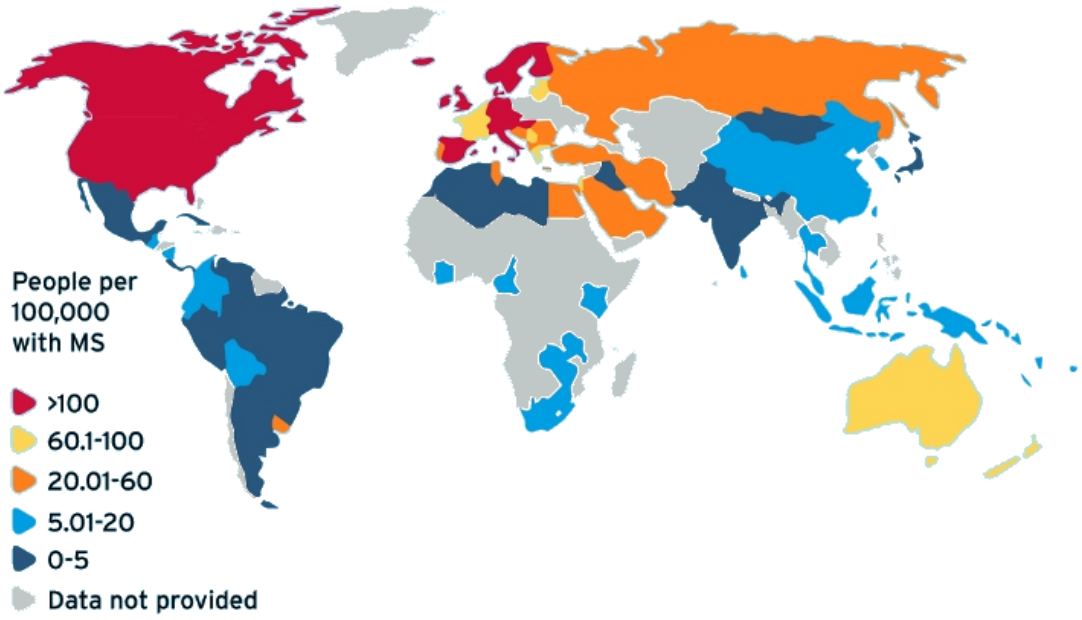


Figure 1.1: Prevalence of multiple sclerosis patients per country. Data from 2013.

scar-like plaques built up around the damaged axons.

1.1.2 Demographics: geography, genetics and environment

Although the course of the disease is highly variable, many people develop irreversible disability, being MS the most common inflammatory disorder of the CNS and a leading cause of neurological disability in young adults. With a prevalence of $50 - 300$ per $100,000$ [4], the total number of people living with MS worldwide is estimated to be $2 - 2.5$ million [1], however, this is likely to be an under-estimate given the relative lack of data from large populations such as India and China.

The disease global distribution generally increases with increasing distance from the equator, although there are exceptions. As can be observed from Figure 1.1, its prevalence varies between < 5 cases per $100,000$ people in tropical areas or Asia and $> 100 - 200$ cases per $100,000$ in temperate areas, especially those with large populations of Northern European origin, including the United States, Canada, New Zealand and parts of Australia [1].

MS typically presents between the ages of 20 years and 50 years, however, approximately 0.5% of adults with this disease have symptom onset at the age of 60 years or older [5], and up to 5% of patients with multiple sclerosis develop their first symptoms in childhood [6]. Although the causes of MS remain unknown, envi-

ronmental risk factors such as vitamin D deficiency, diet, obesity in early life, and cigarette smoking are known to play a part in the development of the disease [4].

The median time to death is around 30 years from disease onset, representing a reduction in life expectancy of $5-10$ years compared to unaffected people [7]. Additionally, similar to many autoimmune disorders, the disease is more common in women. The sex ratio has increased markedly during the last decades ($2.3-3.5:1$) due to an increased incidence of MS among women but not men [8].

1.1.3 MS phenotypes and clinical course

MS takes several forms, with new symptoms either occurring in isolated attacks (relapsing forms) or building up over time (progressive forms). The initial presentation of the disease varies according to both the location of the lesions and the type of symptom onset (relapsing or progressive).

The vast majority of patients who develop MS begin with a single episode, called *Clinically Isolated Syndrome* (CIS), that involves the optic nerve, brainstem, or spinal cord, and resolves over time. If accompanied with white matter (WM) abnormalities detected by magnetic resonance images (MRI) at clinically unaffected sites, the chance of a second attack of demyelination increases, which marks the onset of clinically definite MS.

Patients that have at least two relapses are described as having *Relapsing Remitting MS* (RRMS). RRMS is the most common presentation of the disease, affecting around $85-90\%$ of patients with MS. It is characterized by acute attacks of new or recurrent neurological signs and symptoms, followed by complete or partial recovery, and separated by variable periods of stable neurological condition without clinical disease activity. RRMS typically affects young adults, women being three times more affected than men.

Another possible presentation of the disease is a progressive form from onset, in which case it is termed *Primary Progressive MS* (PPMS). The PPMS course affects around $10-15\%$ of patients and is characterized by a steady accumulation of neurological disability over time, usually without relapses, from disease onset. This presentation of the disease usually occurs later in life than does RRMS and there is no gender bias. Some of these patients may experience superimposed relapses, called *Progressive-Relapsing MS* (PRMS) [1].

With time, some patients with RRMS, may develop a progressive course of the disease, termed *Secondary Progressive MS* (SPMS). SPMS presents a gradual increase in neurological disability, that accumulates progressively between or without further relapses. However, SPMS has been shown to be more similar to PPMS, the differences between them being relative rather than absolute [4].

Finally, patients with incidental MRI findings consistent with MS are classified as suffering from *Radiologically Isolated Syndrome* (RIS). A third of patients with RIS will develop clinical symptoms of MS within 5 years of follow-up, either a relapse or progressive symptoms [6].

When the disease presents in childhood it is almost always a RRMS course. On the other hand, people presenting with multiple sclerosis after the age of 50 years are more likely to have a progressive onset. In this age group, men are over-represented.

1.1.4 Diagnosis

There is no single pathognomonic clinical feature or diagnostic test that can provide a definite diagnosis of MS, and for this reason, the integration of clinical, imaging, and laboratory findings are necessary in practice. However, the rates of misdiagnosis are still as high as 10%.

The diagnosis of MS requires objective evidence of CNS lesions disseminated in time and space. Furthermore, it is also important that there is no better explanation for the clinical presentation, and that alternative diagnoses have been considered and excluded. When neurological symptoms and signs suspected of indicating MS appear, MRI can assist in excluding other conditions that mimic MS (red flags), as well as confirming a definite diagnosis of the disease.

In 2001, a panel of experts on MS came up with a set of diagnostic criteria that included magnetic resonance images for the first time to provide evidence of CNS lesions disseminated in time and space [9]. The so called McDonald criteria have undergone several revisions in recent years [5, 10, 11], with increased certainty with successive versions, and have become the gold standard test for MS diagnosis. The use of the McDonald criteria enables an earlier and more reliable diagnosis of MS than with clinical findings alone, facilitating earlier management, such as initiation of treatment or observation, which is highly important for the patient condition.

The last revision, i.e. 2017, defines lesion dissemination as follows:

- *In space*: this can be demonstrated by one or more lesions that are characteristic of MS in two or more of four areas of the CNS (peri-ventricular, cortical or juxta-cortical, and infra-tentorial brain regions, and the spinal cord).
- *In time*: demonstrated by the simultaneous presence of gadolinium-enhancing and non-enhancing lesions at any time or by a new lesion on follow-up MRI, with reference to a baseline scan, irrespective of the timing of the baseline MRI.

1.2 Brain image analysis in multiple sclerosis

1.2.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive and painless procedure used in radiology to create detailed, cross-sectional images of the organs and tissues within the body. MRI scanners use strong magnetic fields and radio waves to acquire the three-dimensional images without the use of damaging radiation.

Normally, the hydrogen nuclei in water molecules in the body are randomly oriented, but when entering an MRI scanner, the very strong magnetic field causes them to align in one direction. The scanner also produces a radio frequency variation in the magnetic field, which is turned on and off, causing each hydrogen nucleus to change its alignment when switched on and rapidly relax back to its original state when it is switched off. The protons in different types of tissue realign at different speeds and produce distinct signals that can be measured by receivers in the scanner and made into an image. The contrast between different tissues is determined by the rate at which excited atoms return to the equilibrium state. The faster the protons realign, the brighter the image.

The contrast in an MR image can be manipulated by changing the pulse sequence parameters. The most commonly used sequences for brain analysis in MS include T2-weighted (T2-w), Proton Density-weighted (PD-w), Fluid Attenuated Inversion Recovery (FLAIR) and T1-weighted (T1-w), although other modalities such as Magnetization-Prepared Rapid Acquisition with Gradient Echo (MP-RAGE), Double Inversion Recovery (DIR), and Phase-Sensitive Inversion Recovery (PSIR) can also be helpful for disease diagnosis and follow-up.

In MRI, MS plaques appear as focal high signal intensity areas (hyper-intense with respect to gray matter (GM)) on T2-w, PD-w, FLAIR and DIR modalities, whereas they show hypo-intense areas with respect to WM on T1-w, MP-RAGE and PSIR (see Figure 1.2). T2-w, PD-w and FLAIR sequences are often used in white matter (WM) lesion detection and delineation, as these lesions appear brighter than GM and WM in those modalities, favoring an easier differentiation. However, cortical lesions are rarely visualized on conventional MRI sequences and are better seen on DIR, PSIR and MP-RAGE modalities. On the other hand, T1-w and MP-RAGE images show good contrast between tissues, which makes them the most commonly used sequences for tissue and structure delineation.

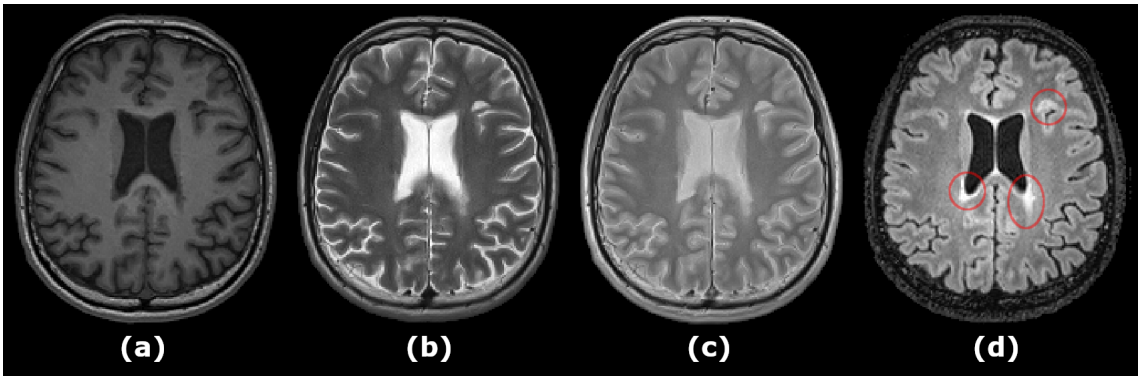


Figure 1.2: Axial view of the same slice on different MRI image modalities. Most frequently used MRI sequences in multiple sclerosis. (a) T1-weighted sequence. (b) T2-weighted sequence. (c) PD-weighted sequence. (d) FLAIR sequence. MS lesions are shown inside red circles on the FLAIR modality.

1.2.2 Automatic segmentation

Manual analysis of brain MRI scans is, in practice, a highly time-consuming task. Given the large number of two-dimensional slices contained in each three-dimensional MR image and the increasing number of images to analyze, the capacity of expert visual analysis is soon exceeded. Furthermore, manual analysis is error-prone and subject to inter- and intra- operator variability. These conditions have led, in recent years, to the development of a wide number of automatic image processing methods, with the aim of reducing the time needed for manual interaction and the inherent variability of manual annotations.

Lesion segmentation

Brain MS lesions are typically small (<1 cm diameter) and ovoid producing a homogeneous signal on T2-w sequences. Furthermore, they are located in characteristic regions of the brain (see Figure 1.3), that include the peri-ventricular (including the corpus callosum), cortical or juxta-cortical (abutting the cerebral cortex), and infra-tentorial regions. Cortical lesions, at the same time, can be classified according to their location within the GM as leuko-cortical (involving the deeper layers of the gray matter as well as the adjacent white matter at the gray/white matter junction), intra-cortical (small demyelinated lesions often centered around blood vessels and confined within the cortex), subpial (extending from the pial surface into the cortex) and lesions that extend to the entire width of the cortex. The largest linear measurement for lesion definition should be 3 mm or more in at least one plane.

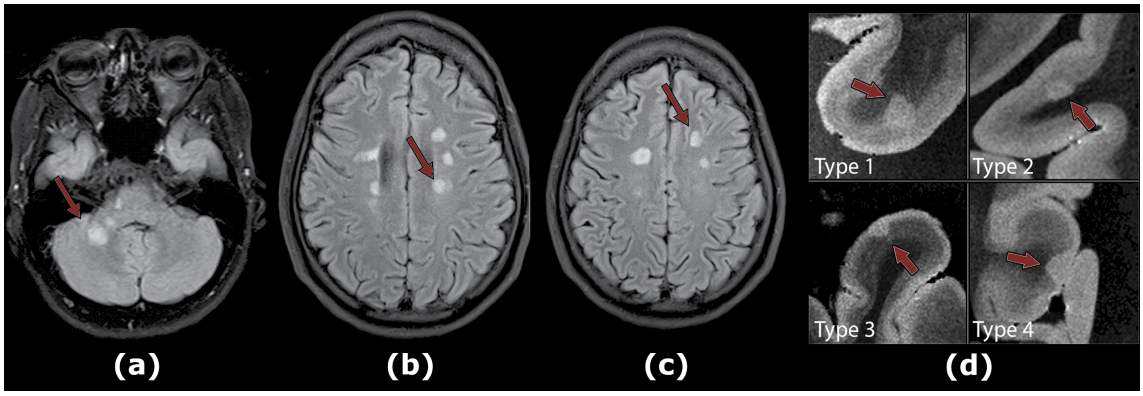


Figure 1.3: Multiple sclerosis lesion location within the brain. Characteristic location of lesions include the infra-tentorial (a), peri-ventricular (b), and juxta-cortical (c) regions. Cortical lesions (d) can be classified into leuko-cortical (type 1), intra-cortical (type 2), subpial (type 3) and lesions that extend to the entire width of the cortex (type 4). Images from [6, 12].

As seen in section 1.1.4, McDonald criteria aims to use MR images in order to provide evidence of lesion dissemination in time and space, conditions that have to be fulfilled for a definite MS diagnosis. For this reason, a need to analyze focal MS lesions quantitatively in individual and temporal studies has emerged in recent years.

A wide number of automated WM lesion segmentation techniques have been proposed over the last few years, the ones that use combinations of different MRI modalities being the most widely used in the literature [13]. Within these methods, two kinds of segmentation strategies can be roughly differentiated. Supervised approaches [14, 15], that use any kind of *a priori* information or knowledge to perform the MS lesion segmentation; and unsupervised strategies [16, 17, 18], where no prior knowledge is used. Interestingly, excellent results have been reported in the last few years within this topic, with methods achieving segmentations that are close to human expert inter-rater variability [19].

Tissue segmentation

Brain tissue segmentation refers to the process of partitioning the brain into its three main tissues (see Figure 1.4 (d)), i.e. white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). In practice, one of the main goals of this technique is to quantify the volume of each of these three regions, so that we can evaluate their evolution over time. When tissue segmentation is applied to brain MR images of MS patients, the neuronal and axonal loss that these patients experience as the

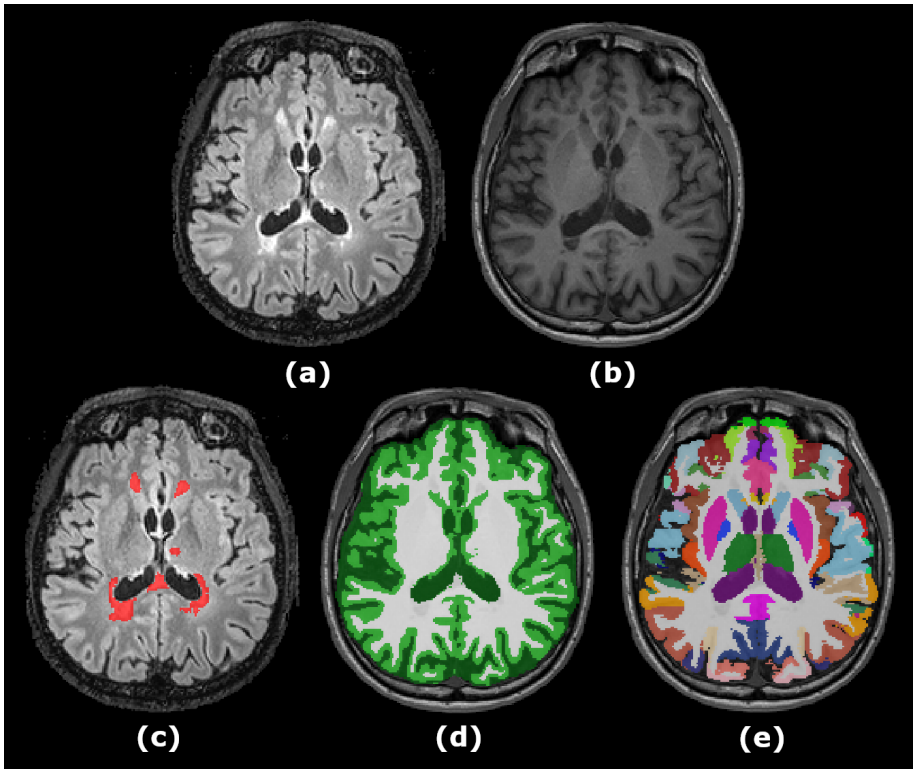


Figure 1.4: Image analysis in multiple sclerosis. (a) FLAIR sequence. (b) T1-w sequence. (c) Lesion segmentation performed on FLAIR modality. (d) Automatic tissue and (e) structure segmentation performed on T1-w modality.

disease progresses, can be quantitatively evaluated. This neurodegeneration can be observed as a reduced brain volume (or brain atrophy) on patient follow-up. This quantification is very useful for practical treatment evaluation, since it has been proved that there is a correlation between brain tissue atrophy measurements and MS disability status [20, 21].

The most commonly used sequence for tissue segmentation is T1-w, as in this modality, there is a clear difference in the intensity distributions of these three tissues. For this reason, most of the unsupervised automatic tissue segmentation methods in the current state of the art rely only on the signal intensity in this sequence [22, 23, 24]. In contrast, supervised learning approaches also combine T1-w sequences with other modalities such as T2-w and PD-w [25, 26, 27]. However, most of these brain tissue segmentation methods are not designed to deal explicitly with MS lesions, which can reduce their accuracy when applied to MS patient images [28, 29, 30].

WM lesions are hypo-intense with respect to normal-appearing WM on T1-w,

with an intensity profile that tends to lie between that of the GM and the WM. In that sense, if lesion voxels are classified as GM, they will distort the overall GM volume, causing a risk that normal-appearing WM voxels with signal intensities similar to the lesions could be mis-classified as GM. On the other hand, if WM lesions are classified as WM, the mean overall signal intensity of the WM will increase, causing GM voxels with signal intensities similar to WM lesions to be also mis-classified as WM.

A commonly used technique to overcome this issue consists of in-painting the lesions on the T1-w sequence with signal intensities of the normal-appearing WM before segmentation. This pre-processing step is commonly referred to in the literature as *lesion filling*. Several lesion filling approaches have been proposed in recent years [29, 30, 31], achieving a significant reduction in the associated errors of WM lesions in tissue volume measurements [32].

Structure segmentation

Although tissue segmentation is highly used in medical practice, it only divides the brain into its three main tissues, which may be insufficient in order to perform more exhaustive analyses of the disease evolution. Brain structure segmentation algorithms, on the other hand, provide a more detailed division of the brain (see Figure 1.4 (e)). In this technique, the three main tissues of the brain, i.e. WM, GM and CSF, are subdivided into their sub-structures, with different levels of detail, depending on the segmentation method.

In a similar way to tissue segmentation, brain structure segmentation provides a tool to quantitatively measure the disease progression. However, this technique also allows focusing on more specific regions of the brain, excluding the ones that provide irrelevant information. This is important in MS, for example, in the case of the GM. This tissue atrophy has been proved to be relevant to the disease progression [33], however, global volume measurement approaches are insufficiently sensitive during the early stages of disease [34]. Some studies have been conducted on isolated groups of structures, such as the deep gray nuclei, concluding that their volume loss in MS patients is predominant compared with that of the periphery [35], and that their atrophy is closely related to the magnitude of inflammation [36]. Besides this, thalamic atrophy itself has been proved to be a clinically relevant biomarker of the neurodegenerative disease process [37].

A large number of automatic brain structure segmentation methods have been proposed during the last two decades. Most of these approaches are designed to segment specific structures [38, 39] or groups of them [40, 41], but strategies to parcellate the whole brain [42, 43] or even the sub-regions contained in some structures [44, 45] are also part of the literature. However, most of these methods are not

intended to segment brains with lesions, such as those of MS patients, which makes their accuracy oscillate when applied to these specific populations.

1.3 Research background

The Computer Vision and Robotics group (VICOROB) of the University of Girona has been working on medical image analysis since 1996, mainly in the segmentation and registration of mammographic images. In 2009, the group started collaborating with several medical teams, experts in MS, with the objective of developing new tools that could be transferred for clinical use. Thanks to the group prior knowledge acquired through previous medical projects, a new line of research emerged, focused on brain MRI analysis. This new line started with the segmentation of MS lesions and has expanded to other fields such as temporal analysis, registration (temporal and inter-subject), atrophy analysis, tissue segmentation and brain structure segmentation.

All these studies have been accomplished inside the framework of several research projects:

1. [2015 - 2017] TIN2014-55710-R: NICOLE: "Herramientas de neuroimagen para mejorar el diagnosis y el seguimiento clínico de los pacientes con Esclerosis Múltiple" awarded in 2014 by the Spanish call Retos de investigación.

2. [2015 - 2019] BiomarkEM.cat: "New technologies applied to clinical practice for obtaining biomarkers of atrophy and lesions in magnetic resonance images of patients with multiple sclerosis". Awarded in 2015 by the Fundació la Marató de TV3.

3. [2016 - 2019] WASSABI: "Automatic brain Structures Segmentation As potential imaging Biomarkers" awarded in 2015 by the Spanish call Retos de investigación - Jóvenes Investigadores.

Since then, the group has published original contributions in different research fields such as image pre-processing [46], lesion segmentation [13, 16, 19, 47, 48, 49], temporal analysis [50, 51], image registration [52, 53], and tissue segmentation [54]. All the projects have been carried out in collaboration with different medical MS teams from:

- The Hospital Vall d'Hebron: Dr. Rovira is the director of the "Unitat de Ressonància Magnètica-Centre Vall d'Hebron" (URMVH) and has participated in numerous research projects funded by public and private institutions in the

last few years, as well as Dr. Pareto and technicians Huerga and Corral. This group is part of the MAGNetic resonance Imaging in MS (MAGNIMS) network, a European network of centers that share an interest in the MS study through MRI.

- The Hospital Josep Trueta: Dr. Ramió-Torrentà, who is the current coordinator of the "Unitat de Neuroimmunologia i Esclerosi Múltiple", as well as Dr. Robles and Dr. Beltrán, who work in the neurology and radiology units, respectively.

- The Clínica Girona / Hospital Santa Caterina: Dr. Vilanova and Dr. Barceló are the codirectors of the "Unitat de Ressonància Magnètica" at the Clínica Girona and are members of several national and international radiology societies.

1.4 Objectives

As part of the BiomarkEM.cat research project framework, the main goal of this PhD thesis is:

to develop novel brain parcellation algorithms capable of computing accurate structure segmentations in images of multiple sclerosis patients.

Different sub-objectives have to be covered in order to fulfill this main goal. Each of the following stages can be considered as sub-objectives that allow us to gain a better knowledge of the problem that we want to overcome. These are:

- **to exhaustively analyze and classify the literature on brain structure segmentation**. This stage aims to review and classify the different brain structure segmentation techniques proposed in the literature. In order to fulfill this goal, we plan to propose a classification of the algorithms based on the segmentation strategy acquired, and analyze their advantages and drawbacks, with the purpose of better understanding each segmentation strategy's strengths and weaknesses. Besides this, we also plan to classify the methods based on their target structures and analyze the clinical applications of the most relevant ones.

- **to quantitatively review and evaluate the state of the art of automatic brain structure segmentation methods**. The aim of this stage is to quantitatively review the methods which are currently the state of the art.

With this goal in mind, we plan to give an overview of the most commonly used databases and metrics for evaluation on brain structure segmentation, as well as a review of the results obtained for each of the methods in the literature, separated by target structure, database, and evaluation metric. Furthermore, we plan to evaluate different publicly available segmentation methods, each relying on a different category of the previously proposed classification, using public databases of healthy subjects that incorporate manual brain structure annotations, which will allow us to perform a fair comparison of the accuracy of the methods.

- **to study and evaluate the effect of MS lesions on different brain structure segmentation strategies**. Although it is known that the presence of MS lesions distorts the measurements of brain volume when using automatic tissue segmentation algorithms, this effect has not yet been studied and compared for structure segmentation methods. In this respect, our third sub-goal focuses on the analysis of the effects of MS lesions on the segmentation result of a set of brain structure segmentation approaches. Our hypothesis here is that a better knowledge of the behavior of different segmentation strategies against lesions may be beneficial for designing a brain structure segmentation method for MS. Thus, we plan to perform a pioneer and exhaustive analysis of this effect based on the segmentation strategy used, the lesion location, the affected structures and the total lesion load.

- **to propose robust brain structure segmentation methods for MS patient images**. Then, we aim to benefit from these sub-objectives in order to propose **novel and robust whole brain parcellation strategies** able to deal with images containing MRI visible lesions, such as those of MS patients. In this fourth stage, we aim to validate the accuracy of the proposed approaches by comparing them to other brain parcellation methods in the current state of the art, as well as to the well-established pre-processing step, lesion filling, which is highly effective in automatic tissue segmentation.

- **to present and evaluate a fully automated pipeline for brain structure segmentation in MS**. In this stage, we aim to develop a segmentation pipeline able to work without the need of human interaction. For this reason, our fifth sub-goal is to fully automate the process of brain structure segmentation, combining the approach proposed in the previous step with automatically segmented lesion masks. Next, we plan to validate the proposed pipeline and compare their segmentation results with those obtained with the current state-of-the-art methods when applying lesion filling.

- **to extend the proposed strategy in order to enable the integration of manual and automatic edits.** To conclude, we believe that integrated

segmentation algorithms, that not only provide a solution based on a closed set of labels, but are also flexible to the addition of other known labels into the final segmentation, could be beneficial. For this reason, our last subgoal aims to extend the theory of the proposed strategies in order to enable the integration of manual and automatic edits, known beforehand, into the segmentation estimation.

1.5 Document structure

The rest of this document is organized as follows:

- **Chapter 2. A review on automatic brain structure segmentation in magnetic resonance imaging [55].** In this chapter, we perform an exhaustive review of the brain structure segmentation methods of the literature. First we divide the algorithms according to their target structures, and then we propose a general classification based on their segmentation strategy. We further discuss each category's strengths and weaknesses and analyze its performance in segmenting different brain structures providing a qualitative and quantitative comparison.

- **Chapter 3. The effect of multiple sclerosis lesions on automatic brain structure segmentation [56, 57].** After reviewing the state of the art on brain structure segmentation methods and assigning each of them a category in our classification, we select a representative method from each main category, and analyze its performance against MS lesions. We compare the methods based on the database used, the affected structures, and the total lesion load.

- **Chapter 4. A multi-atlas approach for brain structure segmentation in the presence of multiple sclerosis lesions [58, 59].** In this chapter, we propose a new non-local correspondence model for intensity-based multi-atlas segmentation, capable of overcoming the weaknesses of this strategy against MS lesions. We integrate this model into two state-of-the-art statistical label fusion methods, and validate its accuracy with respect to other methods in the literature. We further compare our proposal to the well-established lesion filling technique, which is a very popular pre-processing step in automatic tissue segmentation.

- **Chapter 5. A fully automated pipeline for brain structure segmentation in multiple sclerosis.** Since we aim at developing whole brain parcellation strategies for MS patients, which do not require human expert interaction, we present in this chapter a fully automated pipeline to achieve this goal, that

includes the use of automatically segmented lesion masks. Then, we evaluate this pipeline and analyze its results in comparison to the ones obtained with the manual masks. Lastly, we analyze the effect of using the same automatic lesion masks in the segmentation result of the current state-of-the-art methods when applying lesion filling.

- **Chapter 6. Integration of known label masks into the label fusion estimation [59].** In this chapter, we extend the theory of the methods presented in Chapter 4 in order to enable the integration of manual and automatic edits into the target segmentation. Then, we qualitatively evaluate the extended methods, and discuss the results obtained on a set of MS patient images. In addition, we also validate the proposed strategies on patient images of other diseases, including diabetes and cerebral infarction, giving a small overview of how the methods presented in this thesis can be extended to other medical fields.

- **Chapter 7. Conclusions and future work.** Lastly, a comprehensive discussion of the results obtained, as well as the main conclusions based on the contributions of this thesis are defined. Based on these conclusions, we also point out different future investigations to improve and extend the work carried out for this PhD thesis.

Chapter 2

A review on automatic brain structure segmentation in magnetic resonance imaging

2.1 Introduction

MRI of the brain has become a standard tool in medical practice for diagnosis [60], disease follow-up [61], treatment evaluation [62] and brain development monitoring [63]. As it is non-intrusive, painless, fast to acquire and provides good contrast between tissues it provides the best choice for a range of clinical application areas.

Brain segmentation is very useful for clinical analysis, since a qualitative evaluation of brain morphological characteristics is very subjective, and therefore quantified techniques are needed. Given the shortcomings of manual segmentation seen in Chapter 1, the need for accurate automatic segmentation methods has emerged in recent years [40, 64, 65, 66, 67]. Initial automatic methods were focused on segmenting the brain MRI into three different tissues, namely WM, GM and CSF. Most of these methods relied only on signal intensity in T1-w images, where there is a clear difference in the intensity distributions of these tissues. However, segmenting the brain structures is not as straightforward. It cannot be performed based only on image intensities because there is too much overlap between the class distributions, and the structure boundaries are not always clear enough. Hence, more information such as shape, location in the brain or the relative position among structures has to be incorporated into the segmentation algorithms.

The automatic segmentation of the hippocampus has received significant attention in recent years, since the volume reduction of this structure has been related to several diseases, being also an important predictive biomarker for Alzheimer's disease. In the recent work of Dill et al. [68] a review of the evolution of automated methods for the segmentation of the hippocampus in MRI was presented, whereas [69] covered a quantitative comparison of four automatic methods to segment the hippocampus in patients with mesial temporal lobe epilepsy. Another group of structures to which the community has paid attention is the deep GM or sub-cortical structures. Several approaches [40, 70, 71] have been presented focused on automatically segmenting this group of structures, and a recent study reviewed automatic and semi-automatic methods [72]. In addition, Babalola et al. [73] quantitatively evaluated four different algorithms for the task of sub-cortical brain structure segmentation. Other works in brain structure segmentation have been presented recently, such Devi et al. [74], who reviewed several works for neonatal brain segmentation in MRI, and Iglesias et al. [75] who gave an overview of the current state of the art in multi-atlas segmentation of biomedical images. Besides these, Klein et al. [76] quantitatively compared 14 non-linear registration algorithms which were evaluated based on brain structure segmentation. As far as we know, there is no work which reviews the current state of the art and describes automatic methods not concentrating on an exclusive segmentation strategy and that either segment a single structure or the whole brain in MR images.

In this chapter, we present a review of such methods and classify them according to the segmentation strategy used, for which we propose a classification that includes atlas-based, learning-based, deformable, region-based and hybrid categories. Furthermore, to the best of our knowledge, the analysis presented here is the first attempt to review the most relevant works in brain structure segmentation that also presents an analysis of the state-of-the-art results, showing different evaluation metrics, databases and number of test cases, both from the point of view of segmentation strategies and segmented structures.

We reviewed the literature in the following databases: PubMed, Scholar, IEEE Xplore and Scopus. The main search strategy combined four concepts: brain structures, magnetic resonance imaging, segmentation, and automatic methods. From the initial search we discarded those works that were out of the scope of this review, such as tissue segmentation methods, approaches that used contrast MRI or diffusion tensor images, while keeping only the works that strictly proposed an automatic approach for brain structure segmentation in structural MRI. To retrieve other relevant publications, we also examined the reference lists of the selected publications, and we included those works that were related to our aim.

2.2 Clinical applications

Psychiatric and neurodegenerative disorders are frequently associated with structural changes in the brain, such as variations in the volume or shape of the deep GM structures or in the thickness, area and folding pattern of the cortical regions [43]. Because of that, the morphometric analysis of brain structures can be used as an important biomarker of the disease or even as a diagnostic test [77]. Other applications of MRI brain structure segmentation may include pre-operative evaluation and surgical planning [78] for situations in which the procedure requires high accuracy, such as deep brain stimulation [79] or ablation of the appropriate functional regions; longitudinal monitoring for disease progression or remission [80]; or radiotherapy treatment planning [81]. In this section we briefly review the clinical applications of the segmentation of some brain structures such as the hippocampus, caudate nucleus, thalamus, pallidum or brainstem. Table 2.1 provides a summary of the relation of such structures with different diseases and the consequent structure abnormalities.

The hippocampus plays an important role in human memory and orientation. Its atrophy has been shown to be a predictive biomarker for patients with mild cognitive impairment and Alzheimer's disease, but it has also been related to other diseases such as schizophrenia, major depression, bipolar disorder, post-traumatic stress disorder, etc. [83]. Asymmetric atrophy of this structure has also been proved to be a good predictor of epilepsy [85].

The brainstem, which is usually described as including the medulla oblongata, pons, and midbrain (red nucleus and substantia nigra), is especially relevant to primary tauopathies such as progressive supranuclear palsy [95], in which brain atrophy occurs in the midbrain and pons, or corticobasal degeneration. The importance of the brainstem sub-structures in other diseases has also been reported, such as in [120] where the authors proposed a technique for supporting the clinical diagnosis of Parkinson's disease; they claimed that the initial assessment of the neurological condition of a patient should be performed by estimating the area of the substantia nigra [121]. Alzheimer's disease, is another degenerative disease that also affects brainstem structures [96].

The diminished right caudate volume is one of the most replicated findings among attention deficit hyperactivity disorder patients and hence, the ratio between right caudate volume and the bilateral caudate volume is applied as a diagnostic test [77]. Aberrant morphology and function of the caudate nucleus have also been associated with a number of important brain disorders, including Huntington's disease [100], Tourette syndrome [101], autism [102, 122, 123], attention deficit hyperactivity disorder [104, 124], and fragile X syndrome [105, 125].

Table 2.1: Clinical applications. Brain structure abnormalities associated with various diseases.

Structure	Implied Disease	Abnormality
Hippocampus	Alzheimer's	Atrophy [82]
[83, 84]	Temporal lobe epilepsy	Asymmetric atrophy [85]
	Posttraumatic stress disorder	Reduced volume [86, 87]
	Major depression	Reduced volume [88]
	Schizophrenia	Reduced volume [89]
	Bipolar disorder	Non-conclusive volume difference [90, 91]
Brainstem	Progressive supranuclear palsy	Atrophy in midbrain and pons [92, 93, 94]
[95, 96]	Parkinson	Reduced nigral volume [97]
	Alzheimer's	Reduced volume and structure deformation [98]
Caudate	Huntington's disease	Atrophy [99]
[100, 101, 102]	Tourette syndrome	Reduced volume [103]
	Autism	Increased right volume [102]
	Attention deficit hyperactivity disorder	Reduced right volume [104]
	Fragile X syndrome	Increased volume [105]
Thalamus	Multiple sclerosis	Atrophy [37]
[36, 106]	Alzheimer's	Atrophy [107]
	Schizophrenia	Non-conclusive volume difference [108]
	Parkinson	Reduced volume [109]
Corpus Callosum	Multiple sclerosis	Atrophy [110]
[111, 112]	Schizophrenia	Reduced volume [113]
	Autism	Reduced volume [111]
	Alzheimer's	Atrophy [111]
	Multi-Infarct dementia	Atrophy [114]
Amygdala	Schizophrenia	Reduced volume [115]
[84, 116]	Anxiety disorders	Reduced left volume [117]
	Bipolar disorder	Non-conclusive volume difference [89, 91, 118, 119]

The thalamus is associated with a wide range of clinical manifestations including cognitive decline, motor deficits, fatigue, painful syndromes, and ocular motility disturbances in patients with MS. It has also been proved that the atrophy of deep gray nuclei is closely related to the magnitude of inflammation [36]. As stated before, the surgical treatment for many movement disorders, such as essential tremor, Parkinson's disease, drug-resistant epilepsy as well as chronic pain syndromes, involves ablation or electric stimulation of the appropriate functional region within inner-brain structures such as the sub-thalamic nucleus and globus pallidus [126]. Surgical planning for these procedures is often based on pre-operatively acquired MR images, thus segmenting the implied regions would improve the planning and guidance of the surgery.

The corpus callosum is also an important structure due to its vulnerability to environmental toxins, WM diseases (such as MS) and schizophrenia [127]. Effects

on regional callosal structure have been reported in attention deficit hyperactivity disorder, Alzheimer's disease, multi-infarct dementia, and a range of neurodevelopmental disorders and dysplasias [111].

2.3 Methods

Looking at the literature, we have seen that among automatic brain structure segmentation methods, some aim at parcellating the whole brain but, the vast majority are centered on segmenting a few or even only one specific structure. Tables 2.2 to 2.4 show in their 'C' column, for all the reviewed methods that include segmentation results, either graphically or numerically, the type of method we are referring according to the following criteria: (1) methods which parcellate the whole brain, (2) methods that segment a group of structures such as sub-corticals, basal ganglia or those that, according to their results, have been demonstrated that their method can be extended to several structures, (3) structure specific methods, and (4) methods that segment a specific structure and its sub-structures.

In this section, we provide a classification of the state-of-the-art methods according to the strategy used to segment the target structures, for which we mainly distinguish four categories: atlas-based, learning-based, deformable and region-based strategies. A fifth category which combines some of these four approaches is also included in our classification.

2.3.1 Atlas-based methods

In the context of image segmentation, an atlas is defined as the combination of two image volumes: one intensity image (or template) and one segmented image (or labeled image). As stated in Cabezas et al. [128], an atlas can be either topological or probabilistic. Topological or deterministic atlases consist of a single subject volume together with its corresponding, often manual, segmentation. Probabilistic or statistical atlases are constructed on the basis of populations, co-registering all the segmented cases to a standard space and computing the frequency of each voxel to belong to a specific structure.

While the methods presented in this section are all based on topological atlases, statistical atlases can also be used as prior information in statistical image segmentation algorithms, as we will see later in Section 2.3.2.

Table 2.2 shows a summary of the methods found in this category that provided either graphical or quantitative experimental results. The table indicates also the target structures of each method based on the results presented.

Table 2.2: Atlas-based methods. Acronyms from left to right are: hippocampus (HIP), thalamus (THA), caudate nucleus (CAU), putamen (PUT), pallidum (PAL), amygdala (AMY), accumbens (ACC), lateral ventricle (LV), brainstem (BS), corpus callosum (CC), cerebellum (CB), white matter (WM), cortical gray matter (CGM) and cerebrospinal fluid (CSF). Metrics in order of appearance: Kappa index (KI), Jaccard index (JC), Relative overlap (RO), Dice similarity coefficient (DSC), Relative mean squared error (MSE), Relative volume (RV), False negatives (FN), False positives (FP), Hausdorff distance (HD), Similarity index (SI), Mean absolute distance (MAD), Precision (P) and Recall (R). Diseases in order of appearance: Normal controls (NC), Alzheimer's disease (AD), Clinical dementia (CD), Mild cognitive impairment (MCI), Probable Alzheimer's disease (PAD) and First episode psychosis (FEP). The ✓ and × symbols stand for numerical and graphical results, respectively, whereas the ⊛ symbol means that the method perform the internal sub-structure segmentation of the indicated structure. The - symbol indicates that no results have been reported for that particular structure. Column 'C' indicates the segmentation target: (1) whole brain; (2) a group of structures; (3) a single structure; (4) a single structure and its sub-structures. Column 'Ref' shows the reference work from which the results are obtained (in case they do not come from the original work).

	Article	Ref	C	Segmented structures															Metric	Database	Disease
				HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS	CC	CB	WM	CGM	CSF	Others			
Label propagation	Collins (1997) [64]	[129]	1	✓	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	KI;JC	80v (ICBM)	NC
		[130]		✓	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	KI	30v (ICBM); 10v (ICBM)	NC
	Shen (2002) [131]	[132]	1	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	✓	✓	-	✓	JC	11v (IBSR); 36v (Desikan et al.)	AD+NC
	Shattuck (2002) [133]		1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Wang (2005) [134]		2	×	-	-	-	-	-	-	×	-	×	-	-	-	-	-	-	-	-
	Postelnicu (2009) [132]		1	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	✓	✓	-	✓	JC	11v (IBSR); 36v (Desikan et al.)	AD+NC
	Lin (2010) [71]		2	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	KI;RO	15v (IBSR)	NC
	Luo (2011) [135]		2	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	DSC	9v (IBSR)	NC
	Iacono (2011) [126]		4	-	-	-	-	⊛	-	-	-	-	-	-	-	-	-	-	RMSE	1v (MNI152)	NC
	Yousefi (2012) [136]		2	×	×	×	×	×	×	×	×	×	-	-	-	-	-	-	DSC;RV;FN;FP	18v (IBSR)	**NC**
				×	-	×	×	-	-	-	-	-	-	-	-	-	-	-	DSC;RV;FN;FP	40v (LPBA40)	NC
	Joshi (2012) [137]		1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	HD	6v	-
Label fusion	Warfield (2004) [65]	[138]	1	✓	✓	✓	✓	✓	✓	-	✓	-	-	-	✓	✓	-	-	DSC	39v (FS atlas)	AD+CD+NC
		[139]		✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	SI; MAD	17v (IBSR)	NC
	Heckemann (2006) [140]		1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	✓	SI	30v	NC
	Aljabar (2007) [141]		2	×	×	-	-	-	-	-	×	-	-	-	-	-	-	-	DSC	275v (CMA)	-
	Aljabar (2009) [142]		2	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	DSC	275v (CMA)	-
	Artaechevarría (2009) [139]		1	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	SI; MAD	17v (IBSR)	NC
		[143]		✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	DSC	30v (Hammers)	NC
		[143]		✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	30v (ADNI)	AD+MCI+NC
		[144]		✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	DSC	18v (IBSR)	NC
	Lötjönen (2010) [145]		2	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	SI; P; R	18v (IBSR)	NC
				✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	SI	60v (ADNI)	AD+MCI+NC
		[144]		✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	DSC	18v (IBSR)	NC
	Collins (2010) [129]		2	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	KI; JC	80v (ICBM)	NC
	Heckemann (2010) [146]		1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	✓	JC	30v (Hammers)	NC
		[143]		✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	DSC	30v (Hammers)	NC
	Wolz (2010) [147]		3	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	796v (ADNI)	AD+MCI+NC
	Rousseau (2011) [144]		1	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	DSC	18v (IBSR)	NC
	Coupé (2011) [42]		2	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	KI	80v (ICBM)	NC
				-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	KI	80v	AD
		[148]		✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	80v (ICBM); 202v (ADNI)	AD+MCI+NC
	Zhang (2012) [149]		1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	5v (NA0-NIREP)	NC
	Jia (2012) [150]		1	-	-	-	-	-	-	-	✓	-	-	-	✓	✓	✓	-	DSC	50v (ADNI)	MCI+NC
	Cardoso (2013) [143]		1	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	30v (ADNI)	AD+MCI+NC
				✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	DSC	30v (Hammers)	NC
	Wang (2013) [151]		1	✓	✓	✓	✓	-	✓	✓	-	✓	-	-	✓	✓	✓	-	DSC	20v (MICCAI'12)	CD(PAD)+NC
	Asman (2013) [152]		1	×	×	×	×	×	×	-	-	×	-	×	×	×	×	×	DSC	15v (MICCAI'12)	CD(PAD)+NC
	Wang (2014) [153]		1	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	✓	DSC; MAD; HD	20v (MICCAI'12)	CD(PAD)+NC
	Pipitone (2014) [154]		3	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	60v (ADNI1); 81v (FEP)	AD+MCI+NC+FEP
				⊛	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	5v (Winterburn et al.)	NC
	Wu (2015) [155]		1	✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	DSC	30v (Hammers)	NC

Table 2.3: Learning-based methods. Acronyms from left to right are: hippocampus (HIP), thalamus (THA), caudate nucleus (CAU), putamen (PUT), pallidum (PAL), amygdala (AMY), accumbens (ACC), lateral ventricle (LV), brainstem (BS), corpus callosum (CC), cerebellum (CB), white matter (WM), cortical gray matter (CGM) and cerebrospinal fluid (CSF). Metrics in appearance order: Classification rate (CR), Reproducibility (Rep), Dice similarity coefficient (DSC), Modified Hausdorff distance (HD), Hausdorff distance (HD), Precision (P), Recall (R), Relative overlap (RO), Similarity index (SI), Mean absolute distance (MAD), Overlap error (OE), MICCAI'07 score (Score), Overlap by pairs (OBP), Tanimoto coefficient (TAO), Sensitivity (SN), Specificity (SP), Accuracy (Acc), Jaccard index (JC) and Volume (Vol). Diseases in order of appearance: Normal controls (NC), Clinical dementia (CD), Probable Alzheimer's disease (PAD), Alzheimer's disease (AD), Mild cognitive impairment (MCI), Schizotypical personality disorder (SPD), Autism (AU), Parkinson disease (PD), Attention deficit hyperactivity disorder (ADHD), Schizophrenia (SZ), Bipolar disorder (BD), Major depressive disorder (MDD), Elderly controls (EC) and Neonatal normal controls (NNC). The ✓ and × symbols stand for numerical and graphical results, respectively, whereas the ⊛ symbol means that the method perform the internal sub-structure segmentation of the indicated structure. The - symbol indicates that no results have been reported for that particular structure. Column 'C' indicates the segmentation target: (1) whole brain; (2) a group of structures; (3) a single structure; (4) a single structure and its sub-structures. Column 'Ref' shows the reference work from which the results are obtained (in case they do not come from the original work).

	Article	Ref	C	Segmented structures															Metric	Database	Disease
				HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS	CC	CB	WM	CGM	CSF	Others			
Supervised	Pitiot (2002) [156]		2	✓	-	✓	-	-	-	-	-	-	✓	-	-	-	-	-	CR	10v	-
	Deoni (2007) [157]		4	-	⊛	-	-	-	-	-	-	-	-	-	-	-	-	-	Rep	4v (T1&T2)	NC
	Bao (2018) [158]		2	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	DSC	9v (IBSR); 20v (LPBA40)	NC
	Mehta (2017) [159]		1	×	×	×	×	×	×	×	×	×	-	×	×	×	×	-	DSC	15v (MICCAI'12)	CD(PAD)+NC
				✓	✓	✓	✓	✓	✓	-	✓	-	-	-	-	-	-	-	DSC	9v (IBSR)	NC
				✓	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-	DSC	20v (LPBA40)	NC
				✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	DSC	15v (Hammers)	NC
	Shakeri (2016) [160]		2	-	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	DSC	18v (IBSR)	NC
	Dolz (2018) [161]		2	-	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	DSC; MHD	18v (IBSR)	NC
	Wachinger (2018) [162]		1	×	×	×	×	×	×	-	×	×	-	×	×	×	-	×	DSC	10v (MICCAI'12)	CD(PAD)+NC
	Kushibar (2018) [163]		2	✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	DSC; HD	20v (MICCAI'12); 18v (IBSR)	CD(PAD)+NC
	Morra (2008) [164]		2	✓	-	×	-	-	-	-	-	-	-	-	-	-	-	-	P; R; RO; SI; HD	83v (AD)	AD+MCI+NC
	Moghaddam (2009) [165]		2	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	DSC; MAD; HD	6v (IBSR)	NC
	Wels (2009) [166]		2	✓	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	DSC; MAD	18v (IBSR)	NC
				-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	OE; MAD	MICCAI'07	SPD+AU+PD+NC
	Tu (2010) [167]		1	-	-	×	-	-	-	-	-	-	-	-	-	-	-	-	Score	MICCAI'07	SPD+AU+PD+NC
	Traynor (2011) [45]		4	-	⊛	-	-	-	-	-	-	-	-	-	-	-	-	-	OBP; TAO	16v; 18v	NC
	Igual (2012) [77]		4	-	-	⊛	-	-	-	-	-	-	-	-	-	-	-	-	SN; SP; Acc	39v (VH)	ADHD+NC
	Tong (2013) [148]		3	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC	202v (ADNI); 80v (ICBM)	AD+MCI+NC
		[41]		-	-	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	DSC	35v (MICCAI'12)	CD(PAD)+NC
	Kim (2013) [38]		3	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	P; R; RO; SI	20v (7T)	-
	Benkarim (2014) [41]		2	-	-	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	DSC	35v (MICCAI'12)	CD(PAD)+NC
Bayesian	Fischl (2002) [43]	[168]	1	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	DSC	39v (Sabuncu et al.)	AD+NC
		[41]		-	-	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	DSC	35v (MICCAI'12)	CD(PAD)+NC
		[138]		✓	✓	✓	✓	✓	✓	-	✓	-	-	-	✓	✓	-	-	DSC	39v (FS atlas)	CD+AD+NC
		[169]		✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	DSC	30v	-
		[170]		✓	-	✓	✓	-	-	-	✓	-	-	-	-	-	-	-	P; R	14v (LONI28)	NC
		[144]		✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	DSC	18v (IBSR)	NC
	Scherrer (2007) [171]		2	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	DSC	BrainWeb	-
	Scherrer (2007) [172]		2	-	-	-	-	-	-	-	-	-	-	-	×	×	×	-	JC	BrainWeb	-
	Akselrod-Ballin (2007) [173]		1	✓	✓	✓	✓	✓	✓	-	-	✓	-	-	✓	✓	✓	✓	DSC; MAD; HD	18v (IBSR)	NC
		[144]		✓	✓	✓	✓	✓	✓	-	-	✓	-	-	✓	✓	✓	✓	DSC	18v (IBSR)	NC
	Pohl (2007) [66]		1	✓	-	-	-	-	✓	-	-	-	-	-	-	-	-	✓	DSC	50v	SZ+BD+MDD+NC
	Scherrer (2009) [174]		2	-	-	-	-	-	-	-	-	-	-	-	×	×	×	-	DSC	18-20v (IBSR)	NC
				-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	DSC	BrainWeb	-
	VanLeemput (2009) [175]		4	⊛	-	-	-	-	-	-	-	-	-	-	-	-	-	-	DSC; MAD	10v	NC
	Riklin-Raviv (2010) [168]		2	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	DSC	39v (Sabuncu et al.)	AD+NC
	Razlighi (2012) [176]		1	×	×	×	×	×	×	-	×	×	-	×	×	×	×	-	DSC	20v (OASIS)	-
	Iglesias (2013) [177]		4	⊛	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Vol	383v (ADNI)	AD+EC
	Makropoulos (2014) [178]		1	✓	✓	✓	-	-	✓	-	✓	✓	✓	✓	-	-	-	-	DSC	20v (ALBERTs)	NNC
	Iglesias (2015) [44]		4	-	-	-	-	-	-	-	-	⊛	-	-	-	-	-	-	DSC; HD; MAD	10v (BS)	NC

Table 2.4: Deformable methods. Region-based methods. Hybrid methods. Acronyms from left to right are: hippocampus (HIP), thalamus (THA), caudate nucleus (CAU), putamen (PUT), pallidum (PAL), amygdala (AMY), accumbens (ACC), lateral ventricle (LV), brainstem (BS), corpus callosum (CC), cerebellum (CB), white matter (WM), cortical gray matter (CGM) and cerebrospinal fluid (CSF). Metrics in appearance order: Relative agreement (RA), Mean absolute distance (MAD), Kappa index (KI), Hausdorff distance (HD), Similarity (Sim), True positive fraction (TPF), Similarity index (SI), Jaccard index (JC), Comparison (Comp), Dice similarity coefficient (DSC), Overlap (Over), Hausdorff distance 95 (HD95), Mean squared error (MSE), False positive rate (FPR), False negative rate (FNR), Volume difference (VD), Precision (P), Recall (R), Mean distance (MD), Relative volume difference (RVD), Average symmetric surface distance (ASSD) and Root mean square distance (RMSD). Diseases in order of appearance: Schizophrenia (SZ), Normal controls (NC), Elderly controls (EC), Prenatal cocaine exposure (PC), Clinical dementia (CD), Probable Alzheimer's disease (PAD), Alzheimer's disease (AD), Attention deficit hyperactivity disorder (ADHD), Parkinson disease (PD), Fragile X syndrome (FXS), Mild cognitive impairment (MCI) and Autism (AU). The ✓ and × symbols stand for numerical and graphical results, respectively, whereas the ⊛ symbol means that the method perform the internal sub-structure segmentation of the indicated structure. The - symbol indicates that no results have been reported for that particular structure. Column 'C' indicates the segmentation target: (1) whole brain; (2) a group of structures; (3) a single structure; (4) a single structure and its sub-structures. Column 'Ref' shows the reference work from which the results are obtained (in case they do not come from the original work). (*)DSC; JC; P; R; HD; HD95; MAD; ASSD; RMSD.

	Article	Ref	C	Segmented structures																Metric	Database	Disease
				HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS	CC	CB	WM	CGM	CSF	Others				
Deformable	Ghanei (1998) [179]		3	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-		RA	11slices	-
	Kelemen (1999) [180]		2	×	×	-	×	×	-	-	-	-	-	-	-	-	-	-		MAD	21v	SZ+NC
	Duchesne (2002) [130]		2	✓	-	-	-	-	✓	-	-	-	-	-	-	-	-	-		KI	30v;10v (ICBM)	NC
	Ashton (2003) [181]		3	×	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	10v	NC
	Pitiot (2004) [81]		2	✓	-	✓	-	-	-	-	✓	-	✓	-	-	-	-	-		HD; MAD	20v	EC
	Shariatpanahi (2006) [182]		2	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-		Sim; TPF; HD	4v (IBSR)	NC
	Colliot (2006) [183]		2	-	×	✓	-	-	-	-	×	-	-	-	-	-	-	✓		HD; MAD; SI	10v	-
	Babalola (2007) [39]		3	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-		JC	24v	EC+PC+NC
	Zarpalas (2011) [184]		2	×	-	-	-	-	×	-	-	-	-	-	-	-	-	-		Comp	13v (OASIS)	CD(PAD)+NC
	Patenaude (2011) [40]		2	✓	✓	✓	✓	✓	✓	✓	-	✓	-	-	-	-	-	-		DSC	37v;42v;17v;87v;14v;139v	SZ+AD+PC+ADHD+NC
	Cerrolaza (2012) [185]		2	-	-	×	×	-	-	-	×	-	-	-	-	-	-	-		Over	87v	NC
	Gao (2012) [186]		2	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-		DSC; HD95	24v (HIP); 24v (CAU)	-
	Fouquier (2012) [187]		2	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-		MAD	30v (IBSR+OASIS)	NC
	Olveres (2013) [121]		3	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-		DSC; HD	10slices	NC
	García (2014) [188]		3	-	×	-	-	-	-	-	-	-	-	-	-	-	-	-		MSE	4v (DB-UTP)	PD
	Al-Shaikhli (2014) [189]		1	-	-	-	-	-	-	-	×	×	-	×	×	×	×	×		DSC; MAD; HD	BrainWeb; MedPix; NCSUIAL	-
	Ettaïeb (2014) [190]		3	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-		HD	10slices	-
Region	Xue (2000) [191]		2	-	✓	✓	✓	-	-	-	✓	-	-	-	-	-	-	-		FPR; FNR; SI; KI	-	-
	Xue (2001) [192]		2	-	✓	✓	✓	-	-	-	✓	-	-	-	-	-	-	-		FPR; FNR; SI; KI	-	-
	Xia (2007) [125]		3	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-		Over	55v	FXS+NC
	Gui (2012) [193]		1	-	-	-	-	-	-	-	-	✓	-	✓	✓	✓	✓	✓		DSC	10v (newborns)	NC
Hybrid	Zhou (2005) [70]		2	✓	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	-		VD; Over; MAD	17v (IBSR)	NC
	Tu (2008) [194]		2	✓	-	✓	✓	-	-	-	✓	-	-	-	-	-	-	-		P; R; HD; MD	14v	-
	Karsch (2009) [195]		2	✓	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-		DSC; Over	-	-
	Sabuncu (2009) [169]		1	✓	✓	✓	✓	✓	✓	-	×	-	-	-	×	×	-	-		DSC	30v	-
	Sabuncu (2010) [138]		1	✓	✓	✓	✓	✓	✓	-	✓	-	-	-	✓	✓	-	-		DSC	39v (FS atlas)	AD+CD+NC
		[143]		✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-		DSC	30v (Hammers)	NC
		[143]		✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-		DSC	30v (ADNI)	AD+MCI+NC
	He (2011) [196]		2	-	✓	-	-	-	-	-	✓	-	✓	-	-	-	-	-		DSC; Over	20v+25v	AU+NC
	Weisenfeld (2011) [197]		1	-	×	×	×	-	-	-	-	-	-	-	×	×	×	-		DSC	14v	NC
	Iglesias (2012) [198]		1	×	×	×	×	×	×	-	×	-	-	×	×	×	-	-		DSC	8v (multimodal)	NC
	van der Lijn (2012) [199]		2	✓	-	-	-	-	-	-	-	-	-	✓	-	-	-	-		DSC; JC; RVD	18v	-
				✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-		DSC; JC; RVD	18v	-
	Iglesias (2013) [200]		1	✓	✓	✓	✓	✓	✓	-	✓	-	-	-	✓	✓	-	-		DSC	8v (PD)	-
	Liu (2013) [170]		2	✓	✓	✓	✓	✓	✓	-	✓	-	-	-	-	-	-	-		DSC	6v (IBSR)	NC
				✓	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-		DSC; HD	15v (LPBA40)	NC
				✓	-	✓	✓	-	-	-	✓	-	-	-	-	-	-	-		P; R	14v (LONI28)	NC
				✓	-	✓	✓	-	-	-	✓	-	-	-	-	-	-	-		HD; MAD	28v (LONI28)	NC
	Hao (2014) [67]		2	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-		(*)	30v+30v (ADNI); 57v	AD+MCI+NC

Label propagation

Label propagation stands on a very straightforward principle, which allows to automatically segment an image using a single training data set (henceforth an atlas). The basic idea of this technique is to spatially map or deform (i.e. register) the atlas image to the volume we want to segment (target image). This registration produces a deformation field that can be used to propagate (i.e. warp) the atlas labels to this new volume in order to get the final segmentation.

To register two images, it is necessary to find a spatial transformation, mapping the content of one image to the corresponding area of the other in such a way the image similarity is maximized [201]. According to this definition, several works have been presented in the last years focused on label propagation that differ on: the image parts considered to perform the alignment (feature-based vs intensity-based), the transformation function employed (parametric vs non-parametric) or the measure used (sum of squared differences, cross correlation, mutual information (MI), landmark distances, etc.) which defines how similar both images are. Collins and Evans [64] proposed a popular intensity-based registration strategy, called ANIMAL (Automatic Nonlinear Image Matching and Anatomical Labeling), which performed at different spatial scales, starting with very blurred data (where only major structures are apparent, such as temporal lobe, ventricles and longitudinal fissure) and increasing details at each step by using less blurred images, refining the registration at each step. The cross-correlation metric was used as a similarity measure and the deformation was constrained to consist of a linear combination of smooth basis warps that were defined by discrete cosine transforms, performing local translations. Alternatively, Wang and Vemuri [134] employed B-splines to perform the deformation with a previously introduced metric [202], called cross cumulative residual entropy, as a similarity measure.

Combining geometric and intensity features for registration should result in more robust methods. This is actually of current interest and we have seen several methods combining intensity-based and feature-based criteria to establish more accurate correspondences in difficult registration problems [203]. Shen and Davatzikos [131] presented an elastic registration algorithm, called HAMMER (Hierarchical Attribute Matching Mechanism for Elastic Registration), that applied deformation to image sub-volumes rather than voxels, based on the similarity of attribute vectors over the whole sub-volume. These attribute vectors consisted of three individual components: edge type (tissue), intensity and geometric moment invariants, being a more robust way to establish anatomical correspondences in the deformation procedure than considering only a measure derived directly from the intensity.

According to the transformation function, which defines how an image is deformed to match the other, we can mainly distinguish between rigid or non-rigid

transformation, which range from smooth regional variation described by a small number of parameters to dense displacement fields defined at voxel level [203]. Klein et al. [76] presented an evaluation of thirteen non-rigid registration algorithms and stated, corroborating Hellier's evaluation [204], that there was a modest correlation between the number of degrees of freedom of the deformation algorithm and the registration accuracy. Similarly, Carmichael et al. [205] stated in their work that registration methods that produced higher degrees of geometric deformation, produced automated segmentations with higher agreement with manual annotations.

Another common technique is to perform first a coarse registration, and refine in a second step the result with another registration method [71, 126]. This is the case for BrainSuite [133]. In its brain labeling tool (SVReg [137]), subject and atlas surfaces are first smoothed and coarsely aligned in 3D space [206]. After that, a curvature-based alignment, followed by volumetric spatial alignment is performed and, once cortical features are aligned, as sub-cortical features trend to be misaligned, an intensity-based registration refinement is done. In a similar way, Postelnicu et al. [132] combined a feature-based with an intensity-based non-rigid registration method. They first aligned cortical folding patterns and using the resulting deformation as initialization they aligned sub-cortical regions, while preserving the cortical alignment. With a similar idea, Luo and Chung [135] presented a method to segment the sub-cortical structures, for which they first obtained a coarse structure-by-structure segmentation by means of affine registrations, exploiting the spatial dependency relations between the deep brain structures to determine the segmentation order. In a second step, the segmentation result was refined performing a non-rigid registration, that used information about the histogram of the gradient magnitudes lying on the structure boundaries. Following also a two step registration (affine + non-rigid surface-based), Iacono et al. [126] presented a method to segment the internal part of the globus pallidus by means of registering an ultra-high resolution atlas (7T MRI), in which this structure was well defined, to the target image. Yousefi et al. [136] compared different strategies to segment sub-cortical structures based on different registration methods, combining affine and non-rigid registration applied to all brain and sub-cortical area. They concluded that the best results were those obtained by means of an affine transformation applied to the entire brain area, followed by a deformable transformation applied only to the sub-cortical structures.

Label fusion

In general, label propagation suffers from two main drawbacks, which are the fact that a simple atlas cannot sufficiently represent the whole population of potential testing data and that the performance and quality of the obtained results are limited by the accuracy of the pairwise registration method. As an attempt to solve the inherent problems associated with label propagation, label fusion techniques have

been extensively developed in recent years. This approach, also known as classifier fusion or multi-atlas segmentation, consists of registering each training subject (i.e. atlas) to the test subject separately so that each atlas label is propagated to the target image space in the same way as in label propagation. Once all these transferred labels are obtained, they are fused to generate a segmentation result of the target image. Across-subject anatomical variability is better captured here than with a single atlas, and the registration error for a particular propagated atlas is less likely to affect the final segmentation when combined with other atlases [141].

Significant research has been done on multi-atlas segmentation with regard to the influence of several factors that affect the final segmentation such as the atlas selection, the best number of atlases involved in the segmentation or the fusion strategy used. Lötjönen et al. [145] developed and compared different similarity measures, atlas-selection strategies and methods to combine multi-atlas segmentation and intensity modeling. They demonstrated that all these factors play an important role in multi-atlas segmentation and optimizing them is clearly reflected in the brain structure segmentation accuracy.

Several atlas selection strategies have been studied in recent years. Wang et al. [153] first built a graph including all the atlases and the target image and once the graph was built, they grouped the atlases in different clusters by searching the shortest path from each atlas to the target. Finally, they chose from each resulting cluster the atlas with the shortest path to the target to perform the label fusion. Collins and Pruessner [129] also selected the best samples for a given subject from the atlas database, but they used normalized mutual information. Aljabar et al. [141] compared different atlas selection strategies such as image similarity, segmentation similarity or demographics and concluded that image-based selection provided better segmentations than random subsets. Regarding the number of atlases used to perform the fusion, Heckemann et al. [140] showed that the segmentation accuracy increased in a logarithmic way when new random atlases were included in the label fusion up to a limiting value, while Aljabar et al. [142] found that beyond a certain number of ranked atlases (based on a similarity criteria such as image similarity or age-based selection) involved in the segmentation ($15 - 25$ depending on the structure) the accuracy of the resulting segmentation decreased. On the other hand, Pipitone et al. [154] came across in their experiments that when the number of templates was set to an even number, the segmentation performance slightly decreased.

Regarding the fusion strategy, we can mainly distinguish between global and local weighting methods. In global combination strategies, the weight of the contribution of each atlas to the segmentation is the same for every voxel. The simplest global strategy is majority voting, which weights each candidate segmentation equally and assigns to each voxel the label that most segmentations agree on. In spite of its

simplicity, it has been shown to result in highly robust segmentations [129, 140, 141, 154, 207]. Another commonly used fusion method is weighted voting, which can be performed either globally [151] or locally [150]. In this strategy, larger weights are assigned to the atlases that show higher similarity to the target image.

Local weighting methods exploit the fact that different atlases may have achieved a good registration at different parts of the target image, and so it makes sense to borrow labels from different atlases at different target locations. A widely used local combination strategy is the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [65], which weights each segmentation based upon its estimated performance level with respect to other available candidate segmentations. It treats image label fusion as a maximum-likelihood problem, which is solved using Expectation Maximization (EM). Several authors have published STAPLE reformulations [143, 152, 197, 208] that include different advances over the original framework. In [139], Artaechevarría et al. studied the performance of different weighting methods: either globally, using similarity measures for the whole volume, or locally, using a small neighborhood area, concluding that local methods should be preferred in regions that show high contrast with neighbor areas, while global methods should be used in regions that show similar intensities to the surrounding structures. They also stated that there is not a single method that emerged as the best for all regions and images.

In non-local label fusion [42, 144, 149, 155], the labels of all the atlas voxels in the neighborhood of the target voxel have a weight in its label assignment based on their similarity. By means of local search windows, the one-to-one mapping constraint existing in traditional local weighting methods is relaxed. Due to the fact that they explore the neighborhood of each voxel, the registration does not need to be precise, hence it is possible to perform a linear registration instead of non-rigid deformation. Following this idea, Rousseau et al. [144] proposed a patch-based framework based on the construction of a weighted graph of non-local similarities that linked together voxels in the target image and the corresponding neighbor voxels in the atlases. They studied several patch aggregation strategies and also tested the influence of several parameters (neighborhood and patch sizes, number of atlases) in the final segmentation accuracy. Coupé et al. [42] also presented a non-local patch-based segmentation strategy which was conceptually very similar to that of Rousseau et al., differing basically on the patch preselection. In their approach, they first performed an atlas preselection based on the sum of squared differences. Once the best atlases were selected, for each voxel of the target image they discarded the corresponding atlas voxels that were not going to contribute in the label fusion based on their dissimilarity (luminance and contrast). The remaining voxels contributed to the weighting based on their intensity similarity. In a similar way, Wu et al. [155] recently presented a patch-based multi-atlas segmentation strategy in which they introduced a multi-scale image patch that combined both local and global

information. They proposed to dynamically adjust the patch size from large to small during the label fusion procedure, using the global image information to remove the misleading candidate atlas patches and then gradually using more local information to refine the label fusion result.

Recently, several brain structure segmentation methods using graph-based or tree-based intermediate templates guided registration have been presented, achieving effective segmentation results [140, 153]. This strategy is based on the principle that it is generally difficult to obtain accurate registration between images with large shape differences and thus, these methods try to decompose a large registration into smaller ones with the help of intermediate templates. Jia et al. [150] introduced a multi-atlas-based multi-image segmentation (MABMIS) framework to perform simultaneous segmentation of a group of target images based on the construction of a combinative tree. Similarly, Wolz et al. [147] presented a graph based framework called LEAP (Learning Embedding for Atlas Propagation), where the newly segmented images also became candidate atlases to segment the remaining target images. Pipitone et al. [154] proposed the MAGeT-Brain (Multiple Automatically Generated Templates) algorithm, that performed multi-atlas segmentation using a template library built from a subset of target images, constructed via label propagation with each of the available atlases. The authors stated that MAGeT minimized the number of atlases needed whilst still achieving similar agreement to conventional multi-atlas approaches.

2.3.2 Learning-based methods

The goal of learning-based methods or machine learning strategies is to predict the segmentation label S given the input features I. From a probabilistic perspective, the goal is to find the conditional distribution $p(S|I)$, that can be either learned from a training set of labeled images, in which case we have a discriminative approach, or alternatively, the joint distribution $p(I, S)$ can be found and used to evaluate the conditional $p(S|I)$, where the approach is known as generative [209]. These two approaches differ in that discriminative models provide a model only for the target variables conditional on the observed variables, whereas generative models are full probabilistic models of all variables. In what follows, we present different strategies based on such both approaches which we have called supervised and Bayesian methods.

A summary of the reviewed learning-based algorithms can be found in Table 2.3.

Supervised methods

Supervised methods, also known as discriminative models, attempt to directly estimate a label for each voxel given the local appearance of the image around it. For this purpose, these methods extract image features with rich information and use them to train a classification model using supervised learning algorithms. Among these learning algorithms we can find Artificial Neural Networks (ANNs) such as in [156], where a segmentation framework that can be applied to extract various brain structures, composed by two architectures that act in two phases, was presented. The first network classified the textures of the target image while the second one took the output of the first classifier and refined the segmentation correcting possible errors of the initial stage via local shape/texture analysis. Moreover in [165], the authors proposed a two-stage method that combined ANNs with Geometric Moment Invariants (GMIs). At the first stage, a set of multiple Multi-Layer Perceptron (MLP) networks were used for function approximation. There was one MLP for each scale of the GMIs, whose outputs, together with voxel intensities and coordinates, were the input features of the ANN of the second stage. At that stage, the ANN worked as a classifier instead of a function approximator, classifying each voxel as inside or outside the structure of interest.

In this category, deep learning strategies have become very popular in the last few years [210]. Several Convolutional Neural Network (CNN) architectures [158, 159, 160, 161, 162, 163, 211] have been proposed to segment the brain structures in MR images, which, as stated by Bernal et al. [210], can be classified according to: the number of interconnected operating modules, the input patch dimension, the number of predictions at a time or the implicit and explicit contextual information, among others. Based on the number of interconnected operating modules, the strategies in the literature can be divided into single-path and multi-path architectures, where the former corresponds to the cases in which there is a unique flow of information whereas in the second one, independent operative networks are integrated into a single model to capture a more varied set of features. In the single-path category, Wachinger et al. [162] presented an architecture in which a first CNN separated foreground from background on skull-stripped images, whereas a second CNN identified 25 brain structures on the foreground. On the other hand, Brébisson and Montana [211] trained eight networks arranged in parallel, with each network having a seven-layer CNN architecture for whole brain segmentation; Bao and Chung [158] proposed a multi-scale structured CNN architecture with three networks, with different $2D$ input patch sizes around the same pixel, that were arranged in parallel to segment the sub-cortical structures; and Mehta et al. [159], implemented a four network architecture arranged also in parallel, that included local intensity profile and global information in the form of $2.5D$ and $3D$ patches of different sizes, able to parcellate the whole brain.

In the case of the number of predictions the approach performs at a time, we can find CNN and fully CNN (FCNN) architectures. The first one corresponds to the traditional approach in which a single patch is processed by the network, and a single output is returned, whereas the second corresponds to the architectures for which the fully connected layers are replaced by 1×1-kernel (or $1 \times 1 \times 1$-kernel) convolutional layers to obtain a dense prediction. In this second category, Dolz et al. [161] implemented a $3D$ FCNN architecture, that apart from the FCNN implementation, considered multi-resolution information by extracting feature maps from high-resolution layers and merging them with low-resolution information, coming from the main information flow, before the $1 \times 1 \times 1$-kernel convolutional layers.

As can be noticed from the cited works, according to the input patch dimension, the application of $2D$ [160], $3D$ [161, 162], $2.5D$ [163] (patches from the three orthogonal views of an MRI volume), and their combinations [159, 211] including multi-scale patches [158] can be found in the literature.

Finally, apart from implicit information that is provided by the extracted patches from MRI volumes, explicit characteristics distinguishing spatial consistency have been studied. Wachinger et al. [162] introduced explicit within-brain location information through Cartesian and spectral coordinates to provide a distinctive perception of spatial location for every voxel, Brébisson and Montana [211] included distances to centroids to their networks, whereas Kushibar et al. [163] included prior spatial features that came in the form of atlas probabilities.

Besides these, other supervised strategies have been also proposed to segment the brain structures, such as the one by Morra et al. [164], who introduced a learning approach in which they iteratively learned the marginal distribution for each image voxel towards the final segmentation. The classifiers were trained not only on the features from the image patch, but also on the probability patch, and hence, their AdaBoost weak learners were decision stumps on both image and probability maps. In that context, they used the previously trained classifier to compute new classification maps that were used to train the next classifier, repeating this procedure until convergence. More recently, Tu and Bai [167] presented the Auto-Context algorithm, which was also based on this principle. Following this framework, Kim et al. [38] proposed a method to extract the hierarchical feature representation of image patches, from 7.0 T MRI images, based on deep learning. These features were further incorporated into a multi-atlas version of the Auto-Context segmentation framework to improve hippocampus segmentation on such high-resolution images. Moreover, Wels et al. [166] presented a method for segmenting brain sub-cortical structures based on the concept of marginal space learning. At each level of abstraction they built a discriminative model from a labeled set of images, training a probabilistic boosting tree from high-dimensional vectors of Haar and steerable features derived from the image intensities. These models were used to narrow the

range of possible solutions until the final shape can be inferred.

Support Vector Machines (SVM) have also been used as a learning strategy to perform brain structure segmentation. In [77], Igual et al. proposed a method for internal caudate nucleus segmentation, which first delineated the external boundary of the structure by means of the previously proposed algorithm, CaudateCut [212]. After an automatic geometric criterion classification, a SVM classifier based on shape features of the caudate regions was used, to separate head and body caudate regions. Finally a post-processing step based on a decision stump to improve the global classification was applied.

Other learning approaches relying on dictionary learning [41, 148] or Genetic Algorithms (GAs) [45, 157] have also emerged in recent years. Tong et al. [148] introduced a segmentation strategy based on the minimization of patch reconstruction errors, in which they learned a dictionary and a linear classifier simultaneously for every voxel of the target image from predefined neighborhood patches around that voxel and across the training atlases. Benkarim et al. [41] extended this method in their proposed multi-class dictionary learning approach. They learned a single discriminative dictionary and a multi-class linear classifier simultaneously for each target voxel, which allowed segmenting multiple structures at the same time. On the other hand, Deoni et al. [157] presented a method to segment fifteen thalamus internal structures by means of a GA approach that incorporated characteristics of the k-means clustering method. Regarding the fitness function, it gave greater values to those candidate segmentation solutions with maximized cluster (structure) sizes while the variance of the T1 and T2 image modalities was minimized. More recently, Traynor et al. [45] re-evaluated this method under much broader operating conditions.

Bayesian methods

Probabilistic segmentation methods try to infer the most likely segmentation given the observed image, which, according to the Bayes rule, can be approximated to the probability of the image occurring given a certain segmentation $p(I|S)$, together with the prior probability of the segmentation $p(S)$. This can be achieved via Maximum A Posteriori (MAP) estimation. $p(S)$, henceforth the prior, encodes the spatial organization of anatomical structures in the image domain, whereas $p(I|S)$, is a likelihood distribution that predicts how a label image, where each voxel is assigned a unique anatomical label, translates into an intensity image.

There is a large amount of work that relies on this general framework, differing mainly in the way the priors and the likelihood are specified as well as in the optimization method chosen to estimate the model parameters. In brain structure segmentation, priors come very often in the form of probabilistic atlases while

likelihood is commonly modeled as a Mixture of Gaussians (MoG), where the parameters (mean and variance) are usually estimated by means of the EM algorithm. In [175, 177] the prior was a mesh-based probabilistic atlas where the mesh deformation was estimated in a coordinated ascent scheme with the Levenberg-Marquardt algorithm in combination with the Gaussian parameters (EM). On the other hand, Makropoulos et al. [178] obtained their priors from several atlases, whose labels were propagated to the target image and averaged to form, in combination with tissue probability maps, a probabilistic spatial prior for each structure, while their likelihood term was approximated by a MoG. Following this framework, Riklin-Raviv et al. [168] introduced a method for group-wise segmentation of brain structures that avoided the use of statistical atlas by introducing latent atlases, generated from an image ensemble. They proposed to alternate between estimating the MAP segmentations and refining the model parameters, replacing the expectation step by a gradient descent process using a probabilistic level-set formulation.

Markov Random Field (MRF) modelization, that introduces local spatial dependencies between voxels, has also been broadly used among probabilistic methods. Fischl et al. [43] presented a method to segment the whole brain which forms the basis for the well-known software FreeSurfer [213]. In this method the intensity distribution of each structure at each location was modeled as a Gaussian, while the priors, that came in the form of global spatial information given by an atlas and local spatial relationship between anatomical classes, were approximated by an anisotropic non-stationary MRF. Another approach, proposed by Scherrer et al. [171], is the LOcal Cooperative Unified Segmentation (LOCUS) algorithm, that performed tissue and sub-cortical structure segmentation, which cooperate gradually to improve the accuracy. They performed the segmentation partitioning the target image into a set of local sub-volumes and distributed one local MRF per sub-volume. *A priori* knowledge in the form of generic fuzzy spatial relations was introduced in the MRF model to segment the structures. More recently, Razlighi et al. [176] introduced a segmentation method where they used quadrilateral MRF to model both the priors and the likelihood probabilities. In such a model, not only the neighborhood labels were taken into account, but also their intensities, by contrast to the classical MRF model.

Other methods based on this probabilistic principle have been proposed in recent years. Askelrod-Ballin et al. [173] presented a multi-scale algorithm that used a graph representation of the target image which was recursively coarsened to obtain the final segmentation. The posterior probability that two nodes were aggregated was estimated by means of Bayesian formulation, where the priors were given in a form of probabilistic atlas, while the likelihood was estimated from a set of manually labeled atlases. On the other hand, Pohl et al. [66] proposed a hierarchical algorithm guided by a prior information represented within a tree structure. They followed a recursive segmentation process that started at the root, segmenting the image

into its children, using the propagated structure-specific information in a classical Bayesian framework and estimating the solution through EM.

Recently, Iglesias et al. [177] presented a Bayesian segmentation framework that extended the work of Van Leemput et al. [175] based on the statement that traditional Bayesian methods do not fully consider the uncertainty in the model parameters, relying just on point estimates. To overcome this issue, they proposed a Monte Carlo sampling to account for the uncertainty in prior and likelihood free parameters which resulted in an improved approximation of the segmentation posterior.

2.3.3 Deformable methods

Deformable methods start with an initial contour placed in the image, either manually or automatically, which is then iteratively deformed, generating a new contour at each iteration. In the primitive version of a deformable model, known as snakes or Active Contour Models (ACM) [214], the initial contour was deformed under the influence of internal and external forces. The internal forces were related to the surface features and aim to maintain a smooth contour, while the external forces were related to the image features of the adjacent regions to the surface and were responsible for attracting the model towards the structure surface. Ghanei et al. [179] presented an improved discrete deformable model to segment the hippocampus, that addressed some of the associated problems such as optimizing the internal force weight, contour stability and extraction of image features for external energy calculations. They introduced a new external force, which was based on searching for local minima of the image energy in the contour normal direction, to produce better results near multiple and discontinuous edges. Colliot et al. [183] also presented a framework in which they use ACM. They added spatial relations between the different structures (direction and distance) represented as fuzzy subsets which were integrated in the deformable model as a new external force that attracted the model to the edges of the structure being segmented. Fouquier et al. [187] extended that work proposing a criteria to optimize the structure segmentation order and introducing a strategy to evaluate the obtained segmentation quality and detect errors to prevent their propagation. Zarpalas et al. [184] also presented a method for segmenting several structures by means of a mixture of different ACMs, all balanced by the gradient distribution boundaries which tried to differentiate regions that need greater support of prior knowledge from those which can be segmented only by their gray-scale information. They used geodesic ACM on boundary parts with strong gradients and a Chan-Vese model with prior knowledge in parts where the boundary was not well formed or contained weak parts. On the other hand, Shariatpanahi et al. [182] used m-Rep (medial representation) deformable models in a multi-agent framework to segment several brain structures, namely, the thalamus, the caudate

nucleus and the putamen.

An evolution of ACM are the Active Shape Models (ASM) [215], in which the internal energy, besides keeping the contour smooth, avoids also deformations that go beyond the average geometric variation of the structure being segmented, by using shape constraints learned from a collection of training samples. A Point Distribution Model (PDM) is used to build a shape model of the structure of interest, in which shapes are represented by a set of points or landmarks. Kelemen et al. [180] presented a framework that closely followed the seminal work of Cootes and Taylor [216] on ASM, but based on a hierarchical parametric object description rather than a PDM, under the statement that for a large training set containing several anatomical structures, the generation of the PDM parameterization became very tedious and could be a source of errors. Ettaïeb et al. [190] also proposed a segmentation model based on the ASM and a spatial distance relation. At the segmentation stage, the contours evolved iteratively in two steps: they first evolved independently of each other, according to the constraints imposed by the corresponding shape models and then, applying the constraint imposed by the statistical distance model, which was also estimated during the training phase. In [185], the authors introduced an approach referred to as a active hierarchical shape model, that was able to characterize the different inter-structure relationships and to model the particular local variations of each single structure. Gao et al. [186] also proposed a multi-scale representation for the shape using the wavelet transform. Given the initial shape obtained from label fusion, they proposed a segmentation method that alternated data-driven and a multi-scale shape-based process, iteratively evolving the contour until convergence. Alternatively, Olveres et al. [121] evaluated several ways to segment the midbrain including the combination of ASM with LBP (Local Binary Pattern) descriptors and compared them with the classical ASM segmentation, concluding that incorporating information about the texture surrounding the edge (with LBP), the segmentation performance increased, converging also in less iterations.

Another commonly used version of deformable model is the Active Appearance Model (AAM) [217], which incorporates constraints of image intensity variation to the ASM. This appearance model is usually based on the normalized first derivative of fixed-size gray profiles, normal to the surface of the object and centered at each landmark. In their work, Babalola et al. [39] presented an AAM based method, called profile appearance models, that instead of modeling the intensities across an entire region containing the model, as in the original AAM approach, they only modeled the intensities along the profiles that were normal to the boundary of the structure. Duchesne et al. [130] also reformulated the original AAM approach, that initially was not suitable for 3D images. Instead of using the original PDM to characterize the shape, which is impractical in 3D, they proposed to utilize a warp distribution model, that centered on 3D deformation fields from ANIMAL

[64]. On the other hand, Patenaude et al. [40] utilized the principles of the AAM but placed them within a Bayesian framework. The so called Bayesian Appearance Model (BAM) incorporated both shape and intensity information from a training set but, opposed to the original AAM, it used a probabilistic framework to estimate the relationship between shape and intensity making use of conditional probabilities. Based on the learned model, BAM searched the linear combination for the most probable shape given the observed image intensities to find the best fit for the segmentation. This model was implemented as part of the FSL[1] package under the name FIRST.

Other energy minimizing strategies have been proposed in recent years. Pitiot et al. [81] presented a segmentation method that relied on the deformable templates framework which incorporated available *a priori* anatomical expertise, either in the form of implicit knowledge (structure shape and appearance) or of explicit information (relative distance between structures, non-intersection rules) as constraints of the model. Al-Shaikhli et al. [189], on the other hand, proposed an approach for multi-region segmentation using a multi-level set formulation which included a topological graph prior and topological information of an atlas. This topological representation was embedded in the multi-level set energy equation and together with a curvature term, constrained the curve evolution.

Deformable models have to be initialized, either manually or automatically. Atlas registration is a typical practise, but other techniques are also used. In [181] the results of a classifier provided a localization point for the initiation of a deformable template model, whereas García et al. [188] proposed to use a Chan-Vese model to initialize the structure contour.

A summary of the deformable methods described here is presented in Table 2.4.

2.3.4 Region-based methods

Region-based brain structure segmentation methods have also been proposed in recent years. These methods rely on the similarity of different properties of the voxels belonging to the same region. Probably, one of the most well established region-based techniques is region growing, which is the most frequently used in brain structure segmentation. Based on this technique, Xue et al. [192] proposed a method that performed region-wise labeling by means of GAs followed by voxel-wise refinement using parallel region growing. They first over-segmented the target image into three brain tissues (WM, GM and CSF) and also got a coarse location of the structures by registering an atlas to the image. Afterwards, they built a fuzzy model of the regions of interest that represented useful structural knowledge from

[1] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

the atlas (shape, distance and relationship of structures), which was then used to design the objective function of GAs and to guide the region growing. Xia et al. [125] also presented an algorithm to segment the caudate nucleus that performed region growing constrained by anatomical knowledge. They first identified the lateral ventricles, which are easily locatable due to the high tissue contrast of the CSF, and based on their position they determined an initial caudate location by applying region growing from GM voxels adjacent to them. Bounding boxes were defined to reduce potential region growing leakage to other structures. After obtaining this coarse segmentation, caudate boundaries were fine-tuned to be smooth, recognizable and valid, based on anatomical knowledge.

However, there are other strategies within region-based techniques such as successive erosion and dilation operations or the use of the watershed algorithms. For instance, Gui et al. [193] proposed an approach to segment neonatal brain based on the use of general knowledge of neonatal brain morphology, integrating information about tissue connectivity, structure and relative positions. They performed a sequential segmentation of the brain structures that combined well-established segmentation methods (marker-based and similarity-based watershed, region growing and region-based active contours), guided by anatomical knowledge, with morphological operations (openings/closings).

The region-based works reviewed here are summarized in Table 2.4.

2.3.5 Hybrid methods

Several combinations of the previous categories have also been described in the literature. These methods try to combine the strengths of the different strategies in order to improve the segmentation accuracy. A common example is the combination of label fusion and learning-based strategies. For instance, in [138] the authors presented a probabilistic framework that lead to label fusion style segmentation algorithms. Under the assumption that each voxel of the target image is generated from one of the atlases, they constructed the conditional probability of generating the target image and label map where the final segmentation was achieved via MAP estimation. Recently, Iglesias et al. [198, 200] extended this framework to multimodal data. Another approach combining these strategies is the one proposed by Weisenfeld and Warfield [197] in which a multi-classifier fusion algorithm called Learning Likelihoods for Labeling (L3), which combined label fusion and statistical classification, was introduced. They employed each atlas from the training set to train a classifier that was used to generate a classification of the target image. These resulting classifications, generated by a Bayesian segmentation strategy, were later fused with the STAPLE algorithm to produce the final segmentation. Hao et al. [67] also proposed a learning-based label fusion method to segment the hippocampus.

In their approach, for each voxel in the target image, candidate training samples were obtained from voxels of atlases within a spatial neighborhood of the voxel considered, and the image feature vectors were then computed. Once the image features were extracted, a k-NN strategy based SVM classification algorithm was adopted to build a classifier for each voxel, which was applied to each voxel feature vector of the target image to obtain the final segmentation.

Combining discriminative and generative models is also common practice. Van der Lijn et al. [199] presented a segmentation method that combined structures' spatial and appearance information in a posterior probability function which was maximized using graph cuts. The spatial information came in the form of a probability map and the structure appearance was described by a k-NN voxel classifier based on Gaussian scale-space features. Liu et al. [170] proposed a hybrid method that combined a generative and a discriminative model, with feature augmentation and adaptation. Their approach was based on using the estimated segmentation and the parameters of a Bayesian segmentation to normalize the image intensities and extract robust, invariant local features. Afterwards, they used the auto-context algorithm [167] to obtain the final segmentation. Tu et al. [194] also introduced a hybrid model that combined both a discriminative approach to model the appearance and a generative model to describe shape. For appearance modeling they adopted a probabilistic boosting tree framework to learn a multi-class discriminative model while shape information was incorporated through principal component analysis (PCA). Once the system was trained, structure segmentation was obtained performing surface evolution by minimizing an energy function associated with the proposed hybrid model.

Other combinations have also been proposed, such as the ones in [195] and [196] that presented a hybrid method combining both region-based and boundary-based procedures. In these approaches they first applied a clustering technique (k-Means) to generate an initial seed contour and afterwards, the seed was deformed based on a level-set PDE. On the other hand, Zhou and Rajapakse [70] proposed a segmentation method based on fuzzy templates. From a set of labeled samples they obtained three fuzzy maps (intensity, spatial location and relative spatial relations to other structures) based on information obtained from structure histograms. These fuzzy maps were calculated for each structure and in each training image and were fused to obtain a total fuzzy template involving all features for different structures. This total fuzzy template was then registered to the target image, and combining its information with a probabilistic tissue segmentation [218], a fuzzy membership map of each structure was created. Final segmentation was performed by applying alpha-cut thresholding to this final fuzzy map.

Table 2.4 includes an overview of the segmentation targets and evaluation criteria of the methods summarized in this section.

Table 2.5: Pros & Cons. Summary of the advantages and disadvantages of the segmentation strategies reviewed.

	Method	Advantages	Disadvantages
Atlas-based	Label propagation	-Quite fast -Only one atlas is needed -Useful as initialization	-Dependent on the atlas anatomical similarity -Dependent on the registration method
	Label fusion	-Anatomical variation is better captured -Registration error is less likely to affect the segmentation	-Computationally expensive -Usually rely on registration accuracy -Prior segmented images (atlases) needed
Learning-based	Supervised methods	-Good results if test/training data variations are small -Can accurately capture local appearance variations	-Prior segmented images needed (training) -Not easily adapted to capture global shape information -Changes in MRI contrast reduce their performance -Highly dependent on the training set
	Bayesian methods	-Require small amounts of training data -Allows explicit incorporation of prior information -Robust against image artifacts -Flexibility and adaptability	-Can be difficult to define and learn -Can be slow
Deformable	ACM	-No training required -Minimized number of parameters -Perform an accurate adjustment that registration cannot	-Sensitive to initialization -Low contrast boundaries can cause them to fail -Initialization required
	ASM & AAM	-Robust against noise -Avoids deformations that go beyond the average -Handles discontinuities along structure boundaries -Perform an accurate adjustment that registration cannot	-Construction of an explicit model required (training) -Prior segmented images needed -Less flexible -Larger number of parameters
Region-based		-Quite fast -Do not need training data	-Low contrast boundaries can cause them to fail -Additional information to guide the growing is needed -Initialization required

Hybrid methods are not included in the table since they combine the strategies presented here, with their corresponding strengths and weaknesses.

2.4 Pros and cons of the strategies

In Section 2.3 we have presented five main strategies to perform brain structure segmentation, namely atlas-based, learning-based, deformable, region-based and hybrid methods. The main advantages and drawbacks of these strategies are presented here and summarized in Table 2.5.

Atlas-based methods insert robustness to the segmentation strategy, as they overcome the deficiencies of contrast and MRI resolution. Label propagation is the most straightforward atlas-based technique, based on the propagation of the atlas labels to the target image after registration. It is a good technique when only one atlas is available and is also quite fast, but it suffers from two main drawbacks. The first one is the fact that a simple atlas cannot sufficiently represent the whole population of potential test data and the second is that the quality of this approach is limited by the accuracy of the pairwise registration method. This technique is commonly used as a starting point of other methods, which use the propagated labels as a initialization or as a prior.

In order to overcome the problems presented in label propagation, multi-atlas or label fusion methods have emerged in recent years. Several atlases are used to improve the capture of anatomical variability between different scans but at the expense of a high computational time. These methods are less dependent on the

registration as the effect of errors associated with any single atlas propagation is reduced in the combination process.

Supervised methods provide good results when the differences between the training data and the target image are small, capturing accurately local appearance variations. However, they are very dependent on the training set, hence changes in MRI contrast or strong anatomical differences among training and testing images highly reduce their performance, limiting their applicability to images acquired with the same protocol as the images used for training. As such, larger training data sets trend to be beneficial. On the other hand, Bayesian approaches are more flexible and adaptable, as they permit explicitly modeling image artifacts such as the bias field or other image acquisition parameters, making these methods more robust. They also allow the explicit incorporation of *a priori* information by means of the prior term, that frequently comes in the form of a probabilistic atlas, which captures global shape information. However, these methods can be slow and difficult to design and learn, especially for complex structures with inhomogeneous textures.

Deformable methods introduce dynamics to the segmentation (internal and external forces), performing an accurate adjustment that the registration methods cannot perform, but they require robust initialization. In the case of ACM, as they represent a local search, this initialization must be done near the structure of interest in order to avoid falling in a local minima. Furthermore, low contrast boundaries may prevent them to provide the desired solution. By contrast to ACMs, ASMs and AAMs need to be trained to learn an explicit model (shape or appearance constraints) that captures the variation of shape and gray level across a training set. This model guides the contour evolution and avoids deformations that go beyond the average geometric variation of the structure being segmented, making these methods robust against noise, but at the same time less flexible and with a larger number of parameters/constraints than ACMs.

Contrary to these methods, region-based strategies do not require to be trained and are usually quite fast in finding a solution. However, these methods must usually be combined with some other kind of anatomical knowledge and require an initial seed to initialize the growing. Furthermore, and similarly to ACMs, noisy structure boundaries can easily cause them to fail.

Tables 2.2 to 2.4 summarize the approaches reviewed in Section 2.3 and show, for each work, the brain structures segmented as well as the metrics and databases used for evaluation indicating also if they were tested with normal controls or diseased subjects. As shown in these tables, atlas-based methods are the most commonly used to segment the whole brain whereas they are rarely used to segment a single structure. They use to register the atlas/es to the whole volume, which is the hardest and more computationally expensive task, and once it is done, performing label propagation and fusion is as trivial for one label as for all of them. Even among the

hybrid methods, the ones which segment the whole brain are all based on combinations of atlas-based approaches with another segmentation strategy. According to these tables, region-based strategies are the less popular in segmenting brain structures. As stated before, these methods on its own are not sufficient in most of the cases and usually need additional anatomical information or to be combined with another segmentation strategy, which makes them not very attractive. Furthermore, it can be stated that there are not so many methods requiring a training phase, such as supervised, ASM and AAM, which segment the whole brain. This is due to the fact that these trainings use to be task specific instead of generalistic, with the aim of segmenting a single structure or a reduced group of them.

From these tables, we observe that the vast majority of research evaluate their algorithms with non-lesioned brain databases. As an exception, Fouquier et al. [187] used a database composed by 30 healthy cases and 14 pathological (brain tumor) cases. However, they only provide quantitative segmentation results for the healthy cases, which does not provide any information of how these lesions affect the structure segmentation.

To the best of our knowledge, how WM lesions, such as the ones produced in MS or lupus, affect these algorithms has not yet been evaluated. Nevertheless, it is a well known problem among automatic tissue segmentation methods [219], where lesion filling techniques [29, 30, 31] have already been applied improving the accuracy of tissue volume [32, 220]. As far as we know, these techniques have not yet been evaluated in combination with automatic brain structure segmentation algorithms and making these algorithms robust against lesions remains still an open challenge to the research community.

Automatic segmentation of severely atrophied brains, which produces morphological changes in the brain structures, can also make some methods performance, such as the ones relying on registration, oscillate. This could be prevented by means of large atlas databases with high anatomical variability reducing the dissimilarity between the atlases and the target image. However, there are not many public databases with ground truth and furthermore most of them only contain healthy subjects, which makes this solution hard to accomplish.

In order to give an overall quantitative performance estimate of the state-of-the-art methods, in the following section we present an evaluation of the methods reviewed here, grouping the results both per segmentation strategies and target structures and reporting the different evaluation metrics and databases used. Furthermore, the existing challenges in the evaluation of automatic brain structure segmentation methods are discussed and a quantitative evaluation of three well known software tools, each relying on a different category of our classification, is performed.

Table 2.6: Most commonly used public databases for evaluation in brain structure segmentation.

Name	Subjects	Ages	Modality	Structures	Scanner	Volume (mm)	Voxel (mm)
IBSR18[2]	18 (14 ♂, 4 ♀)	7 - 71	T1-w	43	12: GE (1.5T) 6: Siemens (1.5T)	256×256×128	8: 0.94×0.94×1.5 6: 0.84×0.84×1.5 4: 1×1×1.5
LPBA40[3]	40 (20 ♂, 20 ♀)	19.3 - 39.5	T1-w	56	GE Signa (1.5T)	256×256×124	38: 0.86×0.86×1.5 2: 0.78×0.78×1.5
Hammers[4]	30 (15 ♂, 15 ♀)	20 - 54	T1-w	83	GE Signa (1.5T)	192×256×124	0.937×0.937×1.5

Volume and voxel dimensions in native space.

2.5 Validation and quantitative evaluation

2.5.1 Public databases

Comparing the results of the available brain structure segmentation methods is not a trivial task. There are a few publicly available manually labeled images [221, 222, 223] that serve as a ground truth and some authors use their own images for evaluation, which makes quantitative comparison of the structure segmentation algorithms difficult. As we have noticed from the reviewed papers, the most frequently used public databases for evaluation include the Internet Brain Segmentation Repository[2] (IBSR18), the LONI Probabilistic Brain Atlas[3] (LPBA40) [224] and the Hammers Adult atlases[4] [225, 226, 227]. Table 2.6 shows the main features of these databases, that include the name and webpage, the number of images it contains, the image modalities, the number of structures labeled, the scanner used for acquisition, the image resolution, the voxel size and some demographics such as subjects' age and sex.

2.5.2 Evaluation metrics

Apart from the difficulty of having good datasets, there is not a standard metric for evaluation and results are presented using different metrics. Analyzing the literature we have seen that the most commonly used metrics are based either on volume overlap or contour distance. Metrics based on volume compute how much the ground truth and the obtained segmentation volume overlap. The most common metric in this category is the Dice similarity coefficient (DSC) and variants such as the Kappa index (KI), the Similarity index (SI), the relative overlap (RO) or the Jaccard index

[2]https://www.nitrc.org/projects/ibsr

[3]http://www.loni.usc.edu/atlases/Atlas_Detail.php?atlas_id=12

[4]http://brain-development.org/brain-atlases/

(JC), but other metrics are also used such as the specificity (SP), sensitivity (SN), accuracy (Acc), precision (P) and recall (R). Regarding contour distances, which rely on computing how close the ground truth and the obtained segmentation contours are, the most recurrent are the Hausdorff distance (HD) and its variants, and the mean absolute distance (MAD).

As an attempt to standardize this evaluation procedure, three MICCAI Challenges[5] on brain structure segmentation have been proposed during recent years. The first one was the CAUdate SEgmentation 2007 (CAUSE07)[6], which was a competition held as part of the workshop '3D Segmentation in the Clinic: A Grand Challenge' [228], in conjunction with MICCAI 2007. The goal of this competition was to compare different algorithms when segmenting the caudate nucleus from brain MRI scans, for which they provided 67 images (33 for training and 34 for testing) collected from different databases. As a comparative measure, they devised a scoring system that combined several metrics into a single overall score. The second one, was the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling[7] [229]. The challenge was on whole-brain labeling, assuming the majority of the participant methods to be multi-atlas but accepting any method as long as the approach were described in a reproducible manner. They provided 15 images for training and 20 for testing, obtained from the Open Access Series of Imaging Studies (OASIS) project[8] and the primary metric for evaluation was the mean DSC across all brain labels and all subjects in the testing cohort. Finally, the MICCAI 2013 Segmentation: Algorithms, Theory and Applications (SATA) challenge[9] [230], was created to test the limits of the applicability of multi-atlas segmentation. In this case, they proposed two sub-challenges: the free-for-all sub-challenge, which allowed any segmentation framework to be applied, and the standardized registration sub-challenge in which pairwise registrations were provided to remove the impact of the registration algorithm. A collection of 47 images was provided, 45 of which were from the OASIS project and the remaining two were part of the Child and Adolescent NeuroDevelopment Initiative (CANDI)[10], whereas the image labels, that included seven sub-cortical structures (accumbens, amygdala, caudate, hippocampus, pallidum, putamen and thalamus), were provided by Neuromorphometrics, Inc.[11] As an evaluation metric for strategies comparison, they used DSC and the symmetric HD.

[5] http://grand-challenge.org/
[6] http://cause07.grand-challenge.org/
[7] https://masi.vuse.vanderbilt.edu/workshop2012/index.php
[8] http://www.oasis-brains.org/
[9] https://masi.vuse.vanderbilt.edu/workshop2013/index.php
[10] http://www.nitrc.org/projects/candi_share
[11] http://www.neuromorphometrics.com/

2.5.3 Quantitative analysis of the reviewed literature

We present in what follows a quantitative comparison of the works reviewed in this chapter, separately for each brain structure, including: hippocampus, thalamus, caudate nucleus, putamen, pallidum, amygdala, accumbens, lateral ventricles, and brainstem. Table 2.7 summarizes the results, as well as the data and the evaluation metrics obtained from the analyzed approaches. Note that the results are provided with different metrics which have been obtained on different databases with different number of volumes. Ideally, a comparison of the methods should have been done using the same dataset and metrics, however, only a few of those methods share these properties. As shown in the table, the vast majority of the works reviewed here have been tested with real data, most of it from the public databases summarized in Sections 2.5.1 and 2.5.2. Nevertheless, some of them are either tested with private datasets or even with publicly available images for which segmentations have not been made public. As additional information, some of the works have publicly available software, even though a large majority have not, which makes them even more difficult to compare. The choice of the cited brain structures has been done because there were the ones for which a sufficient number of quantitative results to perform analysis was available. The results are shown as averages for each structure (left and right pair combined), except the brainstem which is a unique structure.

From the table we observe that the most commonly used evaluation metrics are those based on volume overlap, in particular DSC. For this reason, we perform a first analysis based on the works that use this metric for evaluation. Another reason of choosing an overlap metric is that the formula used is always the same, whereas the distance metrics are calculated differently in each of the works. In the case of the HD, for example, it can be symmetric or asymmetric and the results can be given in pixels, millimeters or even being not specified, thus the provided distance metric is not always comparable. Hence, to have an overall quantitative estimate of the state-of-the-art methods reviewed here, we focus on the works evaluated based on the DSC, regardless of the data used.

At first glance, we see that the structures on which more works have focused on are the hippocampus and the caudate nucleus, whereas the structures on which less attention has been paid are the brainstem and the nucleus accumbens. Furthermore, in view of the reviewed results, the nucleus accumbens seems to be the most challenging structure to segment, with a mean DSC of 0.69, whereas the structures that seem to get the best results in terms of volume overlap are the brainstem with a mean DSC of 0.88, closely followed by the thalamus (with mean DSC of 0.87) and the putamen (mean DSC of 0.86). The lower results for the accumbens are reasonable since it is the smallest structure, thus small errors in overlap give the highest changes in the DSC. On the other hand, the brainstem presents relatively strong contrast boundaries therefore its segmentation should be easier compared to

Table 2.7: Quantitative results of the reviewed methods. Acronyms from left to right are: hippocampus (HIP), thalamus (THA), caudate nucleus (CAU), putamen (PUT), pallidum (PAL), amygdala (AMY), accumbens (ACC), lateral ventricle (LV) and brainstem (BS). Metrics in order of appearance: Kappa index (KI), Jaccard index (JC), Relative overlap (RO), Dice similarity coefficient (DSC), Similarity index (SI), Mean absolute distance (MAD), Hausdorff distance (HD), Classification rate (CR), Modified Hausdorff distance (MHD), Similarity (Sim), Hausdorff distance 95 (HD95) and Overlap (Over). Column 'Ref' shows the reference work from which the results are obtained (in case they do not come from the original work).

	Article	Ref	Segmented structures									Metric	Database
			HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS		
Label propagation	Collins (1997) [64]	[129]	0.86	-	-	-	-	0.82	-	-	-	KI	80v (ICBM)
		[129]	0.76	-	-	-	-	0.70	-	-	-	JC	80v (ICBM)
		[130]	0.71	-	-	-	-	0.65	-	-	-	KI	30v (ICBM)
		[130]	0.69	-	-	-	-	0.64	-	-	-	KI	10v (ICBM)
	Shen (2002) [131]	[132]	0.49	0.64	0.55	0.52	0.41	0.46	-	0.65	0.72	JC	11v (IBSR)
		[132]	0.62	0.74	0.65	0.72	0.60	0.61	-	0.58	0.79	JC	36v (Desikan et al.)
	Postelnicu (2009) [132]		0.45	0.60	0.53	0.48	0.23	0.41	-	0.67	0.77	JC	11v (IBSR)
			0.63	0.75	0.67	0.75	0.60	0.57	-	0.66	0.73	JC	36v (Desikan et al.)
	Lin (2010) [71]		-	0.81	0.73	0.78	-	-	-	-	-	KI	15v (IBSR)
			-	0.68	0.57	0.64	-	-	-	-	-	RO	15v (IBSR)
	Luo (2011) [135]		-	0.84	0.78	0.80	-	-	-	-	-	DSC	9v (IBSR)
Label fusion	Warfield (2004) [65]	[138]	0.81	0.89	0.83	0.88	0.82	0.79	-	0.86	-	DSC	39v (FS atlas)
		[139]	0.52	0.85	0.68	0.80	0.70	0.65	0.49	0.51	0.85	SI	17v (IBSR)
		[139]	3.73	0.96	1.64	1.04	1.36	1.33	3.74	3.08	1.08	MAD	17v (IBSR)
	Heckemann (2006) [140]		0.82	0.91	0.90	0.90	0.80	0.81	0.71	0.90	0.94	SI	30v
	Aljabar (2009) [142]		0.83	0.91	0.88	0.90	0.82	0.78	0.76	0.91	0.94	DSC	275v (CMA)
	Artaechevarría (2009) [139]		0.75	0.88	0.83	0.86	0.79	0.72	0.67	0.83	0.91	SI	17v (IBSR)
			0.79	0.75	0.64	0.67	0.72	0.85	0.68	0.69	0.69	MAD	17v (IBSR)
		[143]	0.84	0.88	0.88	0.89	0.77	0.80	0.69	-	-	DSC	30v (Hammers)
		[143]	0.87	-	-	-	-	-	-	-	-	DSC	30v (ADNI)
		[144]	0.75	0.88	0.83	0.86	0.79	0.72	0.67	0.83	0.91	DSC	18v (IBSR)
	Lötjönen (2010) [145]		0.81	0.90	0.87	0.91	0.84	0.77	-	-	-	SI	18v (IBSR)
			0.88	-	-	-	-	-	-	-	-	SI	60v (ADNI)
		[144]	0.80	0.89	0.85	0.90	0.83	0.75	-	-	-	DSC	18v (IBSR)
	Collins (2010) [129]		0.89	-	-	-	-	0.83	-	-	-	KI	80v (ICBM)
			0.80	-	-	-	-	0.70	-	-	-	JC	80v (ICBM)
	Heckemann (2010) [146]		0.71	0.80	0.81	0.81	0.63	0.65	0.52	0.83	0.88	JC	30v (Hammers)
		[143]	0.83	0.89	0.89	0.89	0.77	0.79	0.68	-	-	DSC	30v (Hammers)
	Wolz (2010) [147]		0.85	-	-	-	-	-	-	-	-	DSC	796v (ADNI)
	Rousseau (2011) [144]		0.83	0.89	0.89	0.89	0.79	0.75	0.67	0.93	0.93	DSC	18v (IBSR)
	Coupé (2011) [42]		0.88	-	-	-	-	-	-	-	-	KI	80v (ICBM)
			-	-	-	-	-	-	-	0.96	-	KI	80v
		[148]	0.88	-	-	-	-	-	-	-	-	DSC	80v (ICBM)

Table 2.7 continued from previous page

	Article	Ref	Segmented structures									Metric	Database
			HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS		
		[148]	0.85	-	-	-	-	-	-	-	-	DSC	202v (ADNI)
	Jia (2012) [150]		-	-	-	-	-	-	-	0.91	-	DSC	50v (ADNI)
	Cardoso (2013) [143]		0.90	-	-	-	-	-	-	-	-	DSC	30v (ADNI)
			0.84	0.89	0.89	0.89	0.80	0.81	0.70	-	-	DSC	30v (Hammers)
	Wang (2013) [151]		0.87	0.92	0.88	0.91	-	0.82	0.80	-	0.95	DSC	20v (MICCAI'12)
	Wang (2014) [153]		0.80	0.89	0.75	0.88	0.84	0.76	0.71	0.84	0.90	DSC	20v (MICCAI'12)
			0.58	0.52	0.79	0.38	0.42	0.56	0.55	0.61	0.55	MAD	20v (MICCAI'12)
			5.95	4.10	5.36	3.12	2.74	3.65	4.00	8.76	6.85	HD	20v (MICCAI'12)
	Pipitone (2014) [154]		0.87	-	-	-	-	-	-	-	-	DSC	60v (ADNI1)
			0.89	-	-	-	-	-	-	-	-	DSC	81v (FEP)
	Wu (2015) [155]		0.85	0.90	0.90	0.89	0.80	0.82	0.71	-	-	DSC	30v (Hammers)
Supervised	Pitiot (2002) [156]		0.91	-	0.90	-	-	-	-	-	-	CR	10v
	Bao (2018) [158]		0.82	0.90	0.87	0.88	0.80	0.67	-	-	-	DSC	9v (IBSR)
			0.84	-	0.85	0.86	-	-	-	-	-	DSC	20v (LPBA40)
	Mehta (2017) [159]		0.82	0.89	0.87	0.91	0.82	0.74	-	0.93	-	DSC	9v (IBSR)
			0.83	-	0.83	0.84	-	-	-	-	-	DSC	20v (LPBA40)
			0.83	0.90	0.88	0.89	0.80	0.81	0.68	-	-	DSC	15v (Hammers)
	Shakeri (2016) [160]		-	0.87	0.78	0.83	0.75	-	-	-	-	DSC	18v (IBSR)
	Dolz (2018) [161]		-	0.92	0.91	0.90	0.83	-	-	-	-	DSC	18v (IBSR)
			-	0.13	0.21	0.18	0.26	-	-	-	-	MHD	18v (IBSR)
	Kushibar (2018) [163]		0.88	0.92	0.89	0.92	0.85	0.83	0.80	-	-	DSC	20v (MICCAI'12)
			4.12	3.35	3.42	2.69	2.49	2.56	2.47	-	-	HD	20v (MICCAI'12)
			0.85	0.91	0.90	0.90	0.83	0.77	0.75	-	-	DSC	18v (IBSR)
			4.15	7.21	4.10	4.90	3.77	4.31	3.01	-	-	HD	18v (IBSR)
	Morra (2008) [164]		0.70	-	-	-	-	-	-	-	-	RO	83v (AD)
			0.82	-	-	-	-	-	-	-	-	SI	83v (AD)
			4.15	-	-	-	-	-	-	-	-	HD	83v (AD)
	Moghaddam (2009) [165]		-	0.89	0.83	0.88	-	-	-	-	-	DSC	6v (IBSR)
			-	0.90	0.75	0.70	-	-	-	-	-	MAD	6v (IBSR)
			-	2.21	2.40	1.92	-	-	-	-	-	HD	6v (IBSR)
	Wels (2009) [166]		0.73	-	0.80	0.82	0.75	-	-	-	-	DSC	18v (IBSR)
			0.91	-	0.67	0.72	0.79	-	-	-	-	MAD	18v (IBSR)
			-	-	0.66	-	-	-	-	-	-	MAD	MICCAI'07
	Tong (2013) [148]		0.87	-	-	-	-	-	-	-	-	DSC	202v (ADNI)
			0.89	-	-	-	-	-	-	-	-	DSC	80v (ICBM)
		[41]	-	-	0.87	0.90	0.86	-	0.74	-	-	DSC	35v (MICCAI'12)
	Kim (2013) [38]		0.82	-	-	-	-	-	-	-	-	RO	20v (7T)
			0.89	-	-	-	-	-	-	-	-	SI	20v (7T)
	Benkarim (2014) [41]		-	-	0.87	0.91	0.87	-	0.76	-	-	DSC	35v (MICCAI'12)
Bayesian	Fischl (2002) [43]	[168]	0.84	0.88	0.85	0.85	0.80	0.75	-	-	-	DSC	39v (Sabuncu et al.)

Table 2.7 continued from previous page

	Article	Ref	Segmented structures									Metric	Database
			HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS		
		[41]	-	-	0.82	0.79	0.74	-	0.55	-	-	DSC	35v (MICCAI'12)
		[138]	0.85	0.88	0.85	0.84	0.79	0.80	-	0.88	-	DSC	39v (FS atlas)
		[169]	0.79	0.88	0.79	0.81	0.71	0.71	-	-	-	DSC	30v
		[144]	0.75	0.86	0.82	0.81	0.71	0.68	0.58	0.78	-	DSC	18v (IBSR)
	Scherrer (2007) [171]		-	0.80	0.76	0.79	-	-	-	-	-	DSC	BrainWeb
	Akselrod-Ballin (2007) [173]		0.69	0.84	0.80	0.79	0.74	0.63	-	-	0.84	DSC	18v (IBSR)
			1.88	1.44	1.44	1.60	2.43	1.67	-	-	1.62	MAD	18v (IBSR)
			4.57	2.90	3.07	3.36	3.75	3.38	-	-	3.42	HD	18v (IBSR)
		[144]	0.69	0.84	0.80	0.79	0.74	0.63	-	-	0.84	DSC	18v (IBSR)
	Pohl (2007) [66]		0.81	-	-	-	-	0.86	-	-	-	DSC	50v
	Scherrer (2009) [174]		-	0.72	0.83	0.77	-	-	-	-	-	DSC	BrainWeb
	Riklin-Raviv (2010) [168]		0.76	0.85	0.82	0.85	0.78	0.79	-	-	-	DSC	39v (Sabuncu et al.)
	Makropoulos (2014) [178]		0.79	0.90	0.85	-	-	0.83	-	0.84	0.92	DSC	20v (ALBERTs)
Deformable	Duchesne (2002) [130]		0.68	-	-	-	-	0.63	-	-	-	KI	30v (ICBM)
			0.67	-	-	-	-	0.61	-	-	-	KI	10v (ICBM)
	Pitiot (2004) [81]		3.00	-	2.00	-	-	-	-	2.60	-	HD	20v
			2.10	-	1.60	-	-	-	-	1.80	-	MAD	20v
	Shariatpanahi (2006) [182]		-	0.78	0.72	0.74	-	-	-	-	-	Sim	4v (IBSR)
			-	1.54	1.17	1.10	-	-	-	-	-	HD	4v (IBSR)
	Colliot (2006) [183]		-	-	2.20	-	-	-	-	-	-	HD	10v
			-	-	1.00	-	-	-	-	-	-	MAD	10v
			-	-	0.87	-	-	-	-	-	-	SI	10v
	Babalola (2007) [39]		-	-	0.73	-	-	-	-	-	-	JC	24v
	Patenaude (2011) [40]		0.80	0.87	0.84	0.89	0.78	0.73	0.72	-	0.86	DSC	37v (NC+SZ)
			0.83	0.87	0.84	0.87	0.78	0.77	0.71	-	0.86	DSC	42v (NC+AD)
			0.84	0.87	0.87	0.86	0.76	0.76	0.67	-	0.86	DSC	17v (NC+AD)
			0.80	0.85	0.85	0.88	0.72	0.73	0.73	-	0.81	DSC	87v (NC+SZ)
			0.80	0.87	0.84	0.86	0.76	0.74	0.67	-	0.85	DSC	14v (NC+PC)
			0.81	0.86	0.84	0.89	0.79	0.74	0.70	-	0.83	DSC	139v (NC+ADHC+SZ)
	Gao (2012) [186]		0.82	-	0.91	-	-	-	-	-	-	DSC	24v (HIP); 24v (CAU)
			3.32	-	2.36	-	-	-	-	-	-	HD95	24v (HIP); 24v (CAU)
	Fouquier (2012) [187]		-	2.13	3.16	3.25	-	-	-	-	-	MAD	30v (IBSR+OASIS)
Region	Xue (2000) [191]		-	0.94	0.91	0.95	-	-	-	0.98	-	SI	-
			-	0.92	0.90	0.93	-	-	-	0.96	-	KI	-
	Xue (2001) [192]		-	0.94	0.91	0.95	-	-	-	0.98	-	SI	-
			-	0.92	0.90	0.93	-	-	-	0.96	-	KI	-
	Xia (2007) [125]		-	-	0.87	-	-	-	-	-	-	Over	55v
	Gui (2012) [193]		-	-	-	-	-	-	-	-	0.90	DSC	10v (newborns)
Hybrid	Zhou (2005) [70]		0.71	0.84	0.81	0.83	-	0.65	-	-	-	Over	17v (IBSR)

Table 2.7 continued from previous page

Article	Ref	Segmented structures									Metric	Database
		HIP	THA	CAU	PUT	PAL	AMY	ACC	LV	BS		
		0.56	0.32	0.32	0.29	-	0.67	-	-	-	MAD	17v (IBSR)
Tu (2008) [194]		10.50	-	7.35	10.15	-	-	-	6.60	-	HD	14v
		2.40	-	1.45	2.50	-	-	-	1.10	-	MD	14v
Karsch (2009) [195]		0.70	-	-	-	-	-	-	0.80	-	DSC	-
		0.66	-	-	-	-	-	-	0.77	-	Over	-
Sabuncu (2009) [169]		0.81	0.84	0.84	0.89	0.83	0.80	-	-	-	DSC	30v
Sabuncu (2010) [138]		0.87	0.91	0.87	0.89	0.84	0.82	-	0.91	-	DSC	39v (FS atlas)
	[143]	0.82	0.89	0.89	0.87	0.77	0.78	0.67	-	-	DSC	30v (Hammers)
	[143]	0.87	-	-	-	-	-	-	-	-	DSC	30v (ADNI)
He (2011) [196]		-	0.70	-	-	-	-	-	0.80	-	DSC	20v+25v
		-	0.66	-	-	-	-	-	0.77	-	Over	20v+25v
van der Lijn (2012) [199]		0.87	-	-	-	-	-	-	-	-	DSC	18v (Set I)
		0.77	-	-	-	-	-	-	-	-	JC	18v (Set I)
		0.87	-	-	-	-	-	-	-	-	DSC	18v (Set II)
		0.76	-	-	-	-	-	-	-	-	JC	18v (Set II)
Iglesias (2013) [200]		0.80	0.88	0.85	0.89	0.83	0.70	-	0.81	-	DSC	8v (PD)
Liu (2013) [170]		0.78	0.89	0.84	0.87	0.81	0.73	-	0.81	-	DSC	6v (IBSR)
		0.83	-	0.81	0.84	-	-	-	-	-	DSC	15v (LPBA40)
		7.82	-	4.21	6.85	-	-	-	-	-	HD	15v (LPBA40)
		4.90	-	4.89	5.56	-	-	-	42.68	-	HD	28v (LONI28)
		1.25	-	0.91	0.96	-	-	-	0.72	-	MAD	28v (LONI28)
Hao (2014) [67]		0.89	-	-	-	-	-	-	-	-	DSC	30v (ADNI 1.5T)
		3.26	-	-	-	-	-	-	-	-	HD	30v (ADNI 1.5T)
		0.27	-	-	-	-	-	-	-	-	MAD	30v (ADNI 1.5T)
		0.91	-	-	-	-	-	-	-	-	DSC	30v (ADNI 3T)
		1.88	-	-	-	-	-	-	-	-	HD	30v (ADNI 3T)
		0.21	-	-	-	-	-	-	-	-	MAD	30v (ADNI 3T)
		0.91	-	-	-	-	-	-	-	-	DSC	57v (3T)
		2.91	-	-	-	-	-	-	-	-	HD	57v (3T)
		0.25	-	-	-	-	-	-	-	-	MAD	57v (3T)

the rest of the structures.

If we perform an analysis by structure, for hippocampus segmentation we can highlight the work of Hao et al. [67] which we have classified in the hybrid category. Their method has been evaluated with three different databases with a total number of 117 volumes, achieving DSC values ranging from 0.89 to 0.91. Moreover, the works of Cardoso et al. [143] and Pipitone et al. [154], both atlas-based strategies, which have been tested with 60 and 141 volumes respectively (two different databases each) obtained DSC values of $0.84 - 0.90$ and $0.87 - 0.89$. Finally, the learning-based approach proposed by Tong et al. [148], achieved DSC values of 0.87 and 0.89 in two different databases, with a total number of 282 cases.

Regarding the thalamus, some works stand out from the rest, such as the ones by Aljabar et al. [142] and by Wang and Yushkevich [151], which achieved DSC coefficients of 0.91 and 0.92 respectively, when tested with 275 and 20 volumes (both classified in the atlas-based category). Moreover, a remarkable work, due to its good performance on a large number of testing volumes, is the deformable strategy proposed by Patenaude et al. [40], which was tested with 6 different databases with a total number of 336 volumes and obtained DSC values between 0.85 and 0.87. Moreover, the hybrid approach presented by Sabuncu et al. [138] achieved a DSC of 0.89 and 0.91 in two different datasets with a total of 69 testing images.

In caudate nucleus segmentation, the work of Gao et al. [186], classified as a deformable strategy, obtained a DSC of 0.91 when tested with 24 cases. Moreover, the atlas-based approach presented by Wu et al. [155] and tested with 30 cases, achieved a DSC of 0.90. Finally, the hybrid proposal of Sabuncu et al. [138] obtained DSC values of 0.87 and 0.89 with a total of 69 cases.

Segmenting the putamen, some atlas-based strategies have provided good results such as the ones by Aljabar et al. [142], Lötjönen et al. [145] and Wang and Yushkevich [151] that achieved DSC values of 0.90, 0.90 and 0.91 respectively when tested with 275, 18 and 20 volumes. Some learning-based strategies have also obtained good results in putamen and globus pallidum segmentation, which include the approaches of Tong et al. [148] and Benkarim et al. [41], that obtained DSC values for putamen segmentation of 0.90 and 0.91 respectively, and 0.86 and 0.87 for pallidum segmentation, both with a database composed of 35 images. Finally, the deformable approach proposed by Patenaude et al. [40] is also remarkable for the large amount of data used for testing (336v). When segmenting the putamen, the authors achieved DSC values ranging from 0.86 to 0.89.

For amygdala segmentation two learning-based approaches over-perform the others, which are the ones proposed by Pohl et al. [66] and by Makropoulos et al. [178] which achieved DSC values of 0.86 and 0.83 respectively, tested with 50 and 20 cases. In the category of atlas-based strategies we can remark the works of Wang

and Yushkevich [151] and Wu et al. [155], both obtaining a DSC of 0.82 when tested with 20 and 30 volumes respectively.

With regard to nucleus accumbens segmentation, some notable works are those presented by Aljabar et al. [142], Wang and Yushkevich [151] and Benkarim et al. [41]. The first two are atlas-based strategies, while the third is a learning-based approach. The DSC values achieved when segmenting this structure are 0.76, 0.80 and 0.76, respectively (275, 20 and 35 testing volumes).

In segmenting the lateral ventricles and the brainstem, atlas-based approaches provide good results. The work presented by Rousseau et al. [144] obtained a DSC of 0.93 for both structures with a testing dataset of 18 volumes. Furthermore, Aljabar et al. [142] and Jia et al. [150] achieved both a DSC value of 0.91 when segmenting the lateral ventricles with 275 and 50 volumes respectively, whereas Wang and Yushkevich [151] obtained a DSC of 0.95 in brainstem segmentation with a testing cohort of 20 cases. It has to be highlighted that in the other categories there are not too many works that segment these particular structures, however the hybrid approach proposed by Sabuncu et al. [138] achieved good results when segmenting the lateral ventricles (DSC of 0.91 with 39 testing volumes).

If we perform a second analysis based only on the works that use the same database and metric (DSC) for evaluation, we can state that for the IBSR18 database, the approach that seems to perform best in terms of volume overlap for almost all the structures is the one presented by Rousseau et al. [144], which is a patch-based label fusion strategy and therefore classified in the atlas-based category. On the other hand, if we look only at the works that use the Hammers adult atlases database for evaluation, it seems that the strategy that achieved the highest DSC for all the evaluated structures is the one proposed by Wu et al. [155], which is also a patch-based approach (atlas-based) that dynamically adjusts the patch size during the label fusion procedure. Finally, when using the 35 volumes of the MICCAI12 database for evaluation, the approach that performed best is the multi-class dictionary learning approach presented by Benkarim et al. [41].

2.5.4 Quantitative analysis of available software

In this section we evaluate three publicly available and well-known software tools, each relying on a different category of our classification, namely MABMIS [150] (atlas-based), FreeSurfer [43] (learning-based) and FIRST [40] (deformable). The three software tools have been evaluated on the 30 subjects of the Hammers Adult atlases database [225, 226] with default parameters. See Table 2.6 for the details of this database. The DSC and the HD have been computed for the following structures: hippocampus, thalamus, caudate nucleus, putamen, pallidum, amygdala, accumbens, lateral ventricles and brainstem. The obtained results are shown in

Table 2.8: Software tools evaluation on the Hammers adult atlases database.The results show the Dice similarity coefficient (DSC) and the Hausdorff distance (HD) for each structure separately (mean±std). Acronyms from left to right are: hippocampus (HIP), thalamus (THA), caudate nucleus (CAU), putamen (PUT), pallidum (PAL), amygdala (AMY), accumbens (ACC), lateral ventricle (LV) and brainstem (BS). The results are averages for each structure (left and right pair combined), except for brainstem. The - symbol indicates that the FIRST software does not segment the lateral ventricles. (*) The low DSC values reported for the hippocampus are due to the ground truth segmentations of the Hammers atlases database, in which only the hippocampus head is labeled, whereas the evaluated tools segment the whole structure.

Method	**Ms**	HIP (*)	THA	CAU	PUT	PAL	AMY	ACC	LV	BS
MABMIS	DSC	0.65±0.04	0.85±0.02	0.80±0.03	0.84±0.03	0.71±0.07	0.58±0.08	0.50±0.08	0.82±0.04	0.73±0.03
(atlas)	HD	17.75±6.35	6.73±3.03	8.40±2.28	5.11±1.11	4.98±1.08	7.08±2.04	7.76±2.00	37.00±2.59	16.76±1.89
FreeSurfer	DSC	0.61±0.03	0.82±0.03	0.81±0.02	0.80±0.02	0.70±0.06	0.73±0.03	0.48±0.06	0.73±0.07	0.83±0.01
(learning)	HD	19.00±6.17	7.96±2.74	7.00±1.50	5.50±0.82	5.21±1.61	4.37±0.81	8.04±1.97	39.30±2.44	14.30±1.87
FIRST	DSC	0.64±0.02	0.85±0.03	0.85±0.02	0.87±0.02	0.75±0.05	0.74±0.04	0.54±0.08	-	0.77±0.10
(deform.)	HD	19.55±5.72	7.81±2.87	6.34±1.90	4.40±1.13	4.65±1.38	4.18±0.76	9.29±6.91	-	15.63±3.94

Table 2.8 for each of the strategies.

Analyzing each strategy separately, we observe that in terms of both volume overlap (DSC) and contour distance (HD), the deformable method is the one that performs better for caudate nucleus (0.85 mean DSC / 6.34 mean HD, respectively), putamen (0.87 / 4.40) and globus pallidum (0.75 / 4.65), whereas the atlas-based strategy is the one that performs better for thalamus (0.85 / 6.73) and lateral ventricles (0.82 / 7.76) segmentation. We notice from the table that the learning-based strategy provides the lowest performance in segmenting the thalamus (0.82 / 7.96), the putamen (0.80 / 5.50), the pallidum (0.70 / 5.21) and the lateral ventricles (0.73 / 39.30), while the atlas-based strategy provides the poorest results when segmenting the caudate nucleus (0.80 / 8.40). On the other hand, looking only at the HD metric for comparison, we see that for hippocampus (17.75) and accumbens (7.76) segmentation, the atlas-based method performs better than the others.

The results obtained here mostly follow the trend observed in the works highlighted in Section 2.5.3 for each particular structure. For hippocampus segmentation, two of the four enhanced works were atlas-based (Cardoso et al. [143] and Pipitone et al. [154]) while the same happened in thalamus segmentation (Aljabar et al. [142] and Wang and Yushkevich [151]). In caudate nucleus segmentation, the approach that achieved the highest DSC (Gao et al. [186]), with a total of 24 testing cases, was a deformable strategy. On the other hand, in putamen segmentation, the work of Patenaude et al. [40], which is the deformable method analyzed in this section, even not being the one that achieved the highest DSC results, was highlighted due to its results robustness with a large testing dataset (336 volumes). Finally,

either for accumbens or lateral ventricles segmentation, almost all of the highlighted methods were atlas-based [142, 144, 150, 151].

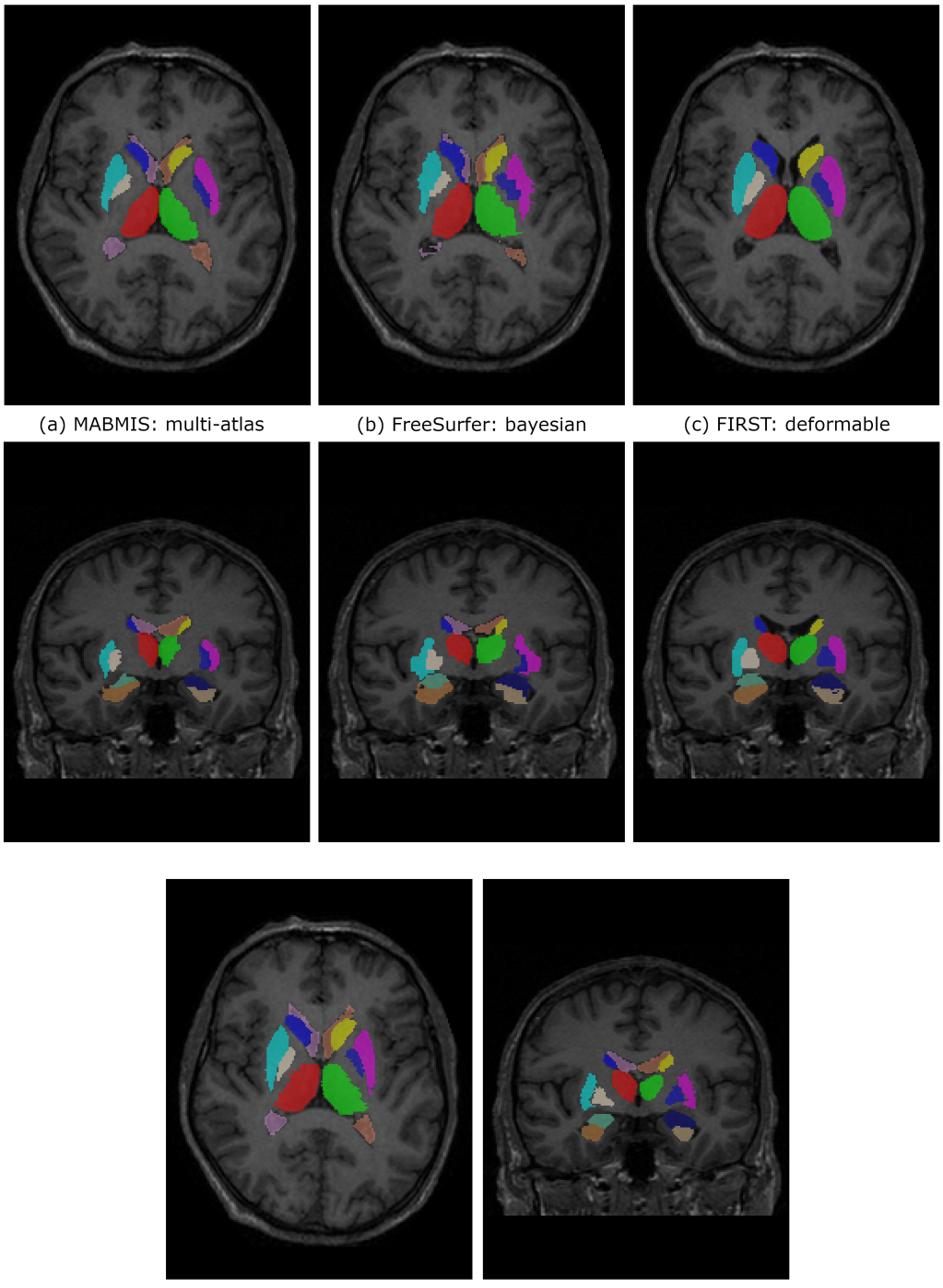
As shown in Table 2.8, when using MABMIS for segmentation, the results for the amygdala and the brainstem are much poorer than the ones of the learning-based and deformable strategies. The reason for this arises from the database used to build the atlases tree in the MABMIS algorithm. In this case we used as atlases the MICCAI'12 database, whose labels for the amygdala and brainstem are not exactly the same as the ones in the testing dataset (Hammers adult atlases). A common practice in the evaluation of multi-atlas algorithms is to use the atlases database, leaving one atlas out for testing and repeating this strategy for all the atlases in the database. This can be a good practice although it can also provide biased results.

Another relevant issue from the performed quantitative evaluation is that the DSC values obtained for the hippocampus when using the same tools are much lower than the ones gathered from the analyzed works in the literature. This is also due to the ground truth segmentations of the Hammers atlases database, in which the hippocampus is labeled following a different protocol. Therefore, the non-existence of a standardized labeling protocol for brain structure segmentation constitutes another open problem in order to evaluate the performance of automatic brain structure segmentation methods.

Figure 2.1 shows some illustrative results of automatic brain structure segmentation obtained with these different strategies. As seen in this figure, the result of the atlas-based approach is similar to the ground-truth, in spite of the atlases used were collected from a different database to that of the target image, whereas the learning-based approach, as already mentioned, is the one that seems to provide the least accurate results, not being able to provide well defined structure boundaries, and including mis-classifications such as the ones in the lateral ventricles (purple label). The deformable method, contrary to the other two, performs a well-defined segmentation, as it is topologically constrained.

2.5.5 Quantitative analysis of the MICCAI Challenges

Analyzing the results of the MICCAI challenges we see that for caudate nucleus segmentation (CAUSE07 challenge) the method that gave the best overall result was the ISICAD [231]. This method performed an adaptive local multi-atlas segmentation that locally decided how many and which atlases were needed to segment a target image and registered only the selected parts of those atlases. The second place in this challenge was for the MIAL-SFU [232] algorithm, that can be classified in the hybrid category, since it used FreeSurfer segmentation for initialization (which has been classified as a learning-based algorithm) and then looked for the best transformation to perform label propagation. Finally, the third position in this



(a) MABMIS: multi-atlas (b) FreeSurfer: bayesian (c) FIRST: deformable

(d) Ground truth

Figure 2.1: Example of automatic structure segmentation using various publicly available softwares. Image a03 from Hammers's work [225, 226]

Table 2.9: MICCAI'07 Challenge (CAUSE07) results. Acronyms from left to right are: volumetric overlap error (OE), relative absolute volume difference (VD), average symmetric surface distance (AD), root mean squared symmetric surface distance (RMD) and maximum symmetric surface distance (MD). OE and VD results are shown in percentage while AD, RMSD and MD are measured in millimeters.

Method	OE	VD	AD	RMD	MD	Score	Strategy
ISICAD [231]	31.09	4.98	0.62	1.03	7.03	79.16	Multi-atlas
MIAL-SFU [232]	25.49	-5.01	0.56	1.41	12.77	75.99	Hybrid
Moghaddam [165]	29.98	-0.96	0.64	1.37	11.79	75.77	Learning-based

Table 2.10: MICCAI'12 Grand Challenge results. Dice similarity coefficient (DSC) results for cortical structures (CS) and non-cortical structures (NCS). The fourth column of the table shows the mean DSC across all brain labels and all subjects in the testing cohort.

Method	DSC (CS)	DSC (NCS)	Overall mean DSC	Strategy
Wang and Yushkevich [151]	0.73	0.83	0.76	Multi-atlas
Non Local STAPLE [152]	0.73	0.82	0.75	Multi-atlas
MALP_EM [233]	0.73	0.82	0.75	Hybrid

caudate segmentation challenge was for the method proposed by Moghaddam and Soltanian-Zadeh [165], a learning-based strategy. The numerical results obtained in the challenge for these three methods are detailed in Table 2.9.

For the MICCAI 2012 Grand Challenge, we observe that the best segmentation strategy was the one presented by Wang and Yushkevich [151] which performed label fusion followed by 'corrective learning'. The second position was for the Non Local STAPLE (NLS) algorithm [152], which is also a label fusion strategy, while the third position was for the MALP_EM [233] algorithm. MALP_EM combined both multi-atlas label propagation and probabilistic segmentation, performing a locally weighted fusion to obtain probabilistic labels that were used as priors in an EM refinement step. It is important to remark that the representation of each category in this challenge may be biased since, as said before, although any reproducible method was accepted, it was assumed the majority of the participant methods to be multi-atlas based. Therefore, it should come as no surprise that the best strategies relied on the atlas-based strategy. The obtained results for these three methods in terms of DSC are shown in Table 2.10.

Finally, in the free-for-all subchallenge of the MICCAI SATA challenge the mean/median DSC results ranged from 0.61/0.63 to 0.86/0.87 and the mean/median Symmetric HD results ranged from 3.30/3.10 mm to 7.68/8.00 mm. The best entry in terms of both DSC and Symmetric HD was the UPENN_SBIA_MAM algorithm [234], which was based on a multi-atlas strategy that performed atlas selection.

2.6 Concluding remarks and future trends

2.6.1 Overview

In this chapter we have reviewed the problem of automatic brain structure segmentation in MR images, presenting a classification of the state-of-the-art methods based on the segmentation strategy used, where we have identified five main categories. The first category includes the approaches that rely on topological atlas registration, either using a single atlas (label propagation) or a set of them (label fusion). The second category in this classification comprises the learning-based strategies which have been subdivided into supervised methods such as classifiers, like ANNs or SVM, and methods based on Bayesian inference. Finally, the last three categories include deformable methods (which include ACM, ASM, AAM and other energy minimizing strategies), region-based approaches, and strategies that combine some of the methods of the previous categories. To conclude this classification, we have discussed the strengths and weaknesses that each category presents.

As we have seen from Tables 2.2 to 2.4, a comparison of the state-of-the-art methods is not an easy task, since there is a lack of standard in both databases and metrics for evaluation. In order to solve this problem, some challenges [228, 229, 230] have been proposed in recent years, but they are either target-specific (caudate nucleus) [228] or methodology-oriented (multi-atlas) [229, 230], which makes the methods included in the challenge not representative of the overall strategies available in the literature. Furthermore, the quantitative results of the challenges are often given in terms of average measures for the whole structures instead of giving evaluation results for each structure independently, which means that an analysis of which method performs better for a given structure cannot be deduced.

We have analyzed the results of the different works reviewed in this chapter, providing an overview of the current state of the art in terms of different evaluation metrics, databases and number of cases used, showing the results both from the point of view of the segmentation strategies and segmented target structures. Furthermore, a quantitative evaluation of three publicly available software tools, each relying on a different category of our classification, have been performed. Lastly, we have commented on the results and the methods presented to the three MICCAI challenges in brain structure segmentation.

2.6.2 Future trends

In recent years, it seems that there exists a trend toward the use of multi-atlas strategies for brain structure segmentation, either as a segmentation method on its own or

combined with other strategies. Given the fact that its greatest weakness is that it is computationally very expensive due to the large number of registrations it has to perform, some strategies have included either atlas selection [129, 141, 151] to reduce the number of registrations, or non-local weighting label fusion [42, 144] in which only affine registrations are needed, which reduces significantly the computational time. All that, in combination with computers becoming more powerful, seems to make multi-atlas approaches a good strategy to follow [151, 155]. Furthermore, hybrid methods have been shown to achieve good results [67, 70, 138] in segmenting brain structures as they can combine the advantages of each category of methods, and try to overcome their weaknesses with the strengths of the other methods in the amalgamation. Because of that, combining multi-atlas methods with any other method in the classification would be a good line of investigation. Atlases give robustness to the method and they are good ways of providing spatial information for structures, thus, we believe that merging atlases with another method which is able to model the structure's appearance, such as supervised classifiers or AAM would lead to improved segmentation results. Van der Lijn et al. [199] presented work relying on this strategy for hippocampus and cerebellum segmentation, achieving mean DSC indexes of 0.87 and 0.95 respectively.

Most of the methods reviewed here are tested with images of non-lesioned brains, either healthy subjects or patients with schizophrenia, epilepsy or attention deficit hyperactivity disorder. One of the biggest problems is that, as far as we know, there is no publicly available database of lesioned brains with ground truth of structure segmentation, to test or train the methods. However, when performing structure segmentation in brain MRI of patients with demyelinating lesions (as in MS or lupus) or space-occupying lesions (such as tumors), the performance of some of these methods is affected. To the best of our knowledge, how demyelinating lesions affect the automatic brain structure segmentation algorithms has not yet been evaluated. For this reason, in the following chapter, we perform an exhaustive analysis of this effect on three of the most popular automatic brain structure segmentation methods reviewed in this chapter, each relying on a different category of the classification proposed in Section 2.3. As a preliminary example of this effect, Figure 2.2 shows how MS lesions can introduce errors in the automatic brain structure segmentation methods. Figure 2.2 (a) shows how either the caudate nucleus (yellow label) and the lateral ventricle (brown label) are over-segmented, due to the fact that the automatic method is interpreting the lesions as part of the structures. The same situation is seen in Figure 2.2 (b), in which MS lesions produce false positives in the putamen (pink label) and lateral ventricle segmentation. As far as we know there has not been a proposal tackling this issue, and therefore there is a need to improve the performance of such methods, trying to make them robust with brain lesions. In order to solve this problem, the proposal of new strategies will be studied later in this PhD thesis. Note that an accurate segmentation is necessary in order to per-

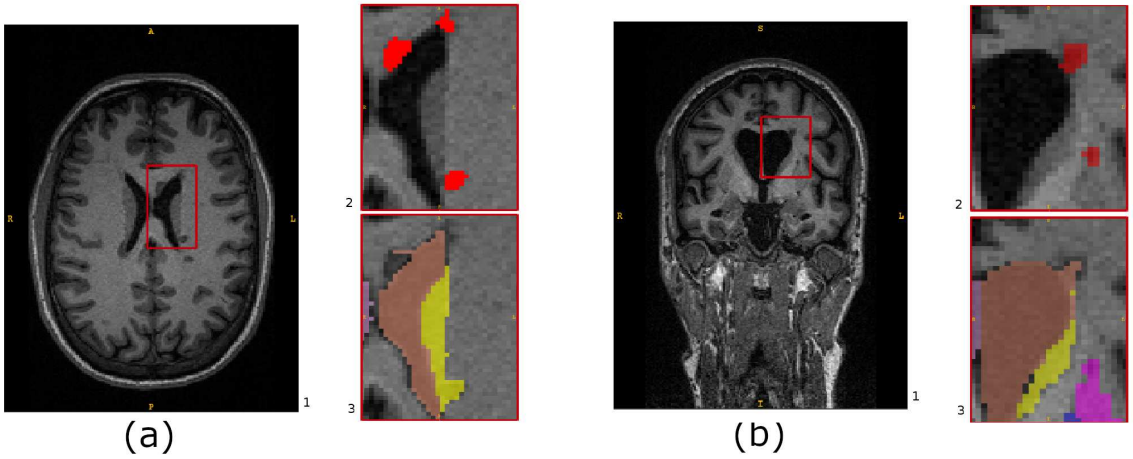


Figure 2.2: Structure segmentation in brains with demyelinating lesions. FreeSurfer [213] segmentation results show that these lesions affect the structure segmentation performance, increasing the number of false positives. Images show (1) the original T1-weighted MRI, (2) the lesion ground truth and (3) the automatic segmentation. Figure 2.2 (a) shows over-segmentation of the caudate nucleus (yellow) and the lateral ventricle (brown) whereas in Figure 2.2 (b) the same situation is shown for the putamen (pink) and the lateral ventricle. Images from Vall d'Hebron Hospital, Barcelona.

form disease follow-up, for instance, in MS patients where deep gray nuclei atrophy is closely related to the magnitude of inflammation. Interestingly, lesion filling techniques [29, 30, 31] to reduce the effect of hypo-intense T1-w MS lesions have already been applied to assess the progression of GM atrophy [32, 220] showing an improvement in the accuracy of tissue volume. Furthermore, integrating automatic lesion segmentation and filling into automatic tissue segmentation pipelines has recently been studied [219], showing very similar results to that of manually segmenting the lesions. This has not yet been integrated as part of any automatic brain structure segmentation pipeline, and indeed opens new challenges to the research community.

Chapter 3

The effect of multiple sclerosis lesions on automatic brain structure segmentation

3.1 Introduction

As seen in Chapter 2, neurodegenerative disorders are frequently associated with structural changes in the brain, such as variations in the volume or shape of the deep GM structures [98, 235, 236]. In the case of MS, the study of the atrophy effect on these structures is an increasing field of research, extensively investigated on patients with CIS and early RRMS, in order to identify the specific structures that are more susceptible to this disease [237, 238, 239, 240].

As we noted in the previous chapter, the vast majority of the proposed automatic brain structure segmentation methods are designed to segment non-lesioned brains, either from healthy subjects or from patients with schizophrenia, epilepsy and other diseases, and these patients typically do not have WM lesions such as those found in MS patients. As seen in Chapter 1, these lesions are hypo-intense in T1-w MRI, and their intensity is very similar to that of the GM, which can make the performance of these automatic methods unreliable.

One of the largest problems when quantitatively evaluating these methods is that training and testing are difficult due to the limitation of having datasets with both structure ground truth and lesion annotations. Here, we overcome this issue with an approach to synthetically generate MS lesions (from cases with lesion manual

annotation) in healthy subjects from whom brain structure ground truth information is available.

In this chapter, we evaluate the effects of simulated MS lesions on the performance of three well-known automatic brain structure segmentation approaches, each of which follows a different segmentation strategy of our classification (see Section 2.3). FreeSurfer [43] follows a learning-based strategy, the FIRST [40] method uses a deformable approach, and the multi-atlas-based segmentation strategy is fused by means of majority vote [139]. To the best of our knowledge, this is the first study of the effect of MS lesions on brain structure segmentation. To evaluate this effect, we generate a set of 100 synthetic MS patients' images with a total of 2174 lesions, using as a base two different databases with structure ground truth (the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database [229] (MICCAI'12) and the Internet Brain Segmentation Repository[1] (IBSR18)). The DSC and volume differences in the automatic segmentations between the healthy and the simulated patient images for the three approaches are analyzed separately by brain structure and lesion location.

3.2 Experiments and evaluation

3.2.1 Segmentation methods

Three publicly available structure segmentation methods are used in this study. The first of these is the segmentation algorithm [43] included in the well-known software FreeSurfer[2] [213], which follows a learning-based strategy. The second method is the Bayesian appearance model proposed by Patenaude et al. [40], which is a deformable strategy and is implemented as part of the FSL[3] package under the name FIRST. The last algorithm follows an atlas-based strategy, more specifically, multi-atlas registration fused by means of the simple and well-known fusion strategy, majority vote. For this last strategy, we follow the procedure described by Artaechevarría et al. [139]. We first perform an affine transformation to align the volumes, followed by a non-rigid B-spline registration using an isotropic grid spacing of 8.0 pixels and mutual information as a similarity metric. As in Artaechevarría et al., Elastix [241] software[4] is used to perform the registrations. Both FreeSurfer and FIRST are executed with default parameters.

[1] https://www.nitrc.org/projects/ibsr
[2] http://freesurfer.net/
[3] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/
[4] http://elastix.isi.uu.nl/

3.2.2 Synthetic MS patient generation

Currently, there is a lack of public database information regarding MS patients with both brain structures and lesion ground truth. Therefore, to evaluate how WM lesions affect the performance of automatic brain structure segmentation algorithms, we present here our method to generate synthetic lesions from MS patient images to healthy subjects with brain structure ground truth information.

In the following sections, we present the steps of this pipeline including lesion dictionary construction, pre-processing and lesion generation. In the first step, a dictionary of individual lesions is built, which are extracted from images of real MS patients with manually annotated lesion masks, and classified based on their type, size, etc. In the second step, the images are prepared for the lesion generation, making sure that both the patient images containing the classified lesions and the target image are in the same coordinate space and intensity range. Finally, in the last step, the lesions from the dictionary are synthetically added to the target.

Lesion dictionary

To build the lesion dictionary, it is necessary to have an MRI dataset from MS patients with a manual lesion annotation (ground truth). This dataset must consist of a set of T1-w volumes and their corresponding lesion delineations.

As MS lesions are usually annotated using either the FLAIR or PD-w sequence and lesions tend to look smaller in T1-w images, we first reduce the lesion masks based on their appearance in the T1-w sequence. To accomplish such a reduction, we perform a tissue segmentation using FSL FAST [242] and discard from the ground truth the voxels classified in the WM class. Once the masks are reduced, we assign an independent label to each lesion in the image with the final objective of evaluating each lesion independently. This is achieved by obtaining the connected components of the lesion mask and considering each mask as a single 3D lesion.

As demonstrated later in this section, we approach strictly WM lesions and WM lesions attached to the lateral ventricles (referred as LV lesions) differently. Therefore, it is important to classify each lesion from the dataset into one of these two groups. To do this, we calculate the 3D Euclidean distance from each lesion contour to the CSF. At this point, we have all of the necessary information to build the lesion dictionary, which includes the following information for each lesion in the dataset: the lesion label, the image of procedence, the lesion type (WM or LV) and, for practical issues, the lesion size.

Pre-processing

After construction of the lesion dictionary, and before applying the lesion generation method as such, some pre-processing steps are required. First, we perform tissue segmentation of the target image, as it is necessary in our lesion generation method to restrict the generated lesion position and avoid overlap with the CSF. Afterwards, we perform both rigid and non-rigid registrations of each MS patient in the dataset to the target image. Both registrations are performed using the original images without any pre-processing (either the MS patient image or the target image). Therefore, we acquire two new volumes in the target image space −one rigidly registered to the target image and another non-rigidly registered− in addition to the corresponding lesion masks. As lesions can affect the non-rigid registration process, their voxels are masked out in order to achieve a more accurate result. The NiftyReg software[5] is used to perform the registrations.

With all of the selected patients in the target image space, we normalize the source image (registered MS patients) intensities to the target image to create realistic lesions. For this purpose, the skull-stripped images of both the MS registered patients and the target image are used to avoid the influence of non-brain intensities on the histogram matching normalization process. To normalize the images, ITK[6] implementation described by Nyul et al. [243] is used.

Lesion generation

We finally generate the new MS lesions in the target image as follows:

$$patch_{target} = patch_{lesion} \times mask + patch_{target} \times (1 - mask)$$

where $patch_{target}$ corresponds to the 3D patch in the target image where the lesion is going to be generated, $patch_{lesion}$ is the 3D patch of the registered and normalized patient containing the lesion, and $mask$ is the corresponding lesion mask to which we apply a Gaussian filter to make the transition between the healthy tissue and the lesion smoother.

As already mentioned, we deal with the lesions separately, generating one lesion at a time and selecting its mask ($mask$) from the corresponding registered image: rigid for WM lesions (since the idea for these type of lesions is to keep the original shape) and non-rigid for LV lesions (given that these lesions have a special shape that depends on the morphology of the LV and we have found necessary to deform

[5]http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg
[6]https://itk.org/

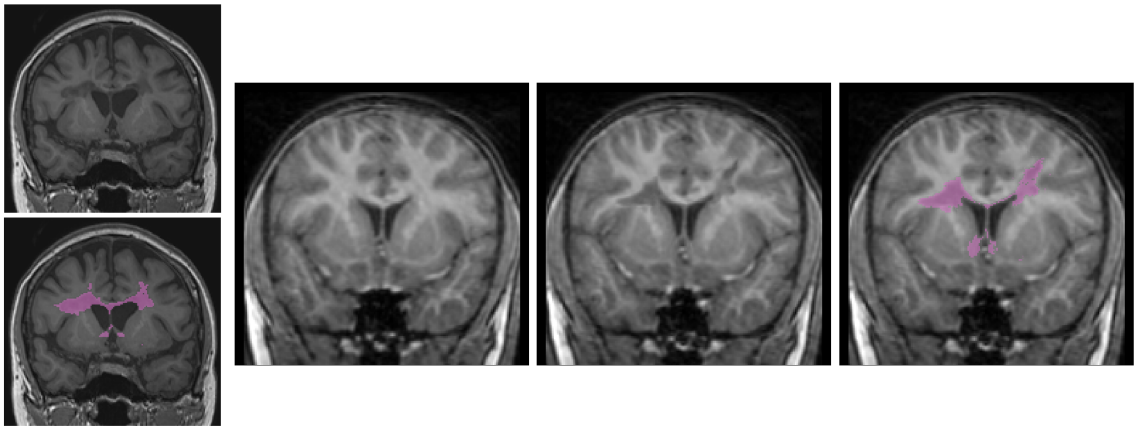


Figure 3.1: Example of MS lesions generation. (a) Original MS patient image and corresponding lesion mask (patient 01016SACH form the MICCAI'16 Challenge database [244]) (b) Healthy image (subject IBSR_17 from the IBSR18 database) (c) Generated image (d) Lesion mask.

the original lesions in order to get adapted to that structure). Therefore, the final lesion mask is composed by the sum of the independent lesion masks proceeding from the registered images. Since the original location of the lesion is better captured in the non-rigid registered images, the lesion location in the target image ($patch_{target}$ center), and therefore in the final lesion mask, is always selected based on the non-rigid position of the lesion center.

Figure 3.1 shows an example of a generated image applying the proposed methodology. Figure 3.1 (a) shows the original MS patient whose lesions have been reproduced, whereas Figures 3.1 (b) to 3.1 (d) show the original healthy subject and the lesion generation results.

3.2.3 Data

To study the performance of the different structure segmentation algorithms, the following two publicly available databases are used: 1) the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database [229] (MICCAI'12) and 2) the Internet Brain Segmentation Repository[7] (IBSR18). The first database consists of 15 T1-w MR images for training and 20 for testing as obtained from the OASIS[8] [223] project and labeled by Neuromorphometrics, Inc.[9], which includes la-

[7]https://www.nitrc.org/projects/ibsr
[8]http://www.oasis-brains.org/
[9]http://neuromorphometrics.com/

Table 3.1: Properties of the four selected healthy subjects.

Name	Database	Age	Scanner	Volume (mm)	Voxel (mm)
1004	MICCAI'12 [229]	23	Siemens (1.5T)	256×256×256	1×1×1
1116	MICCAI'12 [229]	61	Siemens (1.5T)	256×256×256	1×1×1
IBSR_08	IBSR [7]	60	Siemens (1.5T)	256×256×128	1×1×1.5
IBSR_17	IBSR [7]	8	GE (1.5T)	256×256×128	0.84×0.84×1.5

bels for the whole brain. The second database consists of 18 T1-w MR images with expert manual segmentations of 43 individual structures provided by the Center for Morphometric Analysis at Massachusetts General Hospital.

The lesion generation method explained in Section 3.2.2 is used to generate synthetic lesions over the testing subjects. The procedure to select the healthy subjects on which to generate the lesions is as follows. The 38 testing images (20 from MICCAI'12 + 18 from IBSR18) are segmented using the three algorithms presented in Section 3.2.1. The training cohort from the MICCAI'12 database is used as the atlas for the multi-atlas method. Two subjects from each database are selected based on their DSCs achieved for the evaluated structures. According to this metric, we discard the subjects who represent outliers in any of the analyzed structures in relation to the other subjects' segmentation in the same database. From the remaining subjects, the two that best represent each database are chosen with a DSC for almost all of the structures in the median of all the subjects for each of the analyzed segmentation algorithms. Regarding these criteria, subjects 1004 and 1116 from the MICCAI'12 database and subjects IBSR_08 and IBSR_17 from the IBSR18 database are selected. Some details about the selected images can be found in Table 3.1.

Five MS databases including MICCAI'08[10] [245], MICCAI'16[11] [244], ISBI'15[12] [246] and two in-house databases are included in our lesion dictionary, with a total of 140 patients and 4291 lesions. The lesions from each patient are simulated in the selected healthy subjects, and once completed, a second selection is performed to obtain the final images included in our study. Twenty-five MS patient images are selected in such a way that lesions are represented in all of the analyzed structures, and different patient volumes and lesion numbers are achieved. The simulations of these 25 patients in the four selected subjects are chosen as the cohort on which to perform our experiments, which includes a total of 100 simulated images. Details of the original 25 MS patients are shown in Table 3.2. The original cohort has a total of 1429 WM lesions, but for practical issues only lesions larger than 27 mm^3 are simulated.

[10] http://www.ia.unc.edu/MSseg/

[11] https://portal.fli-iam.irisa.fr/msseg-challenge/overview

[12] http://iacl.ece.jhu.edu/index.php/MSChallenge

Table 3.2: Properties of the twenty five selected MS patients. Lesion volumes and number of lesions are calculated from the reduced masks based on lesion appearance in the T1-w sequence.

Database	No.	Scanner	Volume (mm)	Voxel (mm)	Lesion vol (ml)	No. lesions	Lesion size (mm^3)
MICCAI'08 [10]	4	Siemens Allegra (3T)	512×512×512	0.5×0.5×0.5	$6.28 - 14.39$	$51 - 125$	$0.13 - 0.41 \times 10^4$
MICCAI'16 [11]	5	Siemens Verio (3T)	176×256×256	1×1×1	$0.89 - 59.79$	$7 - 69$	$1.00 - 3.28 \times 10^4$
MICCAI'16 [11]	3	Siemens Aera (1.5T)	256×256×176	1.08×1.08×0.9	$1.43 - 35.26$	$19 - 65$	$1.05 - 1.65 \times 10^4$
MICCAI'16 [11]	3	Philips Ingenia (3T)	2: 200×336×336 1: 210×336×336	0.85×0.74×0.74	$4.78 - 26.60$	$51 - 133$	$0.47 - 1.54 \times 10^4$
ISBI'15 [12]	2	Philips (3T)	256×256×120	0.82×0.82×1.17	$12.68 - 25.67$	$42 - 45$	$0.80 - 1.39 \times 10^4$
IN-HOUSE 1	7	Siemens Trio Tim (3T)	6: 128×240×256 1: 128×232×256	1.2×1×1	$1.30 - 14.25$	$20 - 99$	$1.20 - 0.21 \times 10^4$
IN-HOUSE 2	1	Siemens Simphony Quantum (1.5T)	192×256×46	0.98×0.98×3	31.60	9	$48.64 - 2.30 \times 10^4$

The lesions of the same MS patients are replicated on the four selected healthy subjects, leading to simulated MS patients with lesion loads ranging from 0.44 to 59.93 ml and a number of lesions per patient ranging from 1 to 62. The total number of generated lesions over the 100 synthetic images is 2174.

As stated previously, the 25 MS patients are selected in such a way that there are lesions represented in all of the analyzed structures. As a result, the generated cohort contains 27 images with lesions in the left thalamus, 35 with lesions in the right thalamus, 83 in the left caudate, 81 in the right caudate, 25 in the left putamen, 30 in the right putamen, 4 in the left pallidum, 5 in the right pallidum, 41 in the left hippocampus, 64 in the right hippocampus, 27 in the left amygdala, 36 in the right amygdala, 28 in the left accumbens, 23 in the right accumbens and 57 in the brainstem.

Notice that registration inaccuracies affect the lesion generation procedure in different ways. First of all, when we perform registration to take the MS patient image to the healthy image space, if we are moving to a lower resolution space and the lesions are small and only visible in a low number of slices it may happen that these lesions disappear from the registered image. Moreover, registration can make the lesion position displace to an upper/lower slice or even change the lesion position a little, which for some small lesions could result in being or not overlaid on the same structure. Furthermore, in spite of the lesions having been masked out to perform the non-rigid registration, their morphology (shape and size) may have been affected differently from one healthy to another and as we have set the restriction that lesions above 27 mm^3 are simulated, it could had happened that the same lesion is simulated in one healthy but not in the others.

3.2.4 Evaluation

Images from both healthy subjects and their corresponding simulated MS patients are segmented using the three segmentation strategies presented in Section 3.2.1. Since FIRST only provides segmentation results for the sub-cortical structures and the brainstem, the performance of the three algorithms and the effects of the lesions are only evaluated on this subset of brain structures. All of the structures are evaluated separately for the left and right hemispheres, except the brainstem, which is a unique structure.

The DSC [247] and the volume differences between healthy controls and simulated patients are used as the main metrics for evaluation, and other metrics such as false positive Dice (over-segmentation) or false negative Dice (under-segmentation) are also analyzed [73] but finally omitted in this study for the sake of simplicity.

Statistical analysis is performed using the Matlab software package[13]. Differences in the performance of the three analyzed methods when segmenting the healthy subjects of both databases were analyzed using pairwise Wilcoxon rank sum tests. Moreover, the Pearson's linear correlation coefficient was used to compute the correlation between the total lesion volume and the changes in DSC and in structure volume. We also compared the methods robustness with respect to each other when simulated lesions were introduced. In order to rank the methods on both IBSR and MICCAI'12 databases, significant pairwise method permutation tests of the absolute DSC differences and the absolute structure volume differences were performed. Furthermore, we tested for significant differences in the robustness of the three strategies when the lesions were overlaid on the structure of analysis and when they were not. To perform such analysis, series of permutation tests on the absolute structure volume differences with respect to the healthy controls were performed. For our experiments, we have adapted the implementation provided by Klein et al. [76]. For all the permutation tests performed in our experiments, we set the number of comparisons between each pair of methods to $N = 1000$. In all the analysis, hypothesis test with significance level $\alpha = 0.05$ was performed.

3.3 Results

In this section, we analyze the behavior of the automatic brain structure segmentation methods presented in Section 3.2.1. First, we analyze how these methods behave when the healthy subjects from the two databases with ground truth available (IBSR18 and MICCAI'12) are segmented. Then, we compare the automatic segmentations obtained with the three different approaches for both the generated

[13] http://es.mathworks.com/products/matlab

patients and the healthy subject images and perform an analysis of how the simulated WM lesions affect the performance of each software method based on structure and lesion location.

3.3.1 Database performance

Table 3.3 shows the DSC results of the three analyzed segmentation strategies on the healthy subjects from both databases (20 from MICCAI'12 + 18 from IBSR18). As shown in this table, FreeSurfer provides similar results for both databases; however, we can highlight some structures such as the amygdalas ($p \leq 0.001$), the right accumbens ($p < 0.001$), the right thalamus ($p < 0.05$) and the right putamen ($p < 0.01$) that achieve better segmentation results, on average, for the healthy subjects from the IBSR18 database. On the other hand, better results are obtained for the left pallidum ($p < 0.05$), the right pallidum ($p < 0.01$), the left hippocampus ($p = 0.01$) and the brainstem ($p < 0.05$) when the healthy subjects from the MICCAI'12 are segmented.

Smaller differences between the segmentation results in both databases are obtained for the deformable strategy. In this method we can highlight three structures on which this difference is statistically significant: the left caudate ($p < 0.05$), the right caudate ($p < 0.01$) and the right amygdala ($p < 0.01$).

Regarding the majority vote strategy, we see from the table that it provides higher DSC values for MICCAI'12 than for IBSR18 in all of the analyzed structures ($p \leq 0.001$). This difference arises because the atlases used proceed from the training cohort of the MICCAI'12 database, and thus, their similarity in the scanner acquisition configuration and the rank of intensities allow better registration results when segmenting the MICCAI'12 subjects.

In a database-specific analysis, we observe from Table 3.3 that for the MICCAI'12 the differences between the segmentation performance provided by FIRST and majority vote are mostly not significant. However, we can highlight three structures on which these two strategies differentiate, which are the right caudate ($p < 0.01$), the right pallidum ($p < 0.001$) and the brainstem ($p < 0.001$). On the other hand, the results obtained for the IBSR18 database show that for most of the structures, the strategy that provides the best accuracy is FIRST ($p < 0.05$).

3.3.2 Lesion effects per segmentation strategy

The effect of the generated lesions on the three segmentation strategies is analyzed here separately by method performance (DSC) and structure volume.

Table 3.3: Healthies DSC. Structure acronyms are: left thalamus (L.Tha), right thalamus (R.Tha), left caudate (L.Cau), right caudate (R.Cau), left putamen (L.Put), right putamen (R.Put), left pallidum (L.Pal), right pallidum (R.Pal), left hippocampus (L.Hip), right hippocampus (R.Hip), left amygdala (L.Amy), right amygdala (R.Amy), left accumbens (L.Acc), right accumbens (R.Acc) and brainstem (BS). The table shows the DSC values (mean±std) for the MICCAI'12 (M) and IBSR18 (I) databases, separated by segmentation strategy (FreeSurfer, FIRST and majority vote (M.V.)). Highlighted areas show the best segmentation strategy results for a given structure and database. Statistically significant ($p \leq 0.05$) better method performance is shown in bold.

Structure	DB	FreeSurfer	FIRST	M. V.
L.Tha	I	81.53 ± 5.59	**89.34 ± 1.69**	81.68 ± 2.94
	M	83.01 ± 1.77	88.92 ± 1.72	87.73 ± 3.46
R.Tha	I	86.36 ± 2.23	**88.46 ± 1.20**	74.70 ± 4.99
	M	84.88 ± 2.07	89.02 ± 1.83	86.53 ± 4.51
L.Cau	I	79.61 ± 4.96	78.27 ± 4.39	67.39 ± 6.15
	M	80.83 ± 7.89	79.72 ± 11.66	76.64 ± 8.92
R.Cau	I	80.92 ± 4.84	**87.04 ± 2.75**	60.32 ± 6.22
	M	80.11 ± 4.16	**83.66 ± 4.57**	75.82 ± 8.77
L.Put	I	78.88 ± 3.81	**86.88 ± 2.01**	60.25 ± 7.51
	M	77.13 ± 3.86	85.98 ± 7.95	86.24 ± 4.13
R.Put	I	82.92 ± 3.10	**88.05 ± 1.05**	74.29 ± 5.59
	M	79.87 ± 2.62	87.59 ± 6.00	87.75 ± 4.37
L.Pal	I	63.17 ± 17.05	**81.05 ± 3.33**	51.73 ± 10.45
	M	69.25 ± 18.93	81.49 ± 6.04	84.23 ± 2.79
R.Pal	I	77.44 ± 3.23	**80.89 ± 3.70**	68.43 ± 6.34
	M	79.15 ± 8.53	79.93 ± 8.80	**85.23 ± 5.31**
L.Hip	I	76.00 ± 3.58	**80.64 ± 2.31**	62.42 ± 4.82
	M	78.35 ± 5.37	80.85 ± 1.40	78.49 ± 4.48
R.Hip	I	76.66 ± 6.03	**81.68 ± 2.26**	61.13 ± 4.84
	M	79.44 ± 2.54	80.97 ± 2.16	79.32 ± 3.54
L.Amy	I	66.06 ± 6.94	**74.18 ± 6.35**	55.77 ± 7.91
	M	58.47 ± 6.41	72.13 ± 5.44	71.24 ± 7.57
R.Amy	I	69.05 ± 6.73	**75.72 ± 6.18**	46.62 ± 6.19
	M	57.58 ± 7.59	70.67 ± 5.25	73.44 ± 7.14
L.Acc	I	60.42 ± 7.08	**68.40 ± 9.77**	52.44 ± 11.13
	M	62.98 ± 5.51	69.94 ± 8.12	71.03 ± 7.96
R.Acc	I	57.36 ± 7.41	**70.27 ± 7.62**	50.35 ± 9.33
	M	44.26 ± 6.46	67.77 ± 8.87	68.79 ± 10.50
BS	I	84.12 ± 1.96	82.50 ± 2.69	80.79 ± 2.05
	M	85.67 ± 1.96	83.34 ± 1.66	**91.64 ± 1.57**

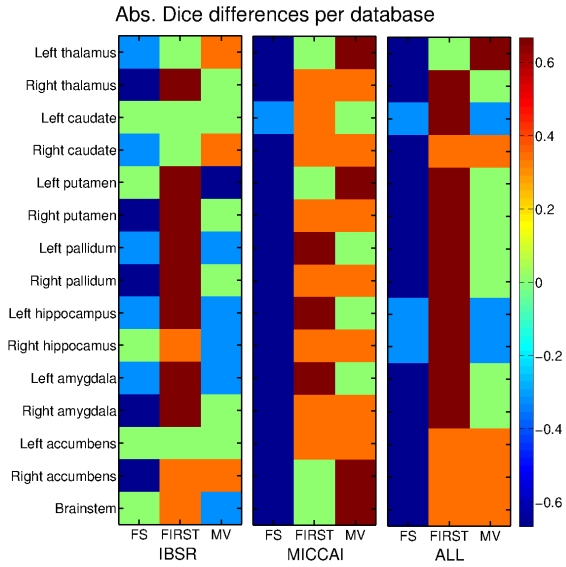


Figure 3.2: Ranking of the segmentation methods separated per database and brain structure obtained from the permutation tests. Color scale that reflects the relative robustness of the segmentation methods when simulated lesions are introduced (with red indicating higher consistency with relation to the healthy segmentation). Each colored square represents the average score for a given method and structure, averaged over 100 segmentations. The scores are values indicating the pairwise robustness of the method relative to each of the other methods, according to DSC differences (generated patient − healthy).

Dice difference

Figure 3.2 presents the relative robustness of the three methods when lesions are introduced as a color-coded table, separated for structure and database. We compare here the three segmentation strategies based on how consistent their segmentations are when lesions are introduced, compared to those obtained for the corresponding healthy subject. From this figure we see that for the IBSR18 database, FIRST provides the most robust results for almost all the structures, however majority vote is more consistent when segmenting the left thalamus and the right caudate. On the other hand, FreeSurfer achieves more unstable results than the other two methods for seven of the analyzed structures, but it is more consistent than majority vote when segmenting the left putamen, the right hippocampus and the brainstem. Regarding the MICCAI'12 database, FreeSurfer shows to be the least robust strategy for all the analyzed structures whereas the other two methods have a similar behavior. Combining both databases FIRST achieves the most consistent results for ten of the analyzed structures whereas the segmentations obtained with FreeSurfer are the

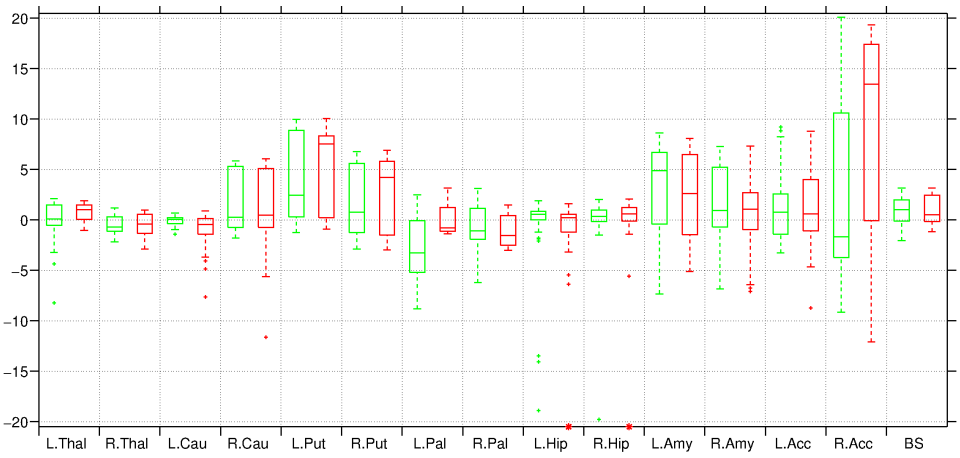
Table 3.4: Permutation tests average ranking based on the method robustness when lesions are introduced. Ranks after conducting permutation tests between absolute DSC differences ($1-|DSC_{generated}-DSC_{healthy}|$) of the generated MS patient images and their corresponding healthy controls (averaged across structures) for each pair of methods, then calculating the percentage of p-values less or equal to 0.05 (of 1000 tests). Members within ranks 1, 2 and 3 have means lying within one, two and three standard deviations of the highest mean, respectively. μ=mean; σ=standard deviation.

	Dice differences					
	IBSR	$\mu \pm \sigma$	MICCAI	$\mu \pm \sigma$	ALL	$\mu \pm \sigma$
Rank 1	FIRST	0.42 ± 0.29	MV	0.33 ± 0.25	FIRST	0.53 ± 0.21
			FIRST	0.31 ± 0.23		
Rank 2	MV	-0.09 ± 0.29				
Rank 3	FS	-0.33 ± 0.28	FS	-0.64 ± 0.09	MV	0.07 ± 0.29
					FS	-0.60 ± 0.14

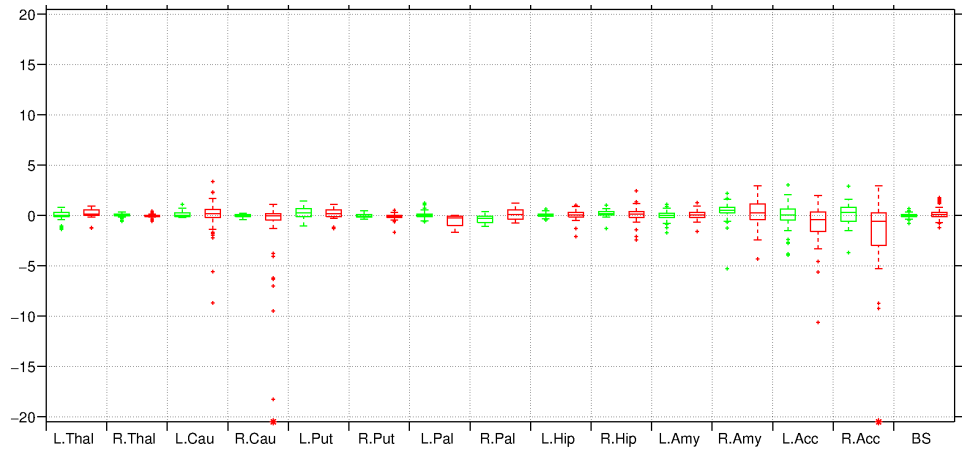
ones that seem more affected by the lesions presence. Table 3.4 presents the ranking of methods separated by database, according to the percentage of permutation tests whose p-values were less or equal to 0.05. Members within ranks 1, 2 and 3 have means lying within one, two and three standard deviations of the highest mean, respectively.

Figure 3.3 shows the boxplots of the differences in terms of DSC between healthy subjects and the corresponding generated MS patients ($patient - healthy$) as separated by segmentation strategy. Green boxplots show the differences for the patients who do not have lesions overlaid on the analyzed structure but do in other parts of the brain. On the other hand, the red boxplots show the differences when these patients have at least one lesion overlaid on that structure, independently of the patient having lesions in other parts of the brain. This figure shows that lesions have an unpredictable effect on the segmentation performance, since in some cases, it might help to improve the segmentation performance, whereas in other cases, it might produce worse overall segmentation results.

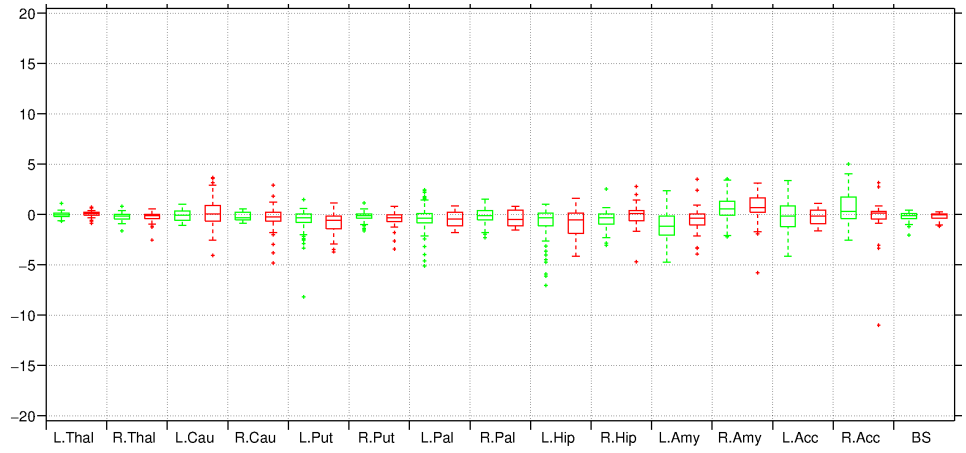
By analyzing each strategy individually, we see that FreeSurfer does not show a clear difference in its performance when lesions are overlaid on a particular structure or not (red vs green). As shown in Figure 3.3 (a), the trends for several structures look similar when lesions are present (in red) or absent (in green), such as in the right thalamus, both putamens, the right hippocampus, the left accumbens and the brainstem. Furthermore, we observe a trend in some of the analyzed structures towards improvement in their segmentation performance when lesions are introduced (anywhere in the brain), as seen in the right caudate, both putamens, the right hippocampus both amygdalas and the brainstem.



(a) FreeSurfer [43]: Bayesian strategy



(b) FIRST [40]: Deformable strategy



(c) Majority vote [139]: Multi-atlas strategy

Figure 3.3: DSC differences (generated patients - healthy) for the 100 generated images using various publicly available softwares. The green boxplots show the differences when there were MS lesions generated on the brain but not on top of that particular structure. On the other hand, red boxplots stand for lesioned structures. The red asterisks on top of the x axis mean that there are more outliers below the -20%. Acronyms from left to right are: left thalamus (L.Thal), right thalamus (R.Thal), left caudate (L.Cau), right caudate (R.Cau), left putamen (L.Put), right putamen (R.Put), left pallidum (L.Pal), right pallidum (R.Pal), left hippocampus (L.Hip), right hippocampus (R.Hip), left amygdala (L.Amy), right amygdala (R.Amy), left accumbens (L.Acc), right accumbens (R.Acc) and brainstem (BS). Notice that the red boxplots for both pallidums and the green boxplots for both caudates contain less than 20 cases.

As shown in Figure 3.3 (b), FIRST seems to be quite robust when lesions are present, showing DSC differences for the non-lesioned structures (in green) that range from -0.36 ± 0.39 (right pallidum) to 0.44 ± 0.94 (right amygdala), whereas the differences for the structures with lesions (in red) achieve values from -2.40 ± 5.54 (right accumbens) to 0.19 ± 0.62 (brainstem). In this strategy, the standard deviations are below 1.50 for all of the structures except the right caudate (3.71) and both accumbens (2.53 left; 5.54 right), but the three of them when lesions are overlaid on these structures (in red), showing that the segmentations provided by this method are consistent, particularly when the lesions in the structure are not present.

In the multi-atlas strategy, the differences between the healthy subjects and the simulated patients are small as shown in Figure 3.3 (c). These differences ranged from -1.24 ± 1.53 (left amygdala) to 0.59 ± 1.45 (right accumbens) for the structures without lesions (in green) and from -0.90 ± 1.16 (left putamen) to 0.59 ± 1.64 (right amygdala) for the structures affected by lesions (in red). As can be deduced from these numbers there is not a clear difference in the method performance when the lesions are overlaid or not, as better segmentation results are achieved for non-lesioned structures in only half of the analyzed cases (8 over 15). In general, majority vote seems to under-perform, on average, when segmenting the simulated patients compared with the healthy subjects for almost all of the structures, independent of whether the lesions are overlaid or not.

On the other hand, in analysis of each structure, we observe that independently of the segmentation strategy, the structure that shows more variability when the lesions are introduced is the nucleus accumbens (1.32 ± 4.00 with FreeSurfer, -1.12 ± 2.53 with FIRST and -0.24 ± 0.81 with majority vote for the left hemisphere, and 9.65 ± 9.87, -2.40 ± 5.54 and -0.48 ± 2.67 for the right hemisphere), whereas those for which the segmentation is more robust are the brainstem (1.03 ± 1.31, 0.19 ± 0.62, and -0.20 ± 0.38) and the thalamus (0.74 ± 0.89, 0.18 ± 0.53 and 0.03 ± 0.35 for the left hemisphere, and -0.48 ± 1.08, -0.04 ± 0.22 and -0.25 ± 0.56 for the right one). On the other hand, the structure for which we have encountered the largest number of outliers is the caudate nucleus, which is also the structure on which we find the largest number of lesions, affected in 83 images (left caudate) and 81 images (right caudate), respectively.

We have also analyzed the extent to which total lesion volume affected the observed changes in DSC for the three evaluated methods. Lesion volume did not correlate with the DSC differences found for FreeSurfer in any of the analyzed structures, except for the left caudate ($r = 0.52$, $p < 0.001$). A similar behavior was found for FIRST, where a significant correlation was seen for both caudates ($r = 0.58$, $p < 0.001$ and $r = 0.51$, $p < 0.001$) and the left hippocampus ($r = 0.47$, $p < 0.001$), while moderate correlation for the right accumbens ($r = 0.38$, $p < 0.001$). Stronger

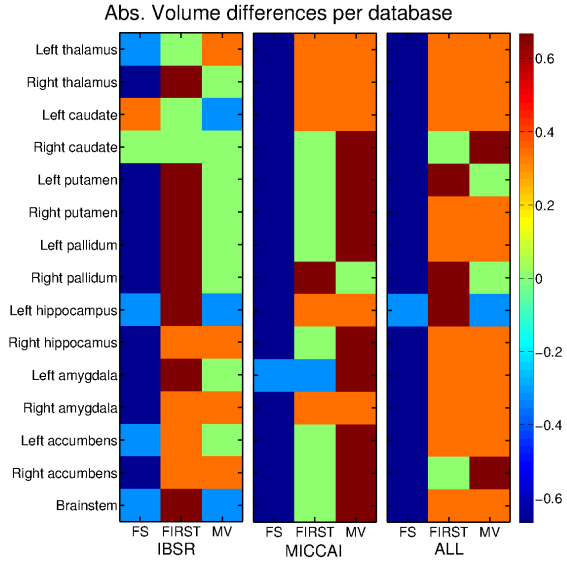


Figure 3.4: Ranking of the segmentation methods separated per database and brain structure obtained from the permutation tests. Color scale that reflects the relative robustness of the segmentation methods when simulated lesions are introduced (with red indicating higher consistency with relation to the healthy segmentation). Each colored square represents the average score for a given method and structure, averaged over 100 segmentations. The scores are values indicating the pairwise robustness of the method relative to each of the other methods, according to volume differences (generated patient − healthy).

correlations were found for the majority vote strategy for the left thalamus ($r = 0.51$, $p < 0.001$), both caudates ($r = 0.73$, $p < 0.001$ and $r = 0.45$, $p < 0.001$), the right putamen ($r = 0.43$, $p < 0.001$) and the left pallidum ($r = 0.42$, $p < 0.001$) whereas a moderate correlation was found for the left putamen ($r = 0.38$, $p < 0.001$), the right pallidum ($r = 0.35$, $p < 0.001$) and both hippocampi ($r = 0.35$, $p < 0.001$ and $r = 0.34$, $p = 0.001$).

Volume difference

Figure 3.4 shows the relative robustness for each method based on each structure volume change, experimented when lesions are introduced to the healthy control. The procedure has been the same as in previous Section 3.3.2. From this figure we see that in the IBSR database, FIRST provides the most robust results for eight of the analyzed structures whereas its consistency is comparable to majority vote in four other structures. Regarding the MICCAI'12 database, the majority vote strategy seems to provide more robust results that the other two methods in nine of

Table 3.5: Permutation tests average ranking based on the method robustness when lesions are introduced. Ranks after conducting permutation tests between absolute volume differences of the generated MS patient images and their corresponding healthy controls (averaged across structures) for each pair of methods, then calculating the percentage of p-values less or equal to 0.05 (of 1000 tests). Members within ranks 1, 2 and 3 have means lying within one, two and three standard deviations of the highest mean, respectively. μ=mean; σ=standard deviation.

	Volume differences					
	IBSR	$\mu \pm \sigma$	MICCAI	$\mu \pm \sigma$	ALL	$\mu \pm \sigma$
Rank 1	FIRST	0.44 ± 0.27	MV	0.51 ± 0.21	FIRST	0.36 ± 0.20
					MV	0.29 ± 0.25
Rank 2	MV	0.02 ± 0.23	FIRST	0.13 ± 0.25		
Rank 3	FS	-0.47 ± 0.30	FS	-0.64 ± 0.09	FS	-0.64 ± 0.09

the fifteen structures whereas it achieves more unstable results than FIRST only for the right pallidum. For this database, majority vote is more robust than FreeSurfer for al the analyzed structures. Combining both databases we see that FIRST and majority vote are comparable in terms on structure volume change consistency, however FreeSurfer seems to be the less robust against lesions achieving only similar results to majority vote for the left hippocampus. Table 3.5 presents the general ranking of the methods based on the results of the permutation tests.

Figure 3.5 presents the relative robustness in terms of structure volume changes for each method comparing the cases on which lesions are overlaid on the structure and they are not (the same case as in Figure 3.3 for red and green boxplots). From this figure we can see that FIRST tends to be significantly more robust when lesions are not overlaid on the structure of analysis whereas FreeSurfer is equally affected wherever the lesions are, except for the right accumbens, on which the segmentation result appears more consistent when the lesions are not overlaid. Regarding the majority vote strategy, it seems there is a trend to be slightly more unstable when lesions are overlaid on the analyzed structure, however this difference is not conclusive.

The relation between the observed change in structure volume and the total lesion volume introduced in the simulated images has also been analyzed for the three evaluated strategies. Significant correlations have not been found for FreeSurfer in any of the analyzed structures whereas the volume differences found in FIRST seem to significantly correlate with the total lesion volume in both caudates ($r = 0.63$, $p < 0.001$ and $r = 0.61$, $p < 0.001$), the right putamen ($r = 0.46$, $p < 0.001$), the left hippocampus ($r = 0.43$, $p < 0.001$) and the right accumbens ($r = 0.45$, $p < 0.001$). A similar behavior than the one seen in the DSC change is seen here for the majority vote strategy. For this method, correlation between structure volume

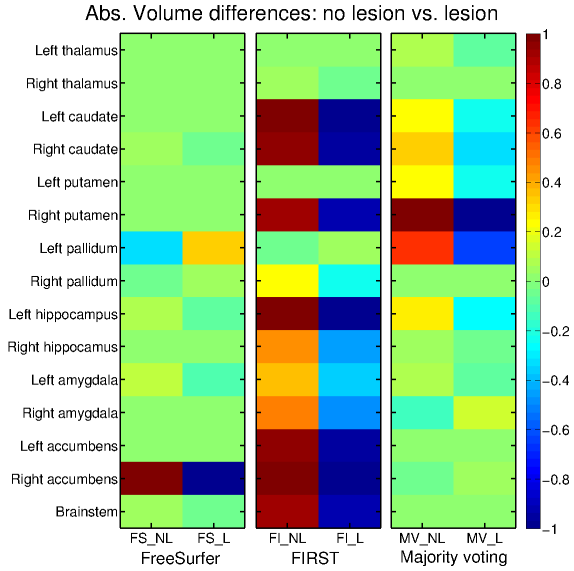


Figure 3.5: Result of the permutation tests. Relative volume consistency of the segmented structures with the three segmentation strategies when lesions are overlaid or not on the structure compared to the healthy controls. Color scale displays the relative robustness for each method (with red indicating higher consistency with relation to the healthy segmentation).

changes and total lesion volume is seen in a higher number of structures than in the other two strategies: both thalami ($r = 0.45$, $p < 0.001$ and $r = 0.48$, $p < 0.001$), both caudates ($r = 0.60$, $p < 0.001$ and $r = 0.56$, $p < 0.001$), both putamens ($r = 0.45$, $p < 0.001$ and $r = 0.45$, $p < 0.001$) and both pallidums ($r = 0.43$, $p < 0.001$ and $r = 0.45$, $p < 0.001$). Also, a trend towards moderate correlation has been observed in the left hippocampus ($r = 0.37$, $p < 0.001$) and the brainstem ($r = 0.32$, $p = 0.001$).

3.3.3 Qualitative results

Figure 3.6 shows a qualitative example of the segmentation results obtained with the three methods. Figures 3.6 (a) to 3.6 (c) show a central slice of the healthy subject and its corresponding simulated MS patient and lesion mask. The automatic segmentation obtained with the analyzed strategies for both a healthy (top row) and simulated MS patient (bottom row) is shown in Figures 3.6 (e) to 3.6 (g), whereas the structure ground truth is shown in Figure 3.6 (d). As seen in Figure 3.6 (e), FreeSurfer improves its segmentation performance for the right caudate (in yellow) and both putamens (in pink and cyan) when the lesions are added, thus reducing the

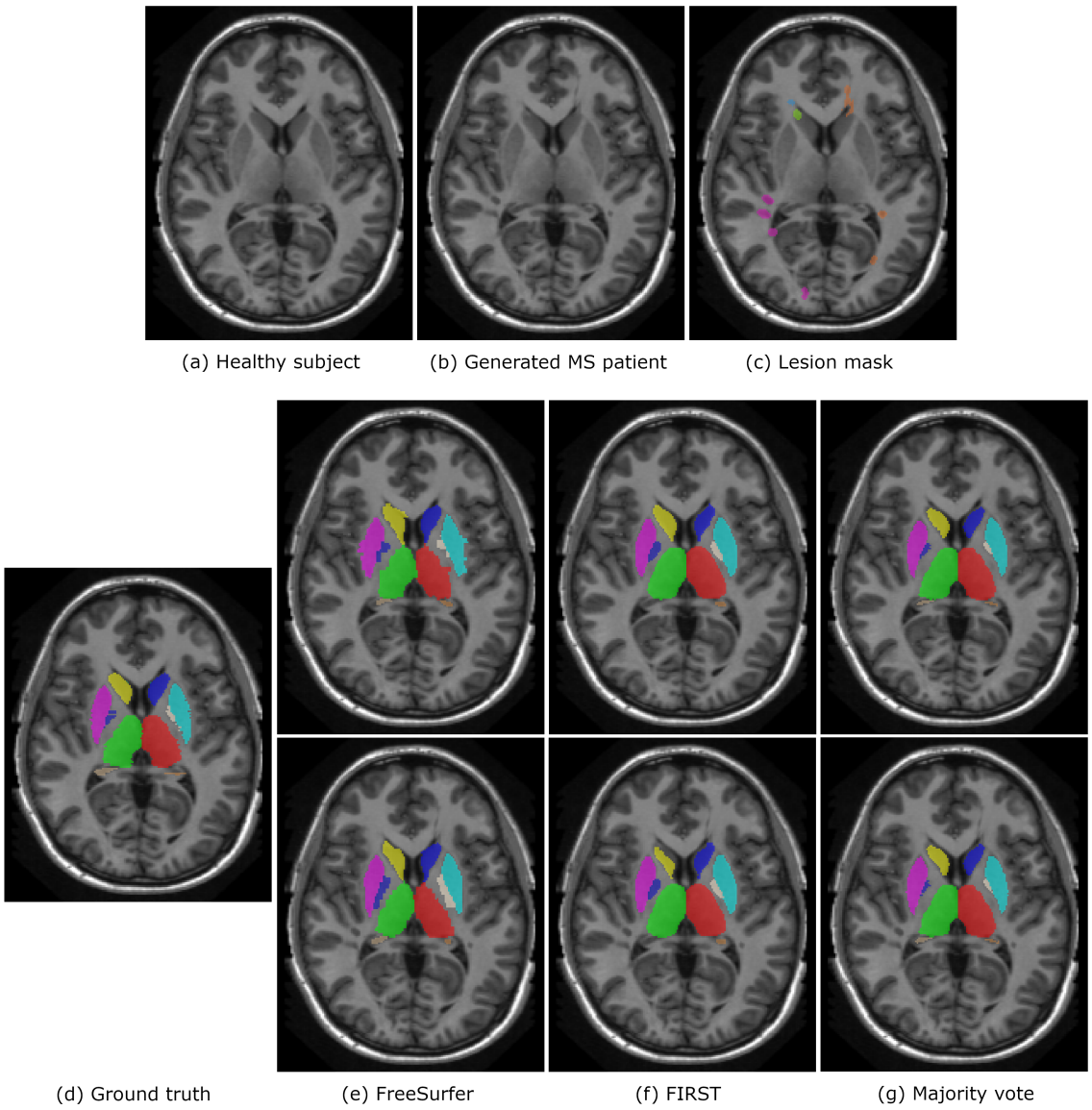


Figure 3.6: Automatic brain structure segmentation. Figures (a)-(c) show the original healthy subject (1004 from MICCAI'12 database), the generated MS patient and the corresponding lesion mask. Figures (d)-(g) show the structure ground truth and automatic segmentation of both images (without lesions in the top row and with lesions bottom row) for the three segmentation strategies.

number of false positives obtained for the healthy images. In this case, the lesions may have modified the global intensity distribution, making the Gaussian chosen to represent the structure intensity more precise and improving the segmentation performance. However, the performance for both pallidums (in blue and white)

decreases when lesions are present, increasing the number of false positives. On the other hand, as shown in the FIRST segmentations in Figure 3.6 (f), we see that opposite to FreeSurfer, local lesions interfere with the segmentation performance. This can be seen for the right caudate (in yellow) where the green lesion shown in Figure 3.6 (c) is constraining the deformation performed by the method to obtain the final segmentation. The results obtained for the rest of the structures are similar for both images, since the generated lesions are far from the structures of interest. Finally, for the majority vote strategy, no changes are visually appreciated for this particular slice in Figure 3.6 (g). As shown, the differences in the segmentation performance of the healthy subject and the generated MS patients tend to be small for this method. Furthermore, the lesions shown in this slice do not necessaryly affect the performance of the structures shown here but may interfere in other parts of the brain, since, as stated before, in this strategy the segmentation performance oscillates independently of the lesion location.

3.4 Discussion

FreeSurfer is the most affected method when MS lesions are present. Despite being a method that deals with WM hypo-intensities (such as MS lesions), its segmentation performance varies significantly when MS lesions are introduced. This may be because this method is based on a Bayesian strategy that tries to infer the most likely segmentation, given the image intensities and prior information in the form of an atlas, and thus, adding MS lesions may affect this method in two different ways. First, the registration of the atlas priors can be affected by the lesions, as seen in the multi-atlas strategy. Moreover, the incorporation of the lesions modifies the image intensities, and consequently the intensity distribution of each structure, which is modeled as a Gaussian, may be affected and produce a different segmentation result.

The segmentation performance of the majority vote strategy is also affected when MS lesions are present. Similar to that of FreeSurfer, the performance of this method oscillates independently of the lesion location. As the atlas registration is performed globally, not only local lesions but also lesions in other parts of the brain have an effect on the registration result. On the other hand, although FIRST also performs registration to align the image and the model, it provides the most robust results. In this case, the method performs a local registration for which it uses a sub-cortical mask that determines whether a voxel is included or not in the calculation of the similarity function, thus allowing the registration to concentrate only on the sub-cortical alignment, and therefore, if the lesions are located outside the mask, they do not interfere with the registration result.

Regarding the effect of the lesion location, we have observed that FreeSurfer and

majority vote are, in general, equally affected whether the lesions are overlaid on the structure of analysis or are placed in other parts of the brain. As these methods provide segmentation for the whole brain, the most logical approach is to perform a unique and global registration, instead of trying to maximize the similarity for each structure independently by means of local registrations, as FIRST does with the group of sub-cortical structures. However, this global registration, despite being quicker, allows registration errors produced by brain irregularities such as WM lesions or tumors to be propagated to other parts of the brain image, and consequently affect the segmentation performance independently of the lesion location. On the other hand, the analyzed results show that in the case of FIRST, the lesion location has a direct effect on the method performance. In this case, the lesions that are overlaid on the structures worsen the segmentation result compared with that of the non-damaged structures, whereas the effects of lesions in other parts of the brain are inappreciable. It should come as no surprise that since FIRST is a locally deformable strategy and segments each structure independently, the shape and intensities of lesions far from the structure of interest do not interfere with the deformation process.

As for the method robustness, FIRST provides the most consistent segmentations. Since this method is based on a deformable strategy constrained by shape and intensity, deformations that exceed the average geometric variation of the structure are avoided, and therefore, only specific lesions attached to the structure may cause the method performance to oscillate. Furthermore, and contrary to the other two analyzed strategies, FIRST only segments the sub-cortical structures and works only with a small region of the brain instead of the whole volume.

Regarding the analyzed brain structures, in terms of DSC, the accumbens is the most affected by the presence of lesions, whereas the thalamus and the brainstem provide the most robust results when lesions are present. However, this behavior may be closely related to the fact that the accumbens is the smallest structure, whereas the brainstem and the thalamus are the largest ones. Thus, small changes in the segmentation result can imply large differences in the DSC when the structure volume is small while having an insignificant effect when the structure volume is large. On the other hand, the structure in which we find the largest number of lesions is the caudate nucleus, which makes sense since the vast majority of MS patients have at least one ovoid peri-ventricular lesion that augments the probability of this structure being affected. The large number of outliers found for this structure when lesions are overlaid can be explained by the number of cases. Although we see that around the fifty percent of the analyzed cases do not show an excessive DSC difference with respect to the healthy image, the segmentation result may have more chances of being affected due the variability in lesions' location and intensity (sometimes similar to that of the caudate) found in the different images. Furthermore, as the caudate nucleus is a small structure, a bigger effect is seen in

the DSC when a small change in the segmentation result is produced.

As we previously saw in Chapter 2, several brain structure segmentation methods following different segmentation strategies [55] have been described in recent years. Despite the large number of studies available in the literature, these approaches have been tested with images of non-lesioned brains, and therefore, how WM lesions affect their performance has not been evaluated. Performing such analysis with real images is not trivial, since there is no publicly available database with both structure ground truth and MS lesion annotation. Because of that, in an attempt to evaluate this effect, we generated a set of synthetic images as done in [29, 30, 248, 249] to evaluate the effect of the lesions on tissue segmentation. The effect of such lesions on automatic tissue segmentation strategies has been widely evaluated, and strategies to reduce this effect, such as masking out the lesions or fill them before segmentation, have been proposed [28, 29, 30].

In spite of the effect of MS lesions having been evaluated on several tissue segmentation methods, as far as we know only Gelineau-Morel et al. [248] have evaluated how these lesions affect the segmentation of the sub-cortical brain structures. In their work, they used FIRST as the baseline method for segmentation, concluding that WM lesions led to an artificial decrease in all the deep GM structure volumes except for the hippocampus. Corroborating their statement, a negative correlation between the lesion volume and the total deep GM volume was found in our experiments when segmenting with FIRST. Despite the experiments performed here not being the same and the fact that they dealt with right and left hemispheres combined, our results are consistent with their findings, except for the left thalamus (for which we saw an average volume increase compared to the healthy controls), and the left hippocampus (which experienced an average volume loss). Furthermore, when analyzing the deep GM structures separately we found that their volume differences (patient − healthy) did not have the same direction of correlation with the lesion volume, which is also consistent with the trend seen in their work [248].

In summary, in this chapter, we have presented an analysis regarding the effect of simulated MS lesions on three well-known automatic brain structure segmentation methods (FreeSurfer, FIRST, and multi-atlas fused by majority vote) based on different segmentation strategies from the ones introduced in Chapter 2. We have proved that MS lesions have a direct effect on the performance of automatic brain structure segmentation methods. FreeSurfer seems to be the most affected algorithm, whereas FIRST is the most robust when lesions are present. The lesions' location does not seem to have a direct effect on the global strategies (FreeSurfer and majority vote), whereas in FIRST, which is a local strategy, the segmentation performance of a brain structure is more affected when there are lesions either overlaid on, or close to the structure. To the best of our knowledge, how different segmentation strategies to automatically segment the brain structures are affected

by MS lesions had not been evaluated until now. This study addresses an important problem of the automatic segmentation methods of deep GM structures, which is related to MS lesion interference for the optimal segmentation. Investigating the influence of the lesions on the segmentation in other diseases is indeed an important aspect of future research for the community.

Chapter 4

A multi-atlas approach for brain structure segmentation in the presence of multiple sclerosis lesions

4.1 Introduction

As shown in previous chapters, among all the approaches proposed in the literature, multi-atlas methods have been proved to be robust and provide good segmentation results on healthy subjects [55, 146, 152]. As explained in Section 2.3.1, in this strategy, a set of manually segmented images (atlases) are non-rigidly registered to the target image. After that, the deformation fields obtained from these registrations are applied to the corresponding manual segmentations in such a way that new pairs of images (structural image and segmentation) are obtained, which are similar to the target. These candidate segmentations of the target are then fused to obtain the final segmentation. Several fusion strategies have been proposed in recent years [139, 151, 250, 251], the ones enhanced by the structural image intensities being the ones that provide the best results.

Multi-atlas label fusion strategies based on intensities exploit the target-atlas similarity under the assumption that images with similar appearance are more likely to have similar segmentations. A successful approach, first introduced by Coupé et al. [42], assumes that registration errors are frequently produced in multi-atlas seg-

mentation due to several factors such as the regularization constraints involved in that process, or the failure to reach a global optimum of the objective function. In order to overcome the registration errors, these methods relax the one-to-one mapping constraint existing in traditional weighting methods and re-compute the correspondences for every voxel / patch of the target image and the atlases before segmentation. This procedure is usually based on patch intensity similarity. However, in the same way as most of the proposed brain structure algorithms in the current state of the art [56], these strategies are designed to segment healthy subjects and, their performance tends to be affected by the presence of MS lesions, as we previously saw in Chapter 3.

Recently, various lesion filling approaches [29, 30, 31, 252] have been successfully proposed in order to minimize the effect of the abnormal MS lesion intensities on the segmentation. These techniques have been proved to improve tissue measurements [29, 30, 31, 252], and have also been used ad-hoc for brain structure segmentation [248, 253]. While the effect of lesion filling has been tested on several tissue segmentation strategies, in which intensity distributions of different tissue classes are modeled, it has not been studied to see how it affects patch based segmentation strategies, in which patch intensities are independent of the global intensity distributions.

In this chapter, we propose a new correspondence search approach for intensity-based multi-atlas label fusion, that can be applied to segment either healthy subjects or patients with MRI-visible lesions. We introduce a new voxel / patch correspondence model, able to deal with brain irregularities, such as MS lesions. Assuming no further improvement on the lesion masks based on intensities, we force the one-to-one correspondence obtained from masked non-rigid registration on those areas, while we also redefine the patch shape on the surroundings of the lesion to prevent the abnormal intensities from interfering in the correspondence search. We integrate this model into two well-known label fusion strategies (Non-local Spatial STAPLE [250] and Joint Label Fusion [251]), reformulating the original methods to improve the correspondences in the lesion areas, while maintaining the original search model in the rest of the brain.

4.2 The model and its integration

Inaccuracies in registration are a substantial source of error in multi-atlas label fusion. Registration is a very complex task, which is considered one of the fundamental problems in medical image processing, and may not always give maximum local similarity between image patches. Because of that, some fusion strategies have successfully tried to reduce these registration errors by means of local search win-

dows, that relax the traditional one-to-one mapping constraint. That is, given the patch centered at voxel i of the target image I ($\wp(I_i)$), it is often possible to find a patch $\wp(A_{i'j})$ centered at voxel i' of the atlas image j which is more similar to $\wp(I_i)$ than to the corresponding patch $\wp(A_{ij})$ centered at voxel i of the atlas image j. This similarity is often computed based on image intensities, which constitutes a challenge when the target image presents visible lesions. The lesions show different intensity profiles to those of healthy tissues in the atlas, making the correspondence search problem more difficult.

In this section, we present a new correspondence search approach for patch-based multi-atlas segmentation (label fusion), that can be applied to segment either healthy subjects or patients with MRI visible lesions. A graphical representation of this idea is depicted in Figure 4.1. In order to avoid the interference of the lesions during the correspondence search, we assume that such correspondence cannot be further improved inside the lesions based on intensities and enforce them on those areas in such a way that we give more weight to the masked registration result. Note that this premise could be applied to any intensity-based multi-atlas label fusion strategy, not only to solve for voxel / patch correspondences, but also to estimate the final voting weights, in weighted voting strategies, that lead to the segmentation result, as we will see in Section 4.2.3.

In the following, we reformulate two well-known label fusion strategies to include this idea: Non-local Spatial STAPLE [250], and Joint Label Fusion [251].

4.2.1 Problem definition

Consider a target gray-level image (with lesions) represented as a vector $I \in \mathbb{R}^{N\times 1}$. Let $T \in \mathcal{L}^{N\times 1}$ be the latent representation of the true target segmentation, where $\mathcal{L} = \{0, \ldots, L-1\}$ is the set of possible labels which can be assigned to a concrete voxel. Let $M \in \{0,1\}^{N\times 1}$ be a binary lesion mask indicating whether a given voxel i of the target image contains or is part of a lesion, hence $M_i = p\,(I_i \in lesion)$. Note that this mask is optional and can be neglected if all voxels in the mask are set to 0, i.e. non-lesion. Consider a set R of registered healthy atlases $\{A, D\}$ with associated gray level images, $A \in \mathbb{R}^{N\times R}$, and propagated label decisions, $D \in \mathcal{L}^{N\times R}$.

4.2.2 Masked Non-local Spatial STAPLE (m-NLSS)

Non-local Spatial STAPLE (NLSS) [250] is a variant of the STAPLE [65] algorithm from a non-local means perspective. As with other STAPLE-family fusion algorithms, NLSS computes simultaneously a probabilistic estimate of the true segmentation and a measure of performance level represented by each atlas segmentation.

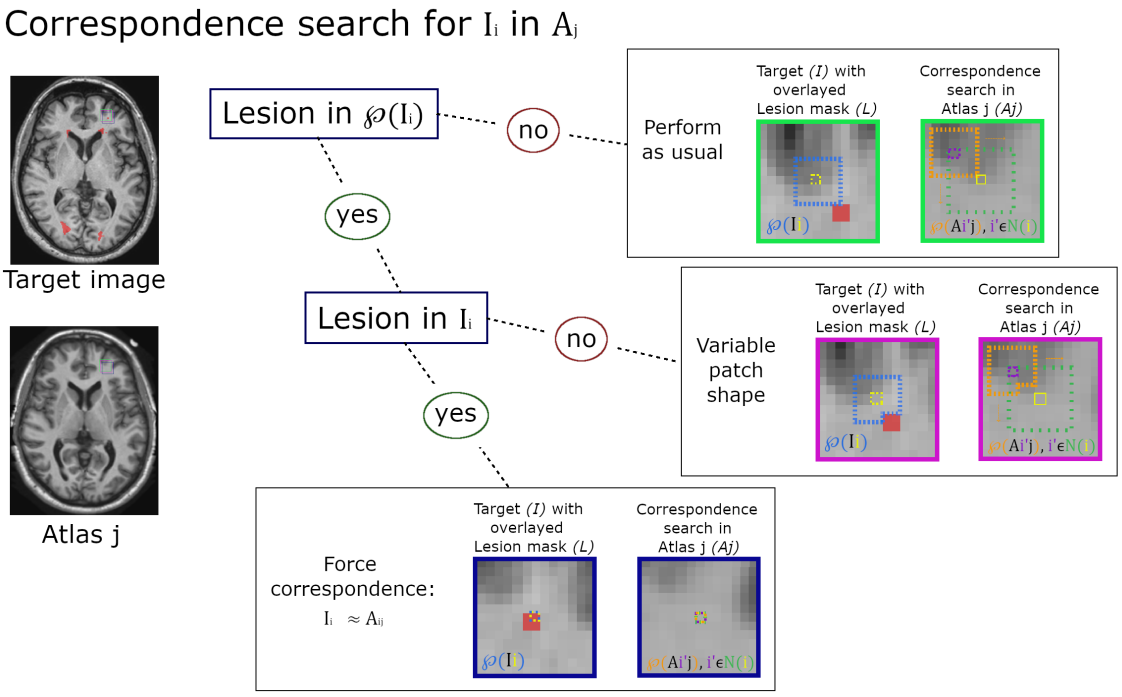


Figure 4.1: Correspondence search scheme. Search for the correspondence of voxel i of the target image I (I_i) on the atlas j (A_j). When there are not lesions in the patch of voxel i, our model performs as the original method (best correspondence is found comparing the target patch of voxel i, i.e. $\wp(I_i)$, to all the atlas patches of the voxels i' that belong to the neighborhood of i, i.e. $\wp(A_{i'j}), i' \in \mathcal{N}(i)$). On the other hand, when there are lesions in the patch, we modify the patch shape to exclude the lesions from the search and use the same patch shape to find the correspondence in the atlas in the same way as before. Finally, when the target voxel i is part of a lesion, we trust the masked registration result and force the correspondence to be A_{ij}. Note that this example is shown in 2D for simplification, where the patch size is set to 5×5 (blue and orange striped squares) and the search neighborhood to 7×7 (green striped square).

The goal of these algorithms is to select the performance parameters, such that they maximize the complete data log-likelihood, corresponding to the observed data and the un-observed latent true segmentation (T). Since T is not available, the performance parameters are estimated through EM framework. In traditional EM terminology, the underlying voxel-wise label probabilities represent the hidden data that is being estimated, and the performance level parameters, represent the hidden model parameters that help to determine the optimal solution for the target segmentation. The estimation of these parameters is accomplished by iterating between the estimation of the voxel-wise label probabilities (E-step) and the estimation of the

performance level parameters that maximize the expected value of the conditional log likelihood function (M-step).

NLSS incorporates the image intensities from both the atlas and the target to this segmentation framework using a patch-based non-local correspondence manner, to account for registration inaccuracies. This is done in such a way that the labels of all the atlas voxels in the neighborhood of the target voxel have a weight in its label assignment based on their intensity similarity. That is, they provide a model in which they learn which label each atlas would have observed given a perfect correspondence with the target and integrate this model into the STAPLE framework.

As stated before, lesion intensities may affect the result of this non-local correspondence model, leading to incorrect voxel correspondences that are sometimes even worse than the ones obtained by the one-to-one mapping resulting from masked registration. For this reason, following the previously stated assumption, we define the probability of correspondence between voxel i of the target image and voxel i' of the j-th atlas ($\alpha_{ji'i} \equiv p\left(A_{i'j}|I_i, M_i\right)$), i.e. the non-local correspondence model, as follows:

$$\alpha_{ji'i} = \frac{1}{Z_\alpha} \cdot e^{-\frac{\|\wp_{M_i} \circ \left(\wp\left(A_{i'j}\right) - \wp(I_i)\right)\|_2^2}{2 \cdot {\sigma_i}^2 \cdot \|\wp_{M_i}\|}} \cdot e^{-\frac{\varepsilon_{i'i}^2}{2 \cdot {\sigma_d}^2}} \cdot (1 - M_i) \; + \; \delta\left(i' = i\right) \cdot M_i \tag{4.1}$$

where $\wp(\cdot)$ is the set of intensities in the patch neighborhood of a given intensity location. In this definition, $\wp_{M_i} = \wp(1 - M_i)$ is the masking term that excludes lesion voxels from the patch calculation and enforces the same patch neighborhood size / shape in both the atlas and the target, $\|\wp_{M_i} \circ (\wp(A_{i'j} - \wp(I_i))\|_2^2$ is the L2-norm between the atlas patch centered at i' and the target patch centered at i, where $\circ$ is de Hadamard product. $\varepsilon_{i'i}^2$ is the Euclidean distance in physical space between i and i', σ_i and σ_d are the standard deviations for the intensity and distance weights, and Z_α is a partition function that enforces the constraint that $\sum_{i' \in \mathcal{N}(i)} \alpha_{ji'i} = 1$, where $\mathcal{N}(i)$ is the set of voxels in the search neighborhood of a given target voxel. $\delta(i' = i)$ is the Dirac delta function, and $\|\wp_{M_i}\|$ is the number of voxels in the patch neighborhood.

Now that the correspondence model is defined, we just have to integrate it into the segmentation algorithm. Therefore, in the same way as in the original approach, i.e. NLSS, $\theta \in [0, 1]^{R \times N \times L \times L}$ will be the performance level parameters of the raters (registered atlases), defined voxel-wise. Being each element of θ, $\theta_{jis's}$, the probability that rater j observes label s' given that the true label is s at a given voxel i and the corresponding voxel i^* on the associated atlas−i.e., $\theta_{jis's} = p\left(D_{i^*j} = s', A_j|T_i = s, I_i, M_i, \theta_{jis's}\right)$, where i^* is the voxel on atlas j that corresponds to the target voxel i.

If the exact voxel correspondences between the target and the atlases (non-local model) were known, the lesion mask, and the target and atlas intensity relationships could be ignored and the spatial STAPLE [254] definition of θ could be used.

$$\begin{aligned}\theta_{jis's} \equiv p\left(D_{i^*j} = s', A_j | T_i = s, I_i, M_i, \theta_{jis's}\right) \\ = p\left(D_{i^*j} = s' | T_i = s, M_i, \theta_{jis's}\right)\end{aligned} \tag{4.2}$$

However, this correspondence is not known and we have to learn it with the model defined above. Note that using this model we can approximate the relationship by taking the expected value of $p\left(D_{i^*j} = s', A_j | T_i = s, I_i, M_i, \theta_{jis's}\right)$ across the raters. Using an assumption of conditional independence between the labels, lesion mask and intensity, we approximate the density function as:

$$\begin{aligned}\mathrm{p}(D_{i^*j} = s', A_j | T_i = s, I_i, M_i, \theta_{jis's}) \approx E\left[p\left(D_j, A_j | T_i = s, I_i, M_i, \theta_{jis}\right)\right] \\ = E\left[p\left(D_j | T_i = s, M_i, \theta_{jis}\right) \cdot p\left(A_j | I_i, M_i\right)\right] \\ = \sum_{i' \in \mathcal{N}(i)} p\left(D_{i^*j} = s' | T_i = s, M_i, \theta_{jis's}\right) \cdot p\left(A_{i'j} | I_i, M_i\right) = \sum_{i' \in \mathcal{N}(i)} \theta_{jis's} \cdot \alpha_{ji'i}\end{aligned} \tag{4.3}$$

E-step

In this step, the weight variables $W^{(t)} \in \mathbb{R}^{\mathrm{L} \times \mathrm{N}}$ are derived from $\theta^{(t)}$, where $W^{(t)}_{si}$ represents the probability that the true label associated with voxel i is s at iteration t of the algorithm given the provided information and the performance level parameters.

$$W^{(t)}_{si} \equiv p\left(T_i = s | D, A, I, M, \theta^{(t)}\right) \tag{4.4}$$

Using Bayes' rule to separate the prior label probability ($p\left(T_i = s\right)$) and assuming independence among the raters, we can rewrite this equation as follows:

$$\mathrm{W}^{(t)}_{si} \equiv \frac{p(T_i = s) \cdot \prod_j p\left(D_{i^*j} = s', A_j | T_i = s, I_i, M_i, \theta^{(t)}_{jis's}\right)}{\sum_n p(T_i = n) \cdot \prod_j p\left(D_{i^*j} = s', A_j | T_i = n, I_i, M_i, \theta^{(t)}_{jis's}\right)} \tag{4.5}$$

Using the non-local correspondence model and the approximated density function, we obtain:

$$W^{(t)}_{si} \equiv \frac{p\left(T_i = s\right) \cdot \prod_j \sum_{i' \in \mathcal{N}(i)} \theta^{(t)}_{jis's} \cdot \alpha_{ji'i}}{\sum_n p\left(T_i = n\right) \cdot \prod_j \sum_{i' \in \mathcal{N}(i)} \theta^{(t)}_{jis's} \cdot \alpha_{ji'i}} \tag{4.6}$$

M-step

In this step, the estimated $W_{si}^{(t)}$ is used to update $\theta_{ji}^{(t+1)}$ by maximizing the expectation of the complete data log likelihood. As the complete data log likelihood is not observable, it is replaced by its conditional expectation given the observable data D, A, I, M using the current estimate θ.

$$\begin{aligned}
\theta_{ji}^{(t+1)} &= \underset{\theta_{ji}}{argmax} \sum_{i' \in \mathcal{B}_i} E\left[\ln\left(p\left(D_j, A_j | T_{i'}, I_{i'}, M_{i'}, \theta_{ji} | D, A, I, M, \theta^{(t)}\right)\right)\right] \\
&= \underset{\theta_{ji}}{argmax} \sum_{i' \in \mathcal{B}_i} \sum_{s} p\left(T_{i'} = s | D, A, I, M, \theta^{(t)}\right) \cdot \ln\left(p\left(D_j, A_j | T_{i'}, I_{i'}, M_{i'}, \theta_{ji}\right)\right) \\
&= \underset{\theta_{ji}}{argmax} \sum_{i' \in \mathcal{B}_i} \sum_{s} W_{si'}^{(t)} \cdot \ln\left(p\left(D_{i^*j} = s', A_j | T_{i'}, I_{i'}, M_{i'}, \theta_{ji}\right)\right) \\
&= \underset{\theta_{ji}}{argmax} \sum_{i' \in \mathcal{B}_i} \sum_{s} W_{si'}^{(t)} \cdot \ln\left(p\left(D_{i^*j} = s', A_j | T_{i'}, I_{i'}, M_{i'} = 0, \theta_{ji}\right)\right. \\
&\qquad \left. + p\left(D_{i^*j} = s', A_j | T_{i'}, I_{i'}, M_{i'} = 1, \theta_{ji}\right)\right) \\
&= \underset{\theta_{ji}}{argmax} \sum_{i' \in \mathcal{B}_i} \sum_{s} W_{si'}^{(t)} \cdot \ln\left(\sum_{i'' \in \mathcal{N}(i'): D_{i''j} = s'} \theta_{jis's} \cdot \alpha_{ji''i'}\right)
\end{aligned} \tag{4.7}$$

As each row of the rater performance level parameters, i.e. θ, must sum one to be a valid probability mass function, we can maximize the performance level parameters for each element by using a Lagrange multiplier (λ) to formulate the constrained optimization problem.

$$0 = \frac{\delta}{\delta \theta_{jin'n}} \left[\sum_{i' \in \mathcal{B}_i} \sum_{s} W_{si'}^{(t)} \cdot \ln\left(\sum_{i'' \in \mathcal{N}(i'): D_{i''j} = s'} \theta_{jis's} \cdot \alpha_{ji''i'}\right) + \lambda \sum_{s'} \theta_{jis's}^{(t+1)} \right] \tag{4.8}$$

By solving this equation, we obtain

$$\theta_{jis's}^{(t+1)} = \frac{\sum_{i' \in \mathcal{B}_i} \left(\sum_{i'' \in \mathcal{N}(i'): D_{i''j} = s'} \alpha_{ji''i'}\right) \cdot W_{si'}^{(t)}}{\sum_{i' \in \mathcal{B}_i} W_{si'}^{(t)}} \tag{4.9}$$

As we can observe, the obtained EM equations remain the same as in the original algorithm.

4.2.3 Masked Joint Label Fusion (m-JLF)

Joint Label Fusion (JLF) [251] is based on the idea that different atlases may produce similar label errors. They assume that the errors produced by the atlases are not independent and address this issue by computing the intensity similarity between the target and each pair of atlases. This, allows them to estimate the probability that a pair of atlases produce the same segmentation error.

In order to estimate this pairwise dependency matrix (P), the authors first solve for the patch correspondences between the target and each atlas, as they also assume registration errors. As these correspondences are computed based on patch intensities, we know that the lesion intensities may interfere on this local patch search, and therefore we redefine the local search correspondence map between the atlas j and the target as follows:

$$i'_{ij} = \underset{i' \in \mathcal{N}(i)}{argmin} \left[\| \wp_{M_i} \circ (\wp(A_{i'j}) - \wp(I_i)) \|^2 \cdot (1 - M_i) \; + \; \delta(i' = i) \cdot M_i \; \right] \tag{4.10}$$

where $\wp_{M_i}$ is the same masking term than in Section 4.2.2, $\| \wp_{M_i} \circ (\wp(A_{i'j}) - \wp(I_i)) \|^2$ is the (masked) sum of squared differences of the non-lesion voxels, $\mathcal{N}(i)$ is the set of voxels in the search neighborhood of a given target voxel, and $\delta(i' = i)$ the Dirac delta function.

In the same way that lesions interfere in the local patch search, they also have to be modeled in the pairwise dependency matrix, P, which estimates how likely two atlases are both to produce wrong segmentation for the target image, given the observed joint patch intensity differences. Following our assumption, we estimate the matrix of expected pairwise joint label differences between the j_1-th and j_2-th atlases, $P_i(j_1, j_2)$, as follows:

$$\begin{aligned} P_i(j_1, j_2) &= p\left(\gamma_i{}^{j_1} \cdot \gamma_i{}^{j_2} = 1 | I, A, M\right) \\ &= p\left(\gamma_i{}^{j_1} \cdot \gamma_i{}^{j_2} = 1 | I, A_{j_1}, A_{j_2}, M\right) \end{aligned} \tag{4.11}$$

where $\gamma_i{}^j$ is the label difference between the j-th atlas and the target image at voxel i. Note that the product $\gamma_i{}^{j_1} \cdot \gamma_i{}^{j_2}$ can only take values 0 or 1, and $\gamma_i{}^{j_1} \cdot \gamma_i{}^{j_2} = 1$ if and only if both atlases produce a label different from the target segmentation.

As in the original approach, if we assume that given the image patches centered around the location in consideration, the pairwise joint label difference term is conditionally independent from distant voxels, we have:

$$P_i(j_1, j_2) = p\left(\gamma_i{}^{j_1} \cdot \gamma_i{}^{j_2} = 1 | \wp(I_i), \wp\left(A_{i'_{ij_1} j_1}\right), \wp\left(A_{i'_{ij_2} j_2}\right), \wp(M_i)\right) \tag{4.12}$$

where $\wp(\cdot)$ is the set of intensities in the patch neighborhood of a given intensity location, with I_i being the target voxel i, M_i the lesion mask voxel i, $A_{i'_{ij_1}j_1}$ the corresponding voxel i'_{ij_1} to target voxel i on atlas j_1 and $A_{i'_{ij_2}j_2}$ the corresponding voxel i'_{ij_2} to target voxel i on atlas j_2.

To predict the label difference between the target and the atlases, the authors used local image information between the two images, adapting the inverse distance function, to estimate the probability of pairwise joint label difference. However, in order to avoid the target lesion intensities interfering in the label difference prediction, we have reformulated this equation as follows:

$$P_i\left(j_1, j_2\right) \propto \left[\frac{\left(\wp_{M_i} \circ \left|\wp\left(I_i\right) - \wp\left(A_{i'_{ij_1}j_1}\right)\right|\right) \cdot \left(\wp_{M_i} \circ \left|\wp\left(I_i\right) - \wp\left(A_{i'_{ij_2}j_2}\right)\right|\right)}{\left\|\wp_{M_i}\right\|}\right]^{\beta} \tag{4.13}$$

where β is a model parameter controlling the weight distribution, $\|\wp_{M_i}\|$ is the number of non-lesion voxels in the patch neighborhood, $\circ$ is the Hadamard product, and, $A_{i'_{ij_1}j_1}$ and $A_{i'_{ij_2}j_2}$ are the corresponding voxels of target image voxel i, i.e. I_i, on atlases j_1 and j_2, respectively.

Then, the voxel voting weights for each atlas can be estimated from $P_i\left(j_1, j_2\right)$, in the same way as in the original formulation [251].

$$w_i = \frac{{P_i}^{-1} \cdot 1_R}{{1_R}^t \cdot {P_i}^{-1} \cdot 1_R} \tag{4.14}$$

where $1_R = [1, 1, \ldots, 1]^t$ is a vector of size R, i.e. number of atlases,
$w_i = [w_{ij_1}, w_{ij_2}, \ldots, w_{ij_R}]^t$, and $\sum_j w_{ij} = 1$. In this definition, t stands for transpose.

Finally, we define the weights that lead to the final segmentation as follows:

$$\mathrm{W}_{si} = p\left(T_i = s | D, A, I, M\right) \tag{4.15}$$

where W_{si} represents the probability that the true label associated with voxel i is s.

The authors apply an average scheme to obtain the final weights W from w, in such a way that the voting weights $w_{i'}$ of all the voxels in the patch of i, i.e. $i' \in \mathcal{N}_p(i)$, contribute equally to W_i. However, applying the same averaging scheme would allow lesion voxels to contribute to the healthy voxels segmentation. For this reason, we redefine W as follows:

$$\mathrm{W}_{si} = (1 - M_i) \cdot \left(\sum_{i' \in \mathcal{N}_p(i): M_{i'}=0} \; \sum_{j: D_{i_{i''_{ij}}j} = s} w_{i'j}\right) + M_i \cdot \left(\sum_{j: D_{i_{i''_{ij}}j} = s} \frac{||\mathcal{N}_p||}{R}\right) \tag{4.16}$$

where $D_{i_{i''_{ij}}j}$ is the decision of the atlas j on the corresponding voxel i''_{ij} for the target voxel i', $\mathcal{N}_p(i)$ is the patch neighborhood of the voxel i, $||\mathcal{N}_p||$ is the number of voxels in the patch neighborhood, and R the number of atlases. Note that with this formulation we avoid propagating the registration errors produced in the lesion areas to the healthy voxels, whereas we apply a traditional majority vote scheme on the lesions.

4.3 Experiments and evaluation

4.3.1 Data

Public databases of patients with lesions and including both brain parcellation and lesion annotations are uncommon. Therefore, to test our approach in a large number of cases, we decided to run the experiments with simulated data. Specifically, we have simulated artificial MS lesions on a database of 45 healthy patients that included brain parcellation annotations. To simulate the lesions, 140 MS patients from five different databases (including MICCAI'08 [245], MICCAI'16 [244], ISBI'15 [246], and two in-house databases) were analyzed, and the 45 patients with larger lesion volume were selected as basis for simulation. The selected 45 patients were paired with the annotated healthy images based on their lateral ventricle size in order to pair similarly atrophied brains. Once the couples were assigned, each MS patient image was non-rigidly registered to its corresponding atlas [255] (previous initial affine registration [256]), masking out the lesion areas for more adjusted registration. Then, the normalized intensities of the registered lesions were copied to the atlases, obtaining a new database of simulated patients with lesion volumes ranging from $1.16-68.43\ ml$, and voxel spacing $1 \times 1 \times 1\ mm$.

The healthy subject dataset consists of 45 T1-w MR images obtained from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database [229]. The images were obtained from Open Access Series on Imaging Studies (OASIS) dataset [223] and labeled according to BrainCOLOR protocol [257], including 133 labels that cover the whole brain: sub-cortical structures, ventricles, cerebral WM, cerebellum, brainstem and 98 regions in the cortex (see Table 4.1 for more information).

Nevertheless, the recent MRBrains 2018 challenge dataset [258] allows us to test our approach also with real data. Although this dataset consists of 30 MRI images obtained from patients with varying degrees of atrophy and WM lesions, only 7 cases have been released with both lesion and tissue segmentation available. For these 7 patients, lesion load ranges from 0.06 to 70.00 ml. As we use the atlases from the previous dataset, the 133 brain structures labels have been combined to

Table 4.1: Labels from the MICCAI 2012 grand challenge and workshop on multi-atlas labeling database [229]

Label	Structure	Label	Structure
0	Background	138	Right MCgG middle cingulate gyrus
4	3rd Ventricle	139	Left MCgG middle cingulate gyrus
11	4th Ventricle	140	Right MFC medial frontal cortex
23	Right Accumbens Area	141	Left MFC medial frontal cortex
30	Left Accumbens Area	142	Right MFG middle frontal gyrus
31	Right Amygdala	143	Left MFG middle frontal gyrus
32	Left Amygdala	144	Right MOG middle occipital gyrus
35	Brain Stem	145	Left MOG middle occipital gyrus
36	Right Caudate	146	Right MOrG medial orbital gyrus
37	Left Caudate	147	Left MOrG medial orbital gyrus
38	Right Cerebellum Exterior	148	Right MPoG postcentral gyrus medial segment
39	Left Cerebellum Exterior	149	Left MPoG postcentral gyrus medial segment
40	Right Cerebellum White Matter	150	Right MPrG precentral gyrus medial segment
41	Left Cerebellum White Matter	151	Left MPrG precentral gyrus medial segment
44	Right Cerebral White Matter	152	Right MSFG superior frontal gyrus medial segment
45	Left Cerebral White Matter	153	Left MSFG superior frontal gyrus medial segment
47	Right Hippocampus	154	Right MTG middle temporal gyrus
48	Left Hippocampus	155	Left MTG middle temporal gyrus
49	Right Inf Lat Vent	156	Right OCP occipital pole
50	Left Inf Lat Vent	157	Left OCP occipital pole
51	Right Lateral Ventricle	160	Right OFuG occipital fusiform gyrus
52	Left Lateral Ventricle	161	Left OFuG occipital fusiform gyrus
55	Right Pallidum	162	Right OpIFG opercular part of the inferior frontal gyrus
56	Left Pallidum	163	Left OpIFG opercular part of the inferior frontal gyrus
57	Right Putamen	164	Right OrIFG orbital part of the inferior frontal gyrus
58	Left Putamen	165	Left OrIFG orbital part of the inferior frontal gyrus
59	Right Thalamus Proper	166	Right PCgG posterior cingulate gyrus
60	Left Thalamus Proper	167	Left PCgG posterior cingulate gyrus
61	Right Ventral DC	168	Right PCu precuneus
62	Left Ventral DC	169	Left PCu precuneus
71	Cerebellar Vermal Lobules I-V	170	Right PHG parahippocampal gyrus
72	Cerebellar Vermal Lobules VI-VII	171	Left PHG parahippocampal gyrus
73	Cerebellar Vermal Lobules VIII-X	172	Right PIns posterior insula
75	Left Basal Forebrain	173	Left PIns posterior insula
76	Right Basal Forebrain	174	Right PO parietal operculum
100	Right ACgG anterior cingulate gyrus	175	Left PO parietal operculum
101	Left ACgG anterior cingulate gyrus	176	Right PoG postcentral gyrus
102	Right AIns anterior insula	177	Left PoG postcentral gyrus
103	Left AIns anterior insula	178	Right POrG posterior orbital gyrus
104	Right AOrG anterior orbital gyrus	179	Left POrG posterior orbital gyrus
105	Left AOrG anterior orbital gyrus	180	Right PP planum polare
106	Right AnG angular gyrus	181	Left PP planum polare
107	Left AnG angular gyrus	182	Right PrG precentral gyrus
108	Right Calc calcarine cortex	183	Left PrG precentral gyrus
109	Left Calc calcarine cortex	184	Right PT planum temporale
112	Right CO central operculum	185	Left PT planum temporale
113	Left CO central operculum	186	Right SCA subcallosal area
114	Right Cun cuneus	187	Left SCA subcallosal area
115	Left Cun cuneus	190	Right SFG superior frontal gyrus
116	Right Ent entorhinal area	191	Left SFG superior frontal gyrus
117	Left Ent entorhinal area	192	Right SMC supplementary motor cortex
118	Right FO frontal operculum	193	Left SMC supplementary motor cortex
119	Left FO frontal operculum	194	Right SMG supramarginal gyrus
120	Right FRP frontal pole	195	Left SMG supramarginal gyrus
121	Left FRP frontal pole	196	Right SOG superior occipital gyrus
122	Right FuG fusiform gyrus	197	Left SOG superior occipital gyrus
123	Left FuG fusiform gyrus	198	Right SPL superior parietal lobule
124	Right GRe gyrus rectus	199	Left SPL superior parietal lobule

Table 4.1 continued from previous page

Label	Structure	Label	Structure
125	Left GRe gyrus rectus	200	Right STG superior temporal gyrus
128	Right IOG inferior occipital gyrus	201	Left STG superior temporal gyrus
129	Left IOG inferior occipital gyrus	202	Right TMP temporal pole
132	Right ITG inferior temporal gyrus	203	Left TMP temporal pole
133	Left ITG inferior temporal gyrus	204	Right TrIFG triangular part of the inferior frontal gyrus
134	Right LiG lingual gyrus	205	Left TrIFG triangular part of the inferior frontal gyrus
135	Left LiG lingual gyrus	206	Right TTG transverse temporal gyrus
136	Right LOrG lateral orbital gyrus	207	Left TTG transverse temporal gyrus
137	Left LOrG lateral orbital gyrus		

obtain an 8-class segmentation: background, cortical GM, basal ganglia, WM, CSF in the extracerebral space, ventricles, cerebellum and brainstem. Voxel spacing for the images in this dataset is $0.9583 \times 0.9583 \times 3\ mm$.

Severely atrophied brains were present in both databases. In the first one, in spite of lesions were simulated on a cohort of healthy subjects, their ages range from 18 up to 90 years old, and therefore, some of them present age-related atrophy.

4.3.2 Pre-processing

The atlases used in our experiments include the 45 images from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database [229]. For the atlas registration, PCA atlas selection was performed and only the 15 most similar atlases were used for segmentation. In the experiments performed on the simulated dataset, since lesions were simulated in images of the same database, the one being analyzed was excluded from the atlas selection, keeping as candidate atlases the remaining 44 images. All the images were histogram normalized and N4 [259] bias field corrected before registration. All the pairwise registrations were performed using an initial affine registration [256] followed by a non-rigid [255] procedure. In all the registrations performed, the lesions were masked-out to avoid their intensities to interfere in the similarity metric calculation. For a fair comparison of the fusion methods, the same registration results were used for all the strategies.

4.3.3 Initialization and priors

The voxel-wise prior $p\left(T_i = s\right)$ in m-NLSS was initialized using the weak log-odds majority vote, as in NLSS. The performance parameters, $\theta_{jis's}$, were initialized assuming each atlas has high performance as: 0.95, if $s = s'$; and $\frac{0.05}{L-1}$, otherwise. Algorithm convergence was detected when the average change in the diagonal elements of θ is below 10^{-4}.

For a fair comparison of the fusion methods, all the parameters in our experiments were set to the same values in all the original and proposed methods. The search

neighborhood $\mathcal{N}(\cdot)$ was set to $7 \times 7 \times 7$, patch $\wp(\cdot)$ dimensions to $5 \times 5 \times 5$ and σ_i, σ_d and β were set to 0.25, 1.5 and 2, respectively.

4.3.4 Evaluation

To evaluate the usefulness of our strategy, in the first experiment, we compare the performances of our approaches (m-NLSS and m-JLF) with respect to the ones obtained by the original algorithms (NLSS and JLF) when testing the following cases: (1) original MRI images of healthy subjects (H), (2) images with synthetic MS lesions (L), and (3) images with synthetic MS lesions after applying a recent lesion filling algorithm [31] (F). Notice that using our proposed strategy, the intensities inside the lesion mask are not relevant and, thus, in the three cases, we obtained the same result. In contrast, the performance of the original algorithms varied in each case (obtaining in what follows, NLSS(H), NLSS(L), and NLSS(F), respectively, and similarly for the JLF algorithm). Figure 4.2 shows the evaluated cases and the corresponding nomenclature.

The lesion filling technique used here [31] replaces the lesion voxel intensities by random values of a normal distribution generated from the mean WM signal intensity of each two-dimensional slice. As stated by their authors, this technique is a compromise between global and local methods, reducing the bias caused by refilled voxels on GM and WM tissue distributions by means of global information from the whole slice, whereas aims to reproduce more precisely the signal variability between slices by means of re-computing the mean signal intensity of the normal-appearing WM at each slice.

In a second experiment, we tested our approaches in the 7 cases of the MRBrains18 dataset. In this case, we compared the proposed correspondence algorithms (m-NLSS and m-JLF) when segmenting the original images (including lesions) to their originals when segmenting: (1) the original images (NLSS(L) and JLF(L)) and (2) the images after applying the lesion filling algorithm [31] (NLSS(F) and JLF(F)). See Figure 4.2 for details.

We quantitatively evaluate the segmentation results using the global DSC across all the structures affected by lesions:

$$\mathrm{DSC}_{global} = \frac{2 \cdot \sum_{l \in L} |T_l \bigcap E_l|}{\sum_{l \in L} |T_l| + \sum_{l \in L} |E_l|} \tag{4.17}$$

where T_l is the ground truth segmentation for label l, E_l is the estimate for the label l, and L is the set of all the available labels.

As the presence of lesions not necessarily affect only the lesion area segmentation

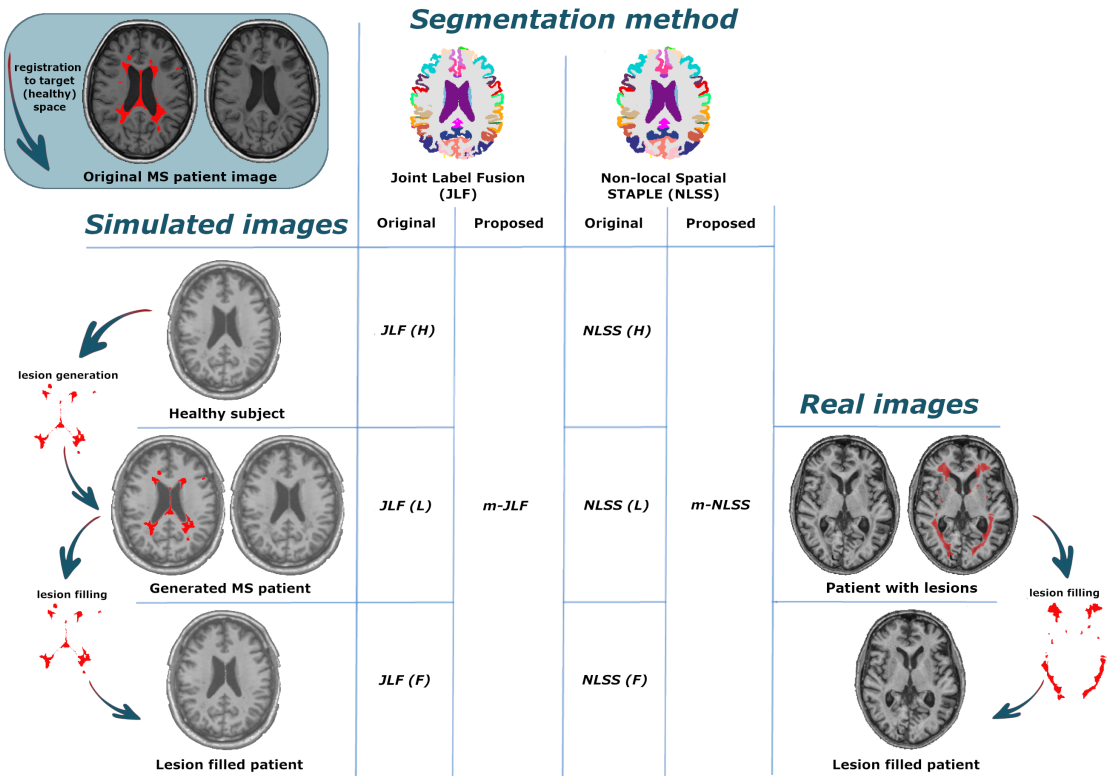


Figure 4.2: Evaluation procedure for both the simulated and the real databases. In the simulated database, the original MS patient is registered to the healthy space and the intensities of the registered lesions are copied to the healthy subject. For evaluation, each of the images shown is segmented with the original methods (NLSS and JLF) and the proposed ones (m-NLSS and m-JLF). The performance of the methods is assessed individually for each segmented image, i.e. healthy / lesioned / filled for the simulated database and lesioned / filled for the real database, based on the DSC difference of the proposed method and the original one. Note that the segmentation result when using the proposed method on the healthy, lesioned and filled images will be the same, i.e. the intensities inside the lesion mask are irrelevant for our method and the atlas registration results are the same for all the images analyzed, thus, we only segment the image with lesions.

itself, but also the surrounding tissues, two measures are calculated: (1) DSC inside the lesion mask and, (2) DSC inside a mask that includes three voxels of the lesion contour. Note that $\mathcal{N}(\cdot)$ was set to $7 \times 7 \times 7$. Besides, to give an overview of the global performance of the strategies, and to evaluate how the lesions affect the segmentation of the whole brain, we also compute the mean DSC across all the structures.

To provide a better comparison between the methods, DSC differences are shown instead of the DSC itself, being the difference of the DSC obtained using our strategy and the original method. Notice that these differences are computed subject by subject, which relates the performance for the same subject with respect to the proposed methods and their original ones.

Statistical analysis is performed using the Matlab software package. Differences in the performance of the analyzed methods are computed using paired-sample t-tests. Moreover, the Pearson's linear correlation coefficient is used to compute the correlation between the total lesion volume and the changes in mean DSC.

4.4 Results

4.4.1 Simulated MS lesions

First, we perform an analysis of the segmentation results on the lesion areas. Figure 4.3 shows the global DSC differences between the proposed strategies when segmenting the images with lesions (m-NLSS and m-JLF) and the original methods when segmenting: (1) the images with lesions (NLSS(L) and JLF(L)) and, (2) the lesion filled images (NLSS(F) and JLF(F)) (see Figure 4.2 for setup details). Notice that each boxplot represents the subtraction of the original method performance from our method's. Hence, positive values indicate an improvement of our proposal with respect to the depicted method. Furthermore, each boxplot significance was assessed independently and represents the relationship with the proposed strategy, and therefore, they are independent to each other. Differences in performances were assessed by means of paired t-tests between the original strategies (shown in Figure 4.3) and the proposed ones. The results have been analyzed separately inside the lesion mask and on a region that includes three voxels of the lesion mask contour. As observed in the figure, inside the lesion masks the proposed correspondence models (m-NLSS and m-JLF) performed significantly better than their originals (NLSS(L) and JLF(L)) when the lesions were present (both inside and around the lesion masks). On the other hand, analyzing the segmentation of the filled images with the original methods (NLSS(F) and JLF(F)), we observed that inside the lesion masks the performance was similar to that of our proposal, while around the lesion areas our strategy performed significantly better.

We also compared our proposals to the best possible segmentation the original algorithms could reach, i.e. segmenting the corresponding healthy subjects. We observed from that experiment that, inside the lesion areas, either segmenting the simulated images (with lesions) with the proposed methods (m-NLSS and m-JLF) or filling the lesions before segmentation (NLSS(F) and JLF(F)) did not reach the

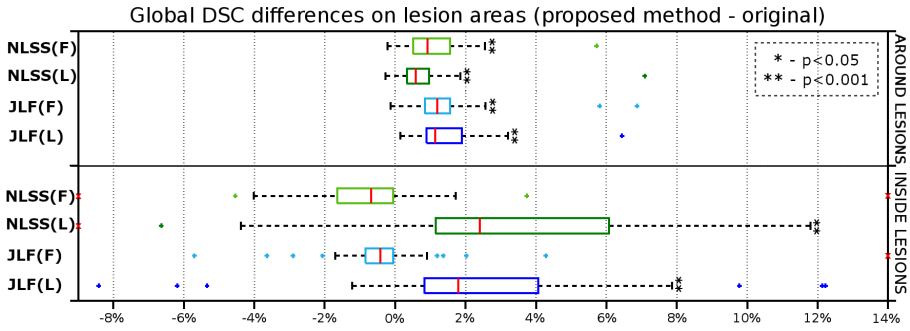


Figure 4.3: Global DSC differences on the MS simulated database. Differences between the proposed strategies (m-NLSS / m-JLF) and their corresponding original methods on the lesion areas: inside the lesion masks and on a mask that includes three voxels of the lesion mask contour. Segmentation differences performed for (1) the lesion filled generated patients (m-NLSS − NLSS(F) and m-JLF − JLF(F)) and, (2) the simulated patients (m-NLSS − NLSS(L) and m-JLF − JLF(L)). Statistical significance assessed independently for each boxplot in the figure, which represents the relationship between the proposed strategy and the original method. By means of paired t-tests, we test the null hypothesis that the true mean DSC difference between both methods (proposed and depicted) is zero.

healthy segmentation performance. This behavior was expected, since the intensities of that areas were "corrupted" both in the image with lesions and in the filled one. However, analyzing the performance around the lesion areas, the proposed methods (m-NLSS and m-JLF) reached similar performance to that of the healthy segmentations (NLSS(H) and JLF(H)). On the other hand, when it comes to the lesion-filled images, both strategies (NLSS(F) and JLF(F)) significantly under-performed the healthy segmentation (NLSS(H) and JLF(H)). Besides this, both original methods (NLSS(L) and JLF(L)) significantly under-performed the best possible segmentations (NLSS(H) and JLF(H)).

In terms of whole brain segmentation performance (not only restricted to the lesion areas), our proposals (m-NLSS and m-JLF) provided significantly better segmentation results than the original methods when segmenting the simulated images (NLSS(L) and JLF(L)) and the filled images (NLSS(F) and JLF(F)), as observed in Figure 4.4. However, whereas the whole brain segmentation performance of our NLSS non-local model (m-NLSS) was similar to that of the original method when segmenting the healthy subjects (NLSS(H)), our proposal for JLF (m-JLF) did not reach the best possible performance of the original method (JLF(H)), which is comprehensible since the real intensity information of the lesions is missing in the simulated images.

To give an overview of the whole brain segmentation performance, the mean

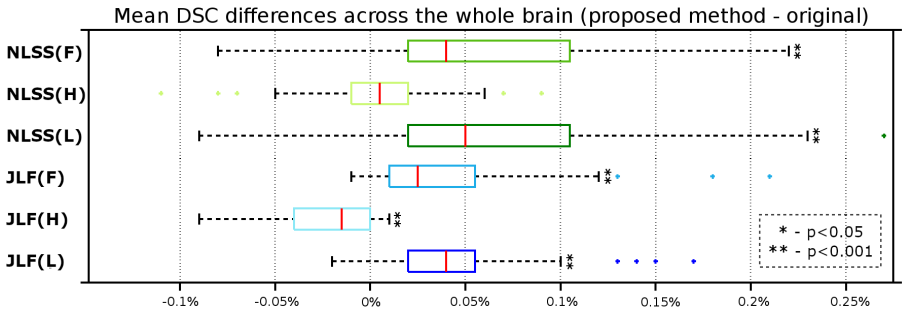


Figure 4.4: Mean DSC differences on the MS simulated database. Differences for the whole brain between the proposed strategies (m-NLSS / m-JLF) and their corresponding original method. Segmentation performed for (1) the lesion-filled generated patients (NLSS(F) / JLF(F)), (2) the healthy subjects (NLSS(H) / JLF(H)) and, (3) the simulated patients (NLSS(L) / JLF(L)). Statistical significance assessed independently for each boxplot in the figure.

DSC achieved by the analyzed methods was: 79.27 ± 1.35 (NLSS(L)), 79.35 ± 1.34 (m-NLSS), 79.35 ± 1.33 (NLSS(H)), and 79.29 ± 1.34 (NLSS(F)), 85.97 ± 1.47 (JLF(L)), 86.03 ± 1.46 (m-JLF), 86.05 ± 1.47 (JLF(H)), 85.99 ± 1.46 (JLF(F)).

In order to see where the bigger DSC changes, due to lesions, occurred within the brain, we performed the same analysis on the sub-cortical and cortical GM, and the WM separately. To do such analysis, the resulting labels were merged before computing the DSC in three groups: (1) cortical labels, (2) left and right cerebral WM, and (3) sub-cortical structures (both thalami, putamens, pallidums, caudates, amygdalas, hippocampi and accumbens).

This second experiment, showed that in both methods, i.e. NLSS and JLF, the structure which experimented more performance variance was the WM, when comparing between our proposal (m-NLSS / m-JLF) and the original method segmenting: (1) the healthy images (NLSS(H) / JLF(H)) and, (2) the simulated images (NLSS(L) / JLF(L)). On the other hand, when checking for differences between our strategy (m-NLSS / m-JLF) and the original method segmenting the filled images (NLSS(F) / JLF(F)), we observed that the GM was more affected than the WM, in particular the sub-cortical structures, where more DSC variance was appreciated. In light of these findings, we see that lesion filling helps in achieving more accurate results than just segmenting the un-pre-processed image, however, the improvement is more visible on the WM than on the GM structures. The results of this analysis are depicted in Figure 4.5.

Figure 4.6 shows some qualitative results obtained with the analyzed methods. As can be observed from this figure, when the lesions are close to the GM, the original methods ((f) and (j)) tend to segment them as part of this tissue. On the

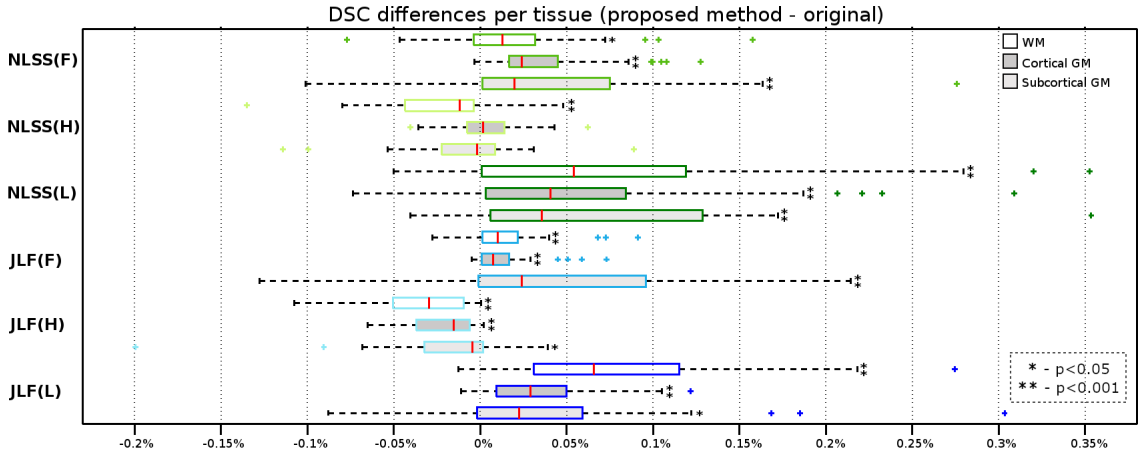


Figure 4.5: Dice differences on the MS simulated database. Differences for the gray matter (GM) (cortical and sub-cortical) and the white matter (WM) between the proposed strategies (m-NLSS / m-JLF) and their corresponding original method. Segmentation performed for (1) the lesion-filled generated patients (NLSS(F) / JLF(F)), (2) the healthy subjects (NLSS(H) / JLF(H)) and, (3) the simulated patients (NLSS(L) / JLF(L)). Statistical significance assessed independently for each boxplot in the figure.

other hand, if the lesions are filled before segmentation ((h) and (l)), GM structures tend to be under-estimated. These two issues seem to be handled correctly by our proposals ((g) and (k)), which results look more similar to the healthy subjects segmentation ((e) and (i)) and the ground truth.

Lastly, we analyzed the extent to which total lesion volume affected the observed changes in DSC for the evaluated methods. Significant correlations were found on the DSC differences of the whole brain between NLSS(L) and m-NLSS ($r = 0.86$, $p < 0.001$), NLSS(F) and m-NLSS ($r = 0.63$, $p < 0.001$), JLF(L) and m-JLF ($r = 0.72$, $p < 0.001$), JLF(H) and m-JLF ($r = -0.50$, $p < 0.001$), and between JLF(F) and m-JLF ($r = 0.57$, $p < 0.001$). However, no correlation with the lesion load was found on the DSC changes between NLSS(H) and m-NLSS. On the other hand, when analyzing the connection between the total lesion load and the performance differences of segmenting the simulated (L) and the lesion-filled (F) images, correlations were found for NLSS ($r = 0.67$, $p < 0.001$), and JLF ($r = 0.52$, $p < 0.001$). A more exhausted analysis, separated by tissue, i.e. sub-cortical structures, cortical GM and WM, is presented in Table 4.2.

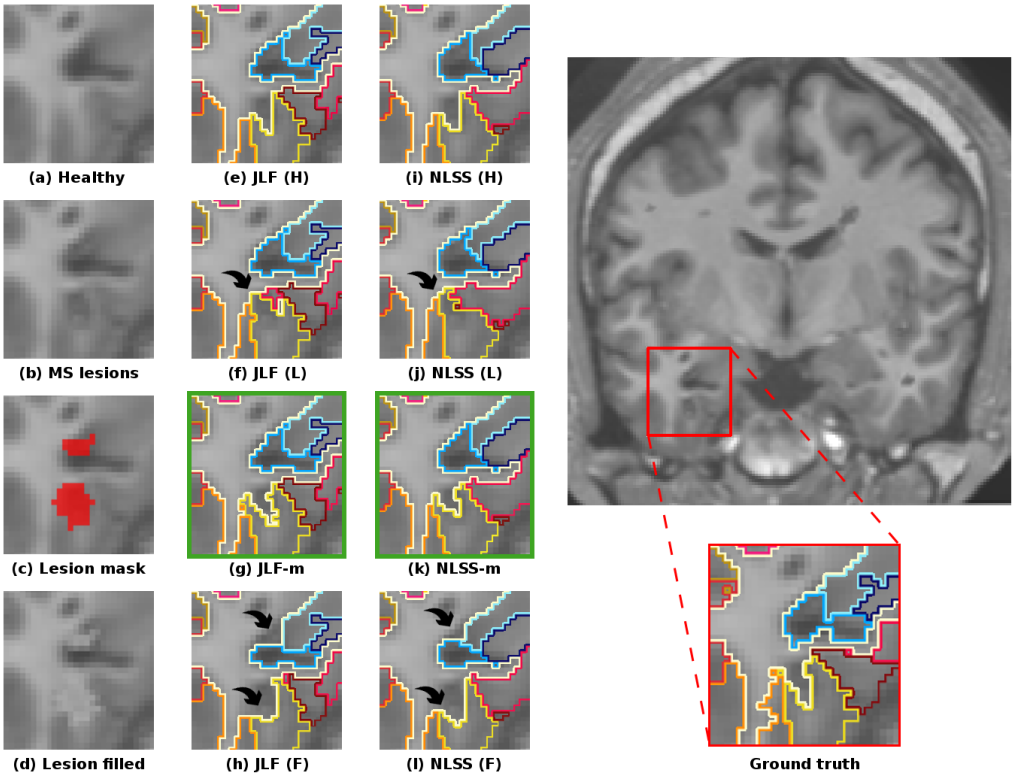


Figure 4.6: Structural images and segmentation results obtained for the analyzed cases. (a) Original healthy T1-w image, (b) simulated MS lesions on the healthy T1-w image, (c) lesion mask, (d) T1-w image after filling the lesions. Segmentation results of JLF on (e) the healthy subject, (f) simulated MS patient, and (h) lesion-filled image; and NLSS on (i) the healthy subject, (j) simulated MS patient, and (l) lesion-filled image. Proposed strategies for (g) JLF and, (k) NLSS are highlighted in green.

4.4.2 MRBrainS 2018 Challenge

On this database, the experiments performed showed that, in terms of global DSC differences, the modified correspondence models provided, in average, better segmentation results on the lesion areas, for both NLSS and JLF methods. On the other hand, analyzing the performance of the methods inside and around the lesion masks separately, we observed that inside the lesion masks, m-NLSS over-performed NLSS(L) on five cases out of seven, whereas m-JLF showed better performance than JLF(L) on six of the seven cases. When it comes to the surroundings of the lesions (around lesions), the proposed correspondence models always provided better segmentation results than their corresponding originals.

In terms of the effect of the lesions on the overall performance of the brain,

Table 4.2: Pearson's correlation between the total lesion load and the DSC differences seen between pairs of methods (ref − other). Values calculated independently for the sub-cortical structures (subcort.), cortical gray matter (cortical) and white matter (WM).

		Lesions (L)			Healthy (H)			Filled (F)			Proposed		
		subcort.	*cortical*	*WM*	*subcort.*	*cortical*	*WM*	*subcort.*	*cortical*	*WM*	*subcort.*	*cortical*	*WM*
N	*p-val*	<0.0001	<0.0001	<0.0001	0.0054	0.0001	0.3220	0.1070	<0.0001	0.2900	ref	ref	ref
L	*R*	0.7772	0.6041	0.7861	-0.4125	0.5622	-0.1528	-0.2463	0.7717	0.1631	ref	ref	ref
S	*p-val*	<0.0001	0.0150	<0.0001	0.8540	0.0013	0.0410	ref	ref	ref	-	-	-
S	*R*	0.8199	0.3644	0.7999	0.0286	-0.4704	-0.3094	ref	ref	ref	-	-	-
J	*p-val*	0.0024	<0.0001	<0.0001	0.0531	0.0003	<0.0001	0.5967	0.0099	0.2573	ref	ref	ref
L	*R*	0.4580	0.6746	0.8405	-0.2936	-0.5234	-0.6243	-0.0820	0.3847	0.1745	ref	ref	ref
F	*p-val*	0.0004	0.0005	<0.0001	0.7262	0.0008	0.0009	ref	ref	ref	-	-	-
	R	0.5099	0.5058	0.8354	-0.0543	-0.4854	-0.4834	ref	ref	ref	-	-	-

we observed that, with the proposed correspondence models, m-NLSS improved its original, in mean, over the 0.09% (74.42 ± 1.95 *vs.* 74.51 ± 1.94), whereas m-JLF improved a 0.1% (77.72 ± 2.15 *vs.* 77.82 ± 2.12). On the other hand, filling the lesions before segmentation improved the original results on 0.03% for NLSS(F) (74.45 ± 1.90) and 0.21% for JLF(F) (77.93 ± 2.11).

Figure 4.7 shows some qualitative segmentation results obtained with the original and the proposed correspondence models for both NLSS and JLF methods. From this figure we observe that the original methods (NLSS(L) and JLF(L)) tend to segment the WM lesions as part of the lateral ventricles, whereas the proposed non-local models (m-NLSS and m-JLF) as well as the original methods when segmenting the filled images (NLSS(F) and JLF(F)) tend to adjust better the edge between the two structures.

4.5 Discussion

In this chapter, we have presented an approach to solve for voxel/patch correspondences on intensity-based multi-atlas label fusion segmentation when MRI-visible lesions are present. We have presented the theory to apply this approach to two well-known label fusion strategies: Non-local Spatial STAPLE (NLSS) and Joint Label Fusion (JLF). Our proposal performs as well as the original strategy when segmenting healthy subjects, whereas the experiments performed showed that it minimizes the effect of the lesions when segmenting lesioned brains, obtaining significantly better results than the original method.

Furthermore, when comparing our approach to the common lesion filling [31] technique, the results obtained for the MS simulated database showed that, masking out the lesions with our approach leads to significantly better segmentation results than filling them before segmentation. On the other hand, when the analysis was performed on the MRBrainS18 dataset, the results showed that filling the lesions

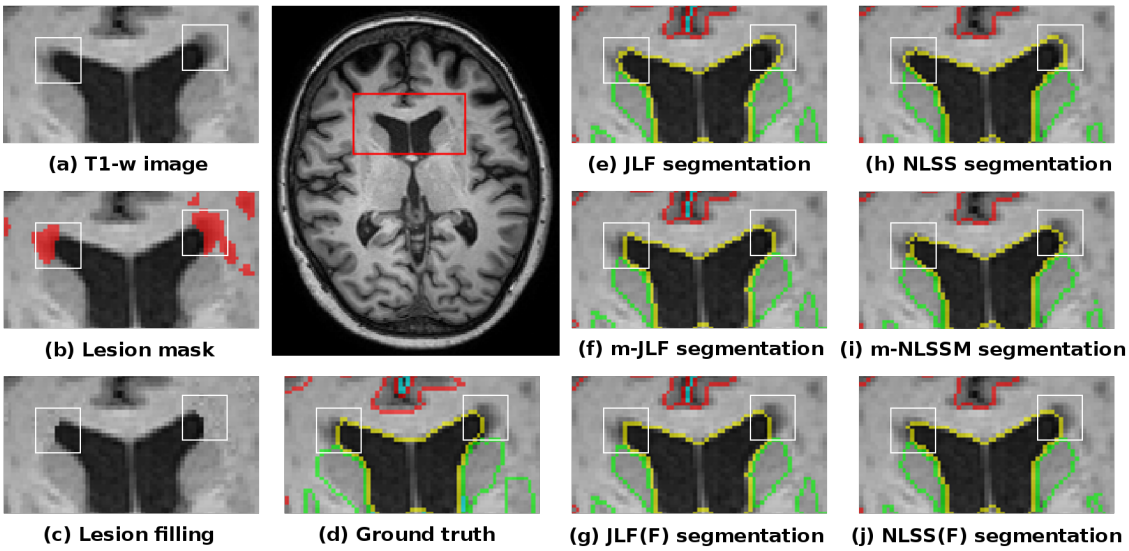


Figure 4.7: Qualitative segmentation results of patient "5" from the MRBrainS18 database obtained with the analyzed methods. The image shows (a) the original T1-w image, (b) the superimposed lesion mask, (c) the lesion-filled [31] T1-w image, (d) the segmentation ground truth, and the segmentation results for the original T1-w image segmented with (e) JLF, (h) NLSS, (f) our proposal for JLF (m-JLF), and (i) our proposal for NLSS (m-NLSS); and the segmentation results for the lesion-filled T1-w image segmented with (g) JLF, and (j) NLSS.

outperformed the segmentation result of our proposal for the JLF method, whereas the proposal for NLSS achieved better results than lesion filling. However, the results on this second database have to be interpreted carefully, since the amount of analyzed data (only seven images) is small.

WM lesions are usually segmented on FLAIR, T2-w or PD sequences, where they appear larger than in T1-w and, sometimes, lesions that are visible in those sequences are not even perceptible in T1-w images. When using filling techniques, one has to be careful, since we might risk corrupting image intensities that were correct, due to inaccurate segmentations. While it has been proved to improve the results for segmentation techniques in which intensity distributions of different tissue classes are modeled, studies have not been made to see how it affects patch based segmentation strategies, in which patch intensities are independent to the global intensity distributions.

Differences in the trend of the results seen in both databases could be explained by the nature of the lesions. In the simulated dataset, lesions were located with no restriction on the affected structures, whereas in the seven images of the MRBrainS18 database, lesions did not affect any other structure besides the WM. Because lesion

filling techniques tend to fill all lesions with WM-like intensities, these inaccurate intensities may be affecting the segmentation result of non-WM lesions on the simulated database. We believe that, for this reason, lesion filling under-performed our proposed method on the simulated database, whereas it worked better on the MRBrains18.

The small improvements seen on the whole brain segmentation are due to the lesion volumes being very small compared to the rest of the brain. Thus, a big improvement on small lesion areas will never have a big impact on the whole brain segmentation result. Furthermore, low values seen in the MRBrainS18 database for the whole brain mean DSC in all the analyzed strategies compared to the state of the art could be caused by the low resolution of the analyzed images and different labeling protocol of the atlases used with respect to the ground truth. However, the purpose of using this database was to compare the modified methods to their originals, which are equally affected by the low resolution and labeling protocols.

The analysis performed on the simulated data, showed that the lesion load has an important effect on the performance differences seen between the proposed methods and the rest of the strategies analyzed. In short, larger improvement was seen in the strategies we proposed compared to the corresponding original methods when segmenting images with larger lesion loads. This makes us believe that either our proposal works better with larger lesion loads or else that the original strategies are strongly affected by large lesion volumes. However, given that strong correlations were also found for healthy-lesions differences when applying the original segmentation strategies, we may conclude that large lesion loads have a bigger effect on the whole brain segmentation results of the original methods. This effect has been shown to be mitigated by our proposal for NLSS, achieving similar results to the original strategy when segmenting the corresponding healthy subjects. On the other hand, although the results obtained with our proposal for JLF are better than those obtained with the original method, they still do not reach the performance obtained for the healthy images.

Regarding the experimental data, it has to be noted that the OASIS database, which is the basis of the MICCAI 2012 challenge dataset, contains images of young, middle aged, non-demented and demented older adults, some of them diagnosed with probable Alzheimer's disease. Although we have referred to these images as healthy subjects, a reduced number of them may contain lesions, which are related to either the age or the disease. However, no ground truth of the lesions neither an image modality to properly segment them is provided, which makes hard to test our algorithms with the original lesion masks. One may think that these lesions could affect the performance of the methods and bias the presented results, however, as we will see in the following chapter, false negative lesions affect both the original and the proposed strategies in the same way.

In all the experiments performed, we used lesion masks that were manually annotated. However, expert-annotated masks are not always available in practice, which generates a need for automated methods capable of producing them. In this regard, automatic lesion segmentation has become a well-studied field, in the medical imaging community [16, 17, 19]. As a proof, several lesion segmentation challenges [245, 244, 246, 260] have been conducted in recent years, with participating strategies [19] able to achieve segmentation results that are close to human expert inter-rater variability. For this reasons, and, with the aim to propose completely automated segmentation strategies, in the next chapter, the effects of replacing the manually annotated lesion mask by automatically segmented ones, as input of the proposed label fusion algorithms, will be extensively addressed.

In conclusion, the results of this chapter show that the proposed correspondence models improve the segmentation results when MRI visible lesions are present, whereas they behave like the original method when they are not. When compared to lesion filling, simulated data show that our proposals perform better overall, while on the seven-case real dataset, our correspondence model only outperforms lesion filling on one of the two methods analyzed (NLSS). Besides this, by using the proposed method we eliminate the pre-processing steps required by lesion filling and make the results more robust, since it is indifferent to the quality of the filling method.

Chapter 5

A fully automated pipeline for brain structure segmentation in multiple sclerosis

5.1 Introduction

In Chapter 4, we presented two label fusion approaches to segment brain structures on images of MS patients. In all of the experiments performed in that chapter, the lesion masks used were annotated manually, however, in practice, manually annotated masks are rarely available since obtaining them is a highly time-consuming task. Furthermore, the use of manually annotated masks requires expert interaction before being able to apply the proposed automatic segmentation algorithms, which is impractical if our objective is to automate the brain parcellation process.

Fortunately, an increasing number of automatic MS lesion segmentation algorithms have been proposed in recent years with very promising results [14, 15, 16, 17, 18, 19], which allows us to automate the lesion mask acquisition, that we will later use in our final objective of segmenting the brain structures on MS patient images.

In this chapter, we present a fully automated pipeline for brain structure segmentation of MS patients. More specifically, in a first step, the co-registered FLAIR and T1-w sequences of the patient are given as input to an automatic lesion segmentation algorithm, which generates the lesion mask. Here, we have selected the SLS algorithm [16], since it is an unsupervised strategy, and hence it does not re-

quire any training, whereas it provides good segmentation results. However, any automatic method could be used instead. After the lesion segmentation, the T1-w image is affinely registered to the MNI305 template and the resulting transformation is applied to the obtained lesion mask. This is done in such a way that both the T1-w image and the lesion mask are moved to a standardized space (MNI) where the brain structure segmentation takes place. Once in MNI space, the patient T1-w image is N4-bias-field corrected, and intensity normalized to a previously built "atlas model" space. Normalizing the target intensities with the atlases is very important since our label fusion strategies depend on correspondence search models based on target-atlas patch-intensity similarity.

In our pipeline, only the 15 atlases more similar to the target, from a cohort of 45 [229], are used for segmentation. These 15 atlases are selected by performing a PCA based atlas-selection strategy [261]. To obtain the PCA manifold from all the 45 atlases, which is done offline, the 3D intensities within the same MNI brain mask of each atlas are converted to a 1D vector. Then, a naïve PCA projection is performed on 1D vectors from all atlases to learn the PCA manifold.

After intensity normalization, the patient T1-w image is projected to the same PCA manifold and the 15 atlases with smallest Euclidean distance to the patient scan are selected to perform the segmentation. Then, the selected atlases are registered to the normalized patient image, using an initial affine registration [256] followed by a non-rigid procedure. The non-rigid registration strategy used [255], is based on the symmetric image normalization method (SyN) proposed by Avants et al., which is implemented as part of the advanced normalization tools (ANTs) package[1]. In all the registrations performed, the automatically segmented lesion mask, that is already in MNI space, is used to mask out the lesion areas in such a way that we avoid their intensities interfering in the similarity metric calculation. The deformation fields obtained from the registration are then applied to the corresponding atlas labels, which are propagated to the patient space, becoming potential brain structure segmentations of the target.

The propagated atlas labels are fused with one of the proposed strategies seen in Chapter 4, i.e. m-NLSS and m-JLF, to obtain the final brain structure segmentation. As previously seen, both m-NLSS and m-JLF require information from the patient T1-w image, the lesion mask and the atlas structural images to obtain the atlas-target correspondences. Thus, both the atlas labels and intensity images, combined with the patient structural image and its corresponding lesion mask are fed to the fusion algorithm, that computes the final segmentation. Note that the obtained segmentation is in MNI space, therefore, the inverse of the transformation resulting from affine registration of the original patient image to the MNI305 template is applied to the fusion result to translate it back to the original patient space. A

[1]http://stnava.github.io/ANTs/

graphical representation of the whole pipeline in presented in Figure 5.1. Note also from this figure that the automatically obtained lesion mask can also be considered as an output of this pipeline in case it were necessary for medical purposes, such as lesion quantification or follow-up.

While we have evaluated the performance of our methods with manual lesion masks in Chapter 4, we do not know the effect of using automatically segmented masks on the brain parcellation result. For this reason, in the following section we perform an analysis of the influence of the lesion masks used on the structure segmentation result of the proposed strategies. To do this, we compare the parcellations obtained with the presented pipeline, for both m-NLSS and m-JLF, to the ones obtained with the same pipeline in which we substitute the automatically obtained lesion mask with the manually segmented lesion mask, i.e. the ground truth. Furthermore, to be consistent with the previous chapter, we also evaluate the effect of such lesion masks on the segmentation results of the original methods, i.e. NLSS and JLF, when we use the automatic and manually annotated masks to in-paint the lesion intensities [31] before brain structure segmentation.

5.2 Experiments and evaluation

In this section we evaluate the effect of automatic lesion masks on the segmentation result provided by the proposed label fusion strategies. For this experiment, we perform brain parcellation on a set of MS patient images and compare the segmentation results obtained for all the fusion strategies evaluated in Chapter 4, when feeding the algorithms with both manual and automatic lesion masks.

5.2.1 Data

The images used for evaluation are from the MICCAI MS segmentation (MSSEG 2016) Challenge database [244]. This dataset consists of 15 MS patients with lesion loads ranging from 0.91 to 68.94 ml. The images of this database are from three different MRI scanners and different manufacturers including those using 3T and 1.5T magnets. For each patient four different MR sequences (3D FLAIR, 3D T1-w, 3D T1-w GADO, 2D DP/T2-w) are available, as well manual lesion delineations from seven different trained experts. From these segmentations, a consensus ground truth segmentation was built for evaluation with the LOP STAPLE algorithm [262]. Demographics are shown in Table 5.1.

The atlases utilized in our experiments include the 45 images from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling database [229], following the same pre-processing steps as seen in Section 4.3.2.

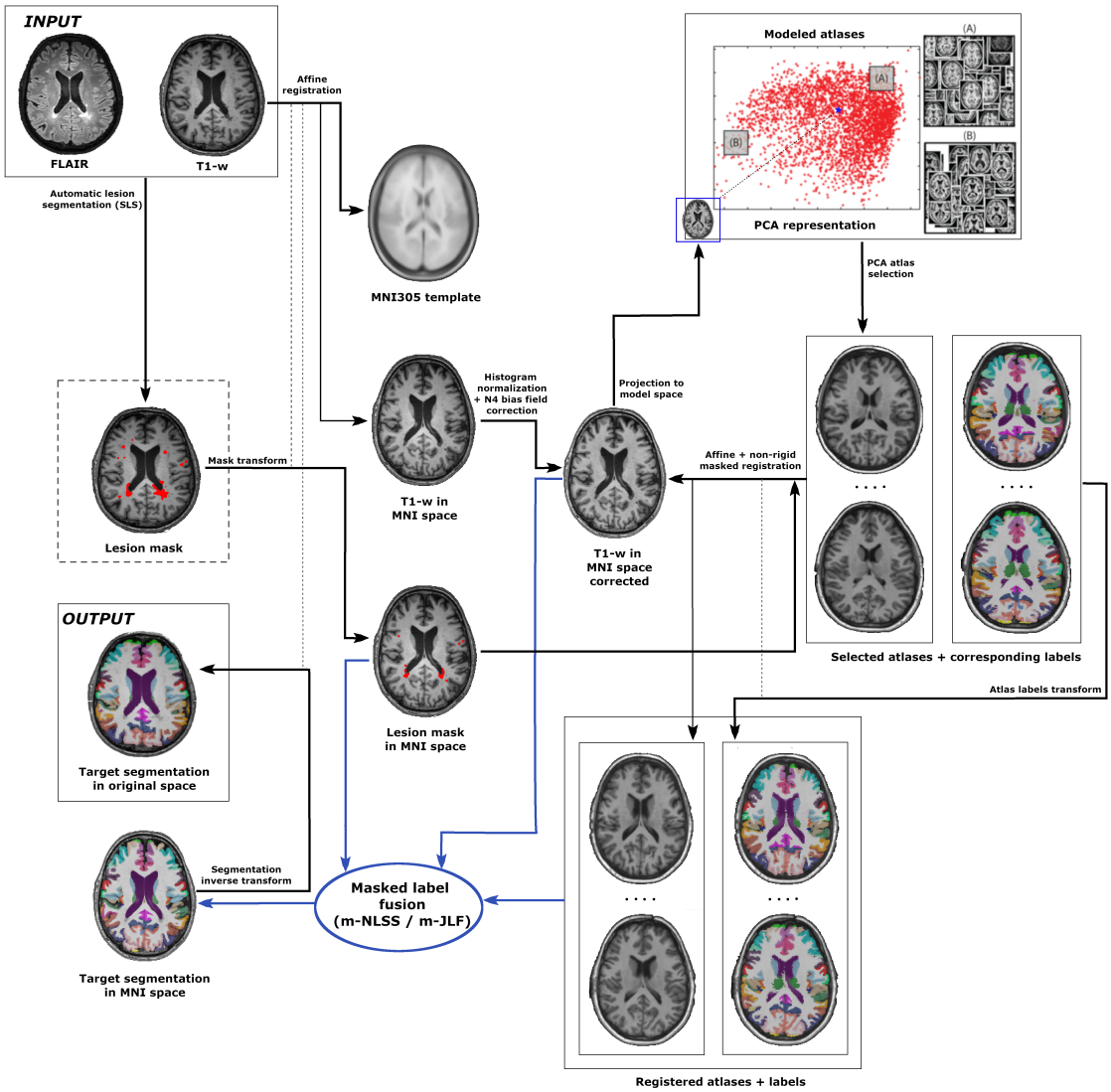


Figure 5.1: Fully automated pipeline for structure segmentation of MS patients. As input of this pipeline, the FLAIR and the T1-w sequences of the patient are required. The automatic segmentation of the lesions is performed by means of the SLS algorithm [16]. Then, the T1-w image and the lesion mask are moved to MNI space, where the structure segmentation is performed. Once in MNI space, the patient T1-w image is bias field corrected and intensity normalized to the atlases model space. After that, the target is projected to the model space and the 15 most similar atlases are selected to participate in the segmentation. Those atlases are affine and non-rigid registered to the target masking out the lesion voxels labeled by SLS. The deformation fields obtained from these registrations are applied to the corresponding atlas labels. Finally, the atlas labels are fused by means of one of the proposed strategies (m-NLSS or m-JLF), for which information from the target, the lesion mask and the atlas intensities is required. The obtained segmentation result is then back-propagated to its original space.

Table 5.1: MICCAI MSSEG 2016 Challenge demographics for the training dataset.

	Patient age		Patient gender			Lesion load (ml)	
Scanner	Mean	Std	Male	Female	Male:female	Mean	Std
Siemens Verio 3T	35.00	10.10	1	4	0.25:1	17.40	24.86
Siemens Aera 1.5T	43.80	8.32	2	3	0.67:1	6.68	17.90
Philips Ingenia 3T	46.00	9.14	4	1	4.00:1	9.41	14.92
Overall	41.60	9.85	7	8	0.88:1	10.30	20.15

5.2.2 Methods

The method of Roura et al. [16] (SLS) was used to automatically segment MS lesions on the evaluated database and generate the automatic lesion masks used in our experiments. We selected this approach due to it is an unsupervised strategy, and produced encouraging results in the MSSEG Challenge, obtaining the second best position in both lesion segmentation and detection tasks [244]. This method consists of outlier segmentation based on brain tissue labeling and post-processing rules. The authors consider the lesions as intensity outliers, that appear as hyperintense regions in the FLAIR sequence, presenting an approach based on two main steps. First, they perform tissue segmentation on the T1-w sequence, which they use to compute the intensity distribution of the GM in the FLAIR image, since it is the brightest healthy tissue in this modality. Then, since the lesions are even brighter than the GM, their intensities are considered to be outliers of this distribution. Thereafter, some post-processing steps are applied to remove false positive lesions that remain after thresholding the FLAIR volume.

To evaluate how automatic lesion masks affect the output of the methods presented in the previous chapter, we compare the segmentation result of the methods when using the consensus masks described in Section 5.2.1, and the automatic SLS masks. To this end, the 15 images are segmented with both methods (m-NLSS and m-JLF) twice (consensus vs. SLS), applying the same lesion masks also for the atlas masked registration. Furthermore, to be consistent with the previous chapter, we also compare how the original strategies (NLSS and JLF) behave when we perform lesion filling [31] with both the consensus and the SLS masks. The filling strategy used is the same as in Chapter 4.

5.2.3 Evaluation

Since brain structure ground truth is not available for the MSSEG 2016 database, we quantitatively evaluate the effect of the automatic lesion masks on the segmentation using the structure volume percentage increase with respect to the manual lesion mask execution, as follows:

$$Vol_{\%increase} = \frac{100 \cdot (Vol_{SLS} - Vol_{manual})}{Vol_{manual}} \tag{5.1}$$

where Vol_{SLS} and Vol_{manual} are the structure volumes when the automatic and consensus lesion masks are fed to the algorithm, respectively.

In order to provide more global results, the 133 structures obtained from the segmentation are combined into the following regions: cortical GM, cerebral WM, sub-cortical GM, CSF, cerebellum GM, cerebellum WM and brainstem.

Statistical analysis is performed using the Matlab software package. We test for significant differences in the structure volumes obtained with the analyzed methods. These differences are computed using one-sample t-tests of the structure volume percentage increase. Moreover, the Pearson's linear correlation coefficient is used to compute the correlation between the lesion volume percentage increase and the structure volume percentage increase achieved when comparing manual vs. automatic lesion masks.

5.3 Results

5.3.1 Quantitative results

In the following, we present a comparative of the structure volume changes observed on the analyzed methods, when using the two different lesion masks. Figure 5.2 shows the structure volume percentage increase obtained for the four parcellation algorithms: (1) Non-local Spatial STAPLE segmentation of the lesion-filled target image (NLSS(F)), (2) masked Non-local Spatial STAPLE of the original target image (m-NLSS), (3) Joint Label Fusion of the lesion filled target image (JLF(F)), and (4) masked Joint Label Fusion of the original target image (m-JLF).

Notice that significance is assessed for each box independently, since the boxes are independent of each other. Each of them represents the volume percentage increase of the analyzed structure when the brain parcellation algorithm is fed with automatic lesion masks (SLS) with respect to the same algorithm fed with manual (consensus) lesion masks. Negative values indicate a volume decrease.

From the figure we observe that with the proposed methods (m-NLSS and m-JLF), the volume change of the analyzed structures was not significant, in a comparison between manual and automatic lesion masks. On the other hand, when comparing the result of segmenting the lesion-filled images with the original methods (NLSS(F) and JLF(F)), we observed that some brain structures presented a significant volume change. In the case of NLSS(F), the CSF overcame a $1.13\% \pm 1.93$

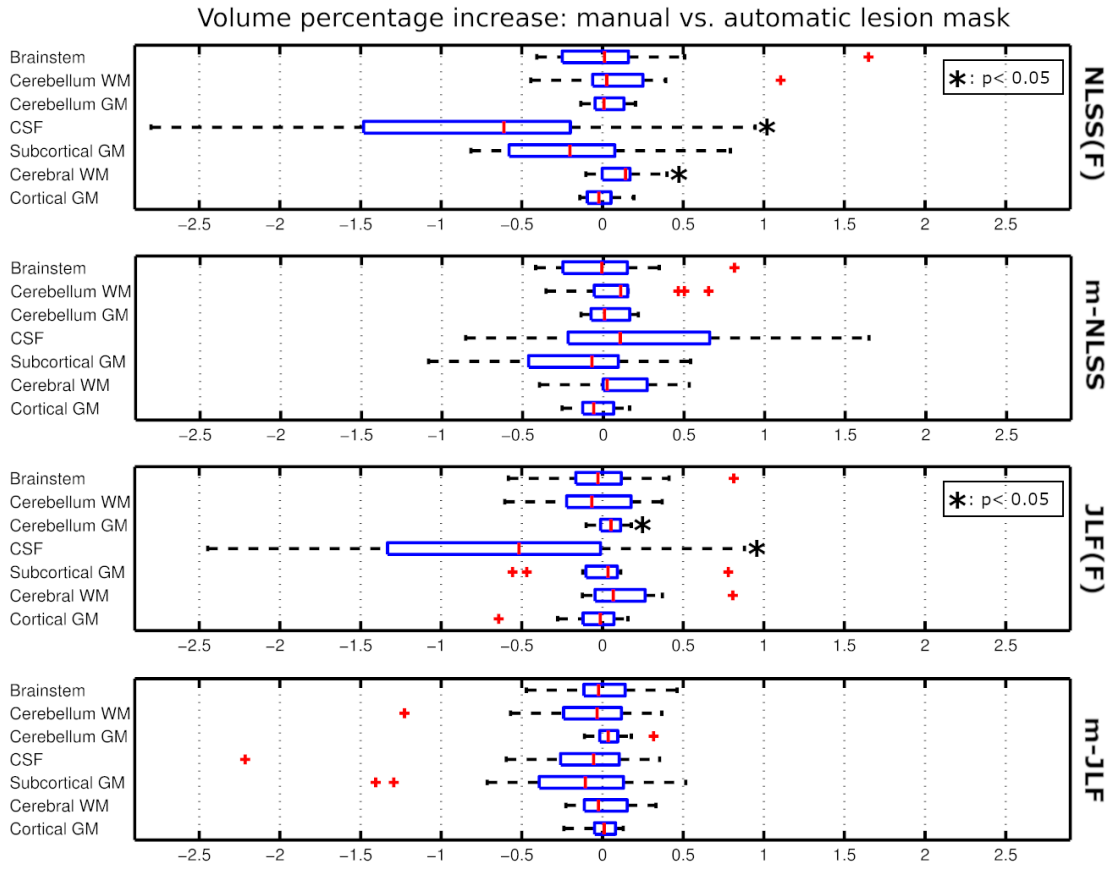


Figure 5.2: Structure volume percentage increase. Comparison of the structure volume change when utilizing manual vs. automatic lesion masks on the analyzed brain structure segmentation strategies: (1) NLSS on the lesion-filled target image (NLSS(F)), (2) masked NLSS on the original target image (m-NLSS), (3) JLF on the lesion-filled target image (JLF(F)) and (4) masked JLF on the original target image (m-JLF). Significance assessed for each method and structure independently, with boxes significance independent to each other.

($p < 0.05$) volume decrease, and the cerebral WM presented a $0.13\% \pm 0.14$ ($p < 0.05$) volume increase. Furthermore, in the case of JLF(F), the cerebellum WM presented a $0.05\% \pm 0.08$ ($p < 0.05$) volume increase, whereas the CSF showed a $0.74\% \pm 1.03$ ($p < 0.05$) volume decrease.

No significant correlations were found between the lesion volume percentage increase and the structure volume percentage increase, except for the cerebral WM, which showed a weak correlation (0.53, $p < 0.05$) with the lesion load when the method analyzed was NLSS(F).

5.3.2 Qualitative results

Some qualitative results of the segmentation outputs obtained in our experiments are shown in what follows. Figures 5.3 to 5.5 present an example of a false positive lesion on the left hippocampus (right side of the image), and an over-segmented lesion on the right hippocampus (left side of the image). Figure 5.3 shows the (a) FLAIR and (b) T1-w sequences of the target image and the superimposed lesion masks, in red, of (c) the experts' consensus delineation and (e) the automatic segmentation obtained with SLS. In addition to this, the resulting images of applying lesion filling to the target image with (d) the manual and (f) the SLS masks are depicted in the image. As as appreciated in the FLAIR sequence, the hippocampus area (yellow arrows) looks brighter than the normal-appearing GM and the automatic lesion segmentation method has mis-classified it as an outlier (lesion). In both of the lesions shown in that figure, the over-segmented voxels pertain to the hippocampus, which is a GM structure. Therefore, the resulting "extra" filled lesion voxels (Figure 5.3 (f)) obtained with the automatic mask show abnormal WM-like intensities on that area. On the other hand, when the manual mask is used to fill the lesions, both hippocampi seem to conserve their original intensities and shapes.

Regarding the effect of this mis-classification on the segmentation output of the analyzed methods, we observe from Figures 5.4 and 5.5 that independently of the lesion mask used, the proposed strategies (Figures 5.4 − 5.5 (c) and (f)) present similar structure classification results. However, in the case of the filled images we observe a similar trend with both segmentation methods (NLSS and JLF). In the case of the automatic lesion mask, the abnormal WM-like intensities are producing an under-segmentation of both hippocampi (Figure 5.4 (e) and Figure 5.5 (e)) when compared to the manual lesion mask (Figure 5.4 (b) and Figure 5.5 (b)).

Another example is illustrated in Figures 5.6 to 5.8. These figures show a case of false negative lesions, two of them surrounded by WM and one peri-ventricular lesion, i.e. attached to the lateral ventricle. From Figure 5.6 (d) we observe that some lesions have not been detected by the automatic segmentation method (yellow arrows), and therefore after applying the filling on the target image with the SLS mask (Figure 5.6 (f)), their abnormal intensities have not been replaced by normal-appearing WM intensities, contrary to Figure 5.6 (e).

When analyzing the behavior of the multi-atlas segmentation strategies (Figures 5.8 to 5.7) we observed that both families of methods (NLSS and JLF) achieved similar results. In this case, the three lesions should be classified as WM. However, if we pay attention to the peri-ventricular one (right-inferior arrow) we realize that when using the automatic lesion mask, in which that lesion has not been detected, the four analyzed strategies (Figures 5.7 (e), 5.7 (f), 5.8 (e) and 5.8 (f)) classify it as part of the lateral ventricles. On the other hand, when the consensus manual

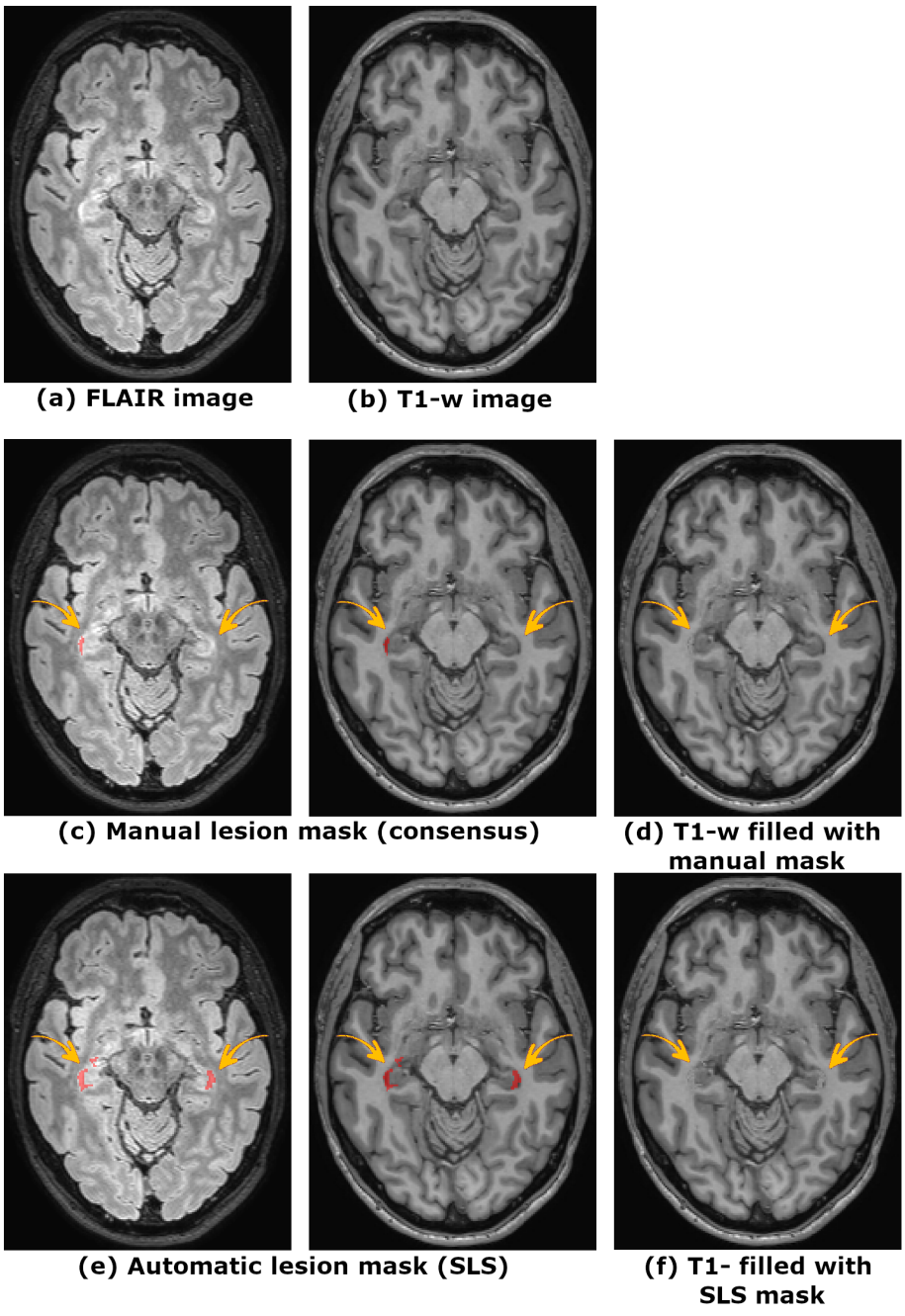


Figure 5.3: False positive and over-segmentation example. Axial slice of the original (a) FLAIR and (b) T1-w sequences, (c) the super-imposed consensus lesion mask, (d) the resulting T1-w image after applying lesion filling on the consensus lesion mask, (e) the automatic lesion mask, and (f) the resulting T1-w after filling the lesions found by SLS.

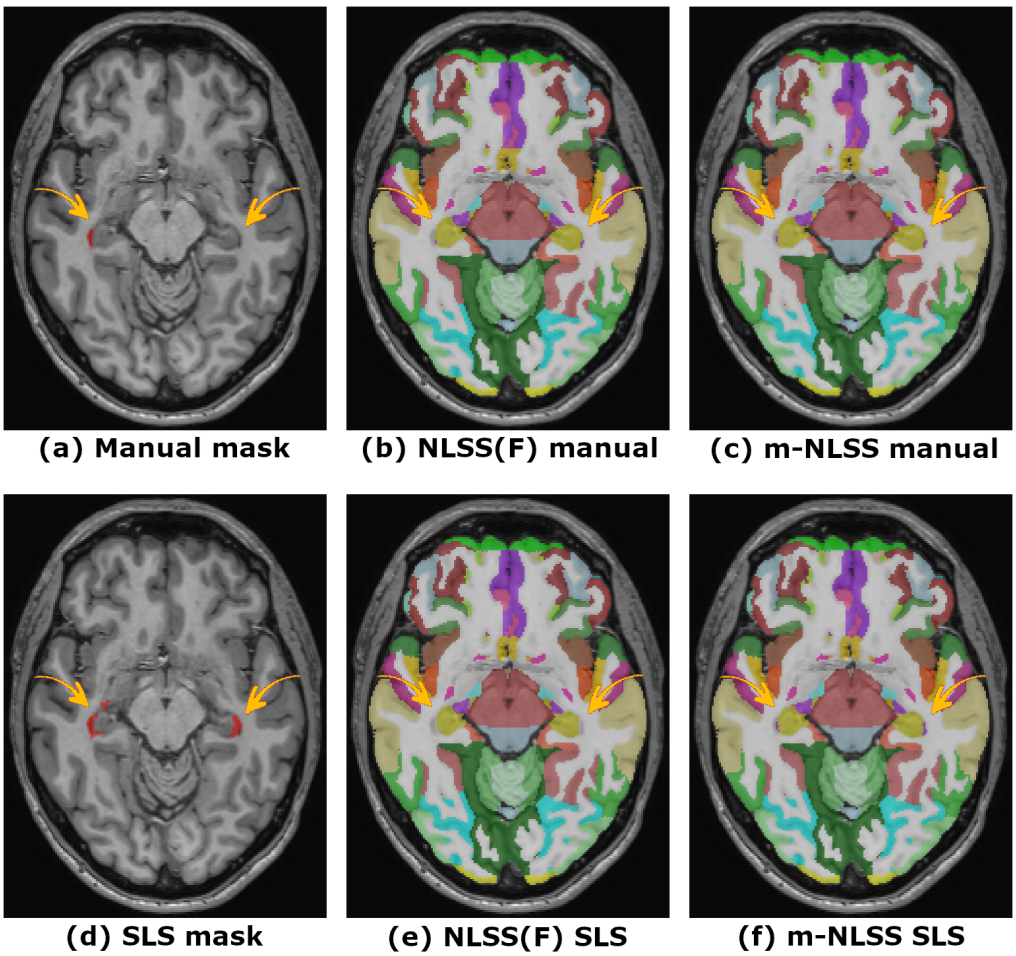


Figure 5.4: Non-local Spatial STAPLE family segmentation results. False positive and over-segmented lesions. T1-w target image with super-imposed (a) consensus and (d) automatic lesion masks. NLSS segmentation on the, target image filled (NLSS(F)) with (b) the manual and (e) automatic lesion masks. Segmentation result of m-NLSS for the target image with (c) the manual and (f) the automatic lesion mask as input. Note that (b) and (e) are the result of segmenting the lesion-filled images, however the original target is shown under the segmentation as a reference.

mask is used, both the original methods previous lesion filling (Figures 5.7 (b) and 5.8 (b)) and the proposed strategies (Figures 5.7 (c) and 5.8 (c)), properly classify the lesion as WM.

Regarding the other two lesions (two arrows on the top), they are totally surrounded by WM. In this case, when the manual lesion mask is used, either to fill the lesion intensities before performing the structure segmentation with the original methods (Figures 5.7 (b) and 5.8 (b)), or as input to the proposed strategies

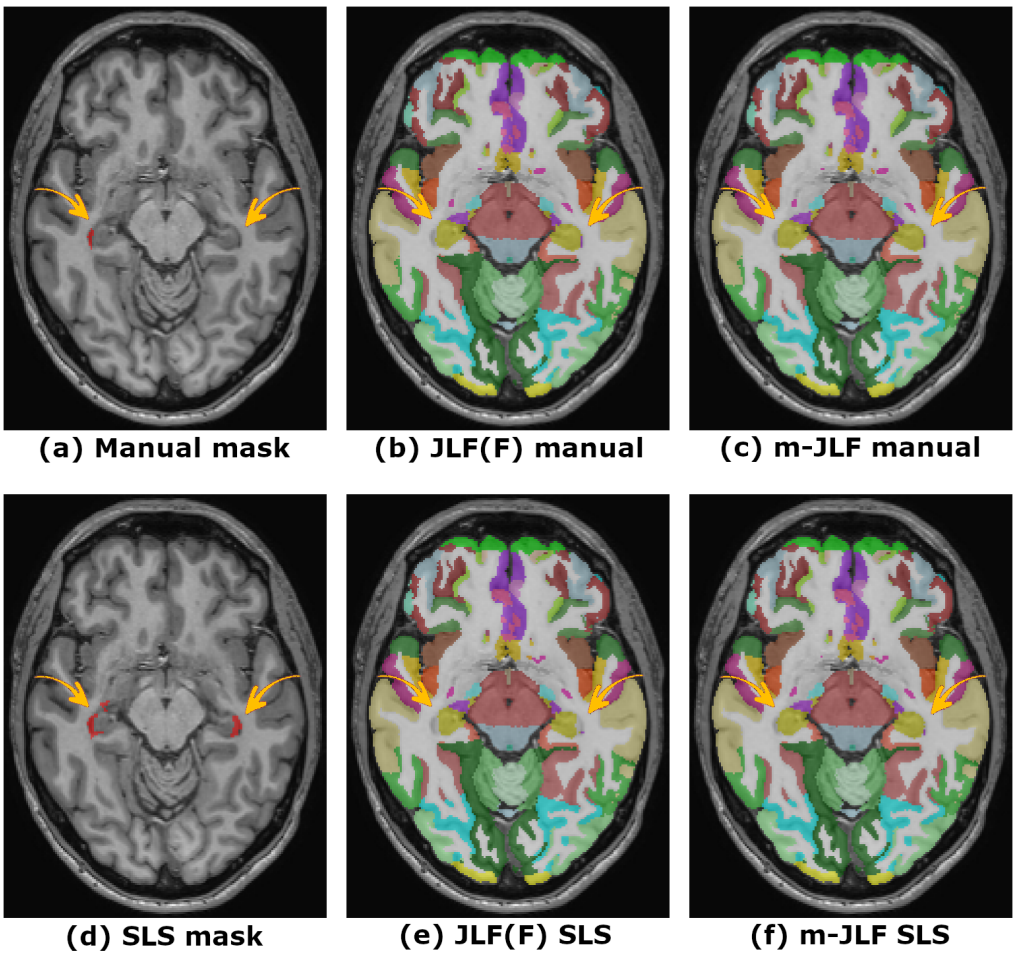


Figure 5.5: Joint Label Fusion family segmentation results. Example of false positive and over-segmented lesions. T1-w target image with super-imposed (a) consensus and (d) automatic lesion masks. JLF segmentation on the, target image filled (JLF(F))) with (b) the manual and (e) automatic lesion masks. Segmentation result of m-JLF for the target image with (c) the manual and (f) the automatic lesion mask as input. Note that (b) and (e) are the result of segmenting the lesion-filled images, however the original target is shown under the segmentation in order to allow an easier comparison with (c) and (f).

(Figures 5.7 (c) and 5.8 (c)), the lesions are correctly classified as WM, as expected. On the other hand, when the automatic lesion mask is used, where the lesions have not been detected, all the methods surprisingly segment the lesions as part of the WM. Even though this behavior may seem disconcerting, it is totally normal, since the original methods (NLSS and JLF) do not always mis-classify the lesion areas. In particular, they tend to succeed when lesions appear surrounded by WM. Given

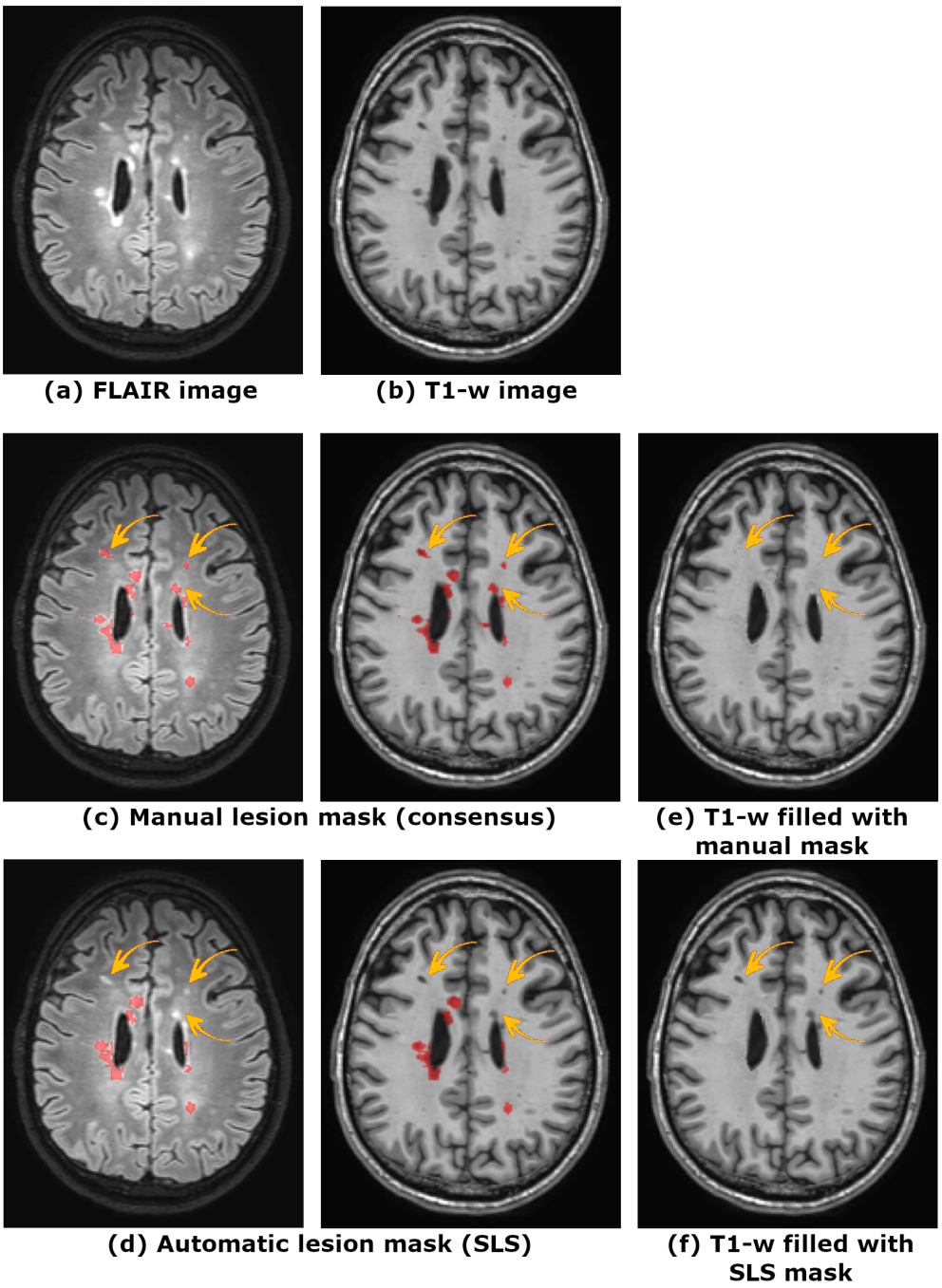


Figure 5.6: False negative example. Axial slice of the original (a) FLAIR and (b) T1-w sequences, (c) the super-imposed consensus lesion mask, (e) the resulting T1-w image after applying lesion filling on the consensus lesion mask, (d) the automatic lesion mask, and (f) the resulting T1-w after filling the lesions found by SLS.

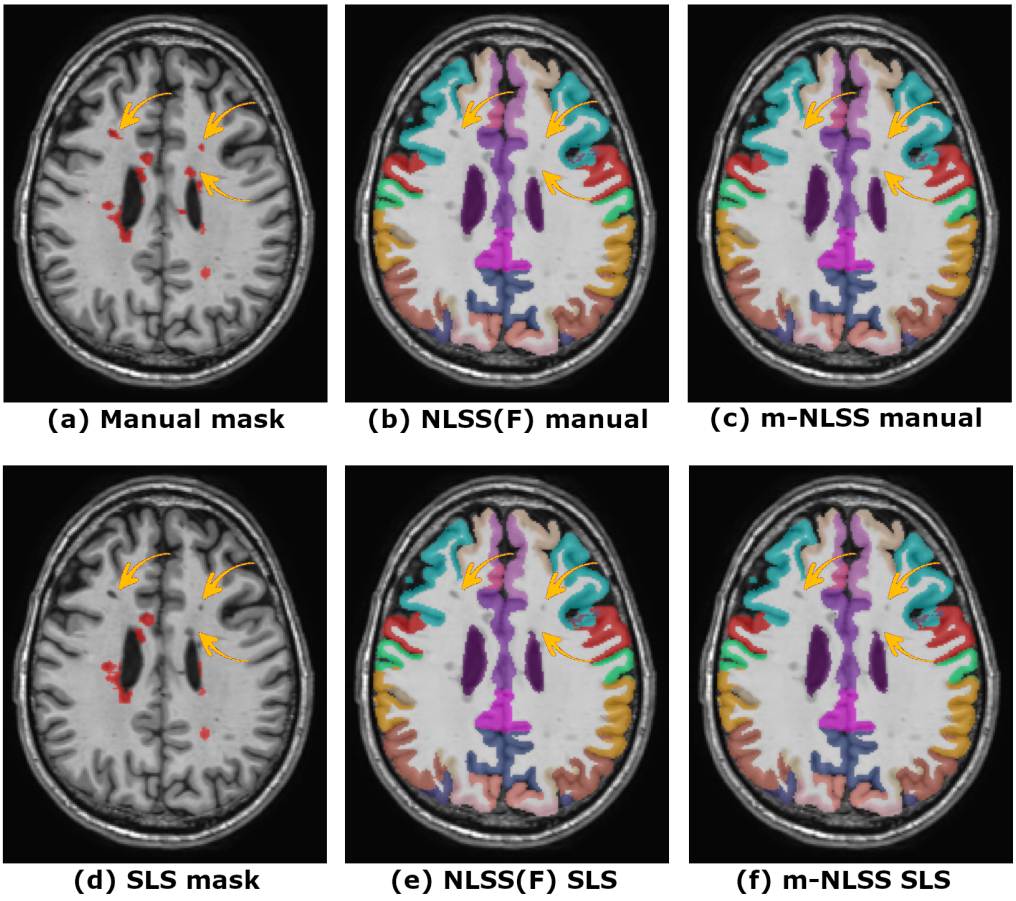


Figure 5.7: Non-local Spatial STAPLE family segmentation results. Example of false negative lesions. T1-w target image with super-imposed (a) consensus and (d) automatic lesion masks. NLSS segmentation on the, target image filled (NLSS(F))) with (b) the manual and (e) automatic lesion masks. Segmentation result of m-NLSS for the target image with (c) the manual and (f) the automatic lesion mask as input. Note that (b) and (e) are the result of segmenting the lesion-filled images, however the original target is shown under the segmentation in order to allow an easier comparison with (c) and (f).

that NLSS and JLF correctly classify these two lesions in the original (not filled) T1-w image, the analyzed methods, which behave exactly as their originals when no lesion mask applies, also succeed in their classification.

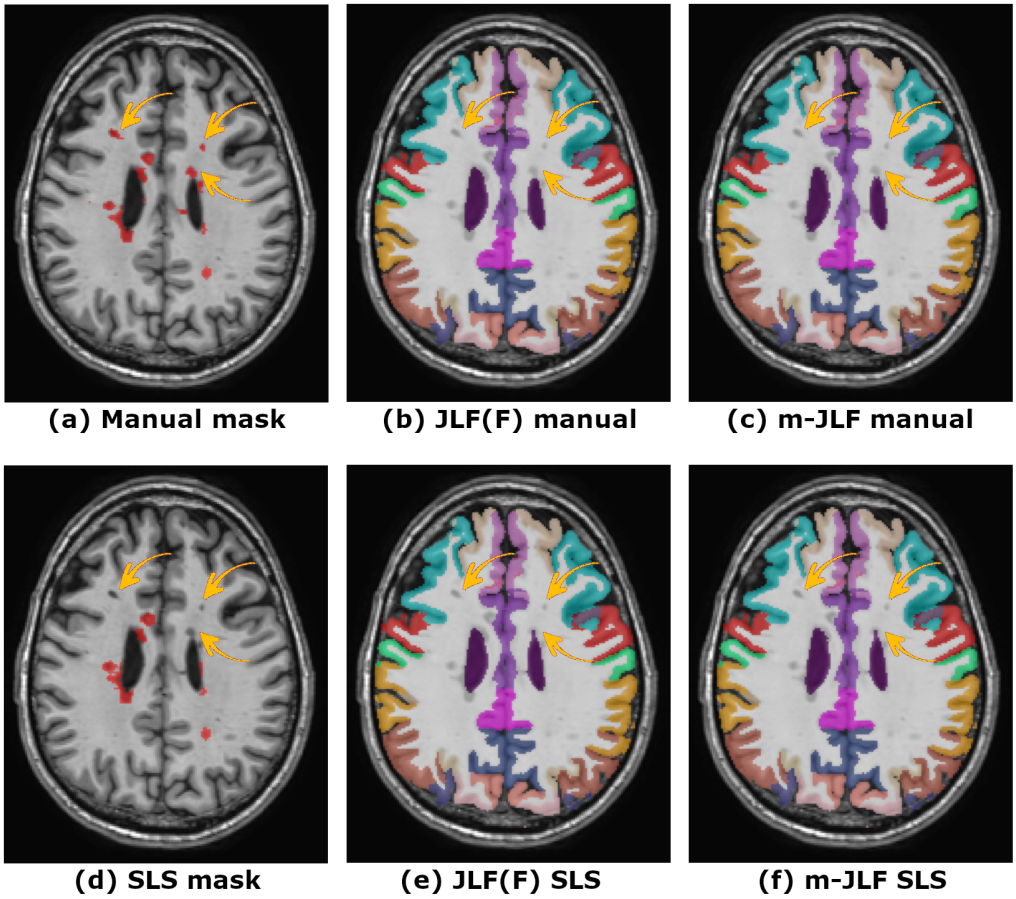


Figure 5.8: Joint Label Fusion family segmentation results. Example of false negative lesions. T1-w target image with super-imposed (a) consensus and (d) automatic lesion masks. JLF segmentation on the, target image filled (JLF(F))) with (b) the manual and (e) automatic lesion masks. Segmentation result of m-JLF for the target image with (c) the manual and (f) the automatic lesion mask as input. Note that (b) and (e) are the result of segmenting the lesion-filled images, however the original target is shown under the segmentation for reference with respect to (c) and (f).

5.4 Discussion

We have evaluated the effect of using automatic lesion masks on the brain structure segmentation results of both the proposed masked strategies (m-NLSS and m-JLF) and on their corresponding originals before applying lesion filling (NLSS(F) and JLF(F)). The results obtained showed that with the proposed strategies, no significant structure volume differences were found with respect to the use of manual lesion masks. On the other hand, when the original methods were used, significant

volume differences on the CSF, the cerebellum GM and the cerebral WM appeared, when comparing the segmentation result obtained for both the lesion-filled images (manual vs. automatic lesion masks).

Regarding the nature of the lesions, when false positives are found by the automatic algorithm, either in the form of a new lesion or as an over-segmentation of an existing lesion, the use of lesion filling might be risky. Note that SLS segments lesions as hyper-intense outliers in the FLAIR sequence, and therefore, since the GM is the brightest healthy tissue in FLAIR, it is probable that the false positives belong to this tissue. If this happens, filling the healthy GM tissue with normal-appearing WM intensities might be self-defeating, causing the automatic brain structure segmentation methods to mis-classify those areas, as seen in Section 5.3.2. On the other hand, with our proposal, the structure classification in the "fake lesion areas" relies only on the label information of the healthy atlases. Since we do not take the intensities into account, it becomes a "traditional" label fusion problem, i.e. fusion without structural image information, which has been proved to achieve very competitive results on healthy subjects [139].

Alternatively, the effect of false-negative lesions will depend mostly on the behavior of the original algorithms (NLSS and JLF) against that particular lesion, but also on the (masked) registration result. In particular, when the lesion is totally surrounded by the structure to which it belongs, for example WM, the original approaches tend to correctly classify it as part of this structure, as we saw in Section 5.3.2. Note that the effect will be the same when a lesion of this kind is detected as a false positive. However, in the case of juxta-cortical, i.e. attached to the cerebral cortex, or peri-ventricular lesions, i.e. abutting the lateral ventricles, the result is more uncertain, since the lesion intensity and morphology play an important role in the segmentation result. In any case, false negative lesions will have the same effect on both the original previous lesion filling and the proposed strategies, given that the proposed methods behave as their originals in the absence of lesions.

In conclusion, the proposed segmentation strategies, m-NLSS and m-JLF, have shown to be robust against variations in the lesion mask used. The robustness of these strategies when used in combination with automatically segmented lesion masks makes the presented fully automated pipeline totally suitable for medical practice, obtaining similar results to the ones achieved with the use of expert segmented lesion masks. However, the same conclusion cannot be applied to the original strategies when segmenting the lesion-filled images. Even though the volume change observed for most of the structures is low, the results show that these changes are significant when comparing the segmentation of the target image filled with the manual and automatic lesion masks.

Chapter 6

Integration of known label masks into the label fusion estimation

6.1 Introduction

The proposal to create integrated segmentation algorithms, that not only segment tissue, but also lesions, is strongly supported by expert radiologists in MS [263]. Furthermore, the possibility of dynamically integrating new labels to the target segmentation which are not available in the atlases, such as a lesion class or a new sub-structure, would make the methods more versatile.

In some situations, it may happen that the "true" segmentation for a concrete structure or region of the target image is known beforehand. Either because it comes from a manual segmentation performed by an expert or from an automatic segmentation algorithm in which we are very confident. Figure 6.1 depicts an example of this situation, in which the "true" labels for both thalami are known beforehand. If we perform a traditional multi-atlas segmentation (see Figure 6.1 (e)), in which the label information is obtained from the registered atlases, the "true" thalamus label information that we know beforehand would be wasted. However, if we could include this known thalami mask information into the label fusion estimation process, the obtained multi-atlas segmentation (see Figure 6.1 (f)) would be better adjusted for these structures (green label), which are already modeled in the atlases, and the estimation for the surrounding structures, i.e. WM (white label) and third ventricle (brown label) would also benefit from this information. On the other hand, in this example, the internal segmentation of this structure, that includes seven thalamic subparts in each hemisphere, is also available. In this case, the segmentation has

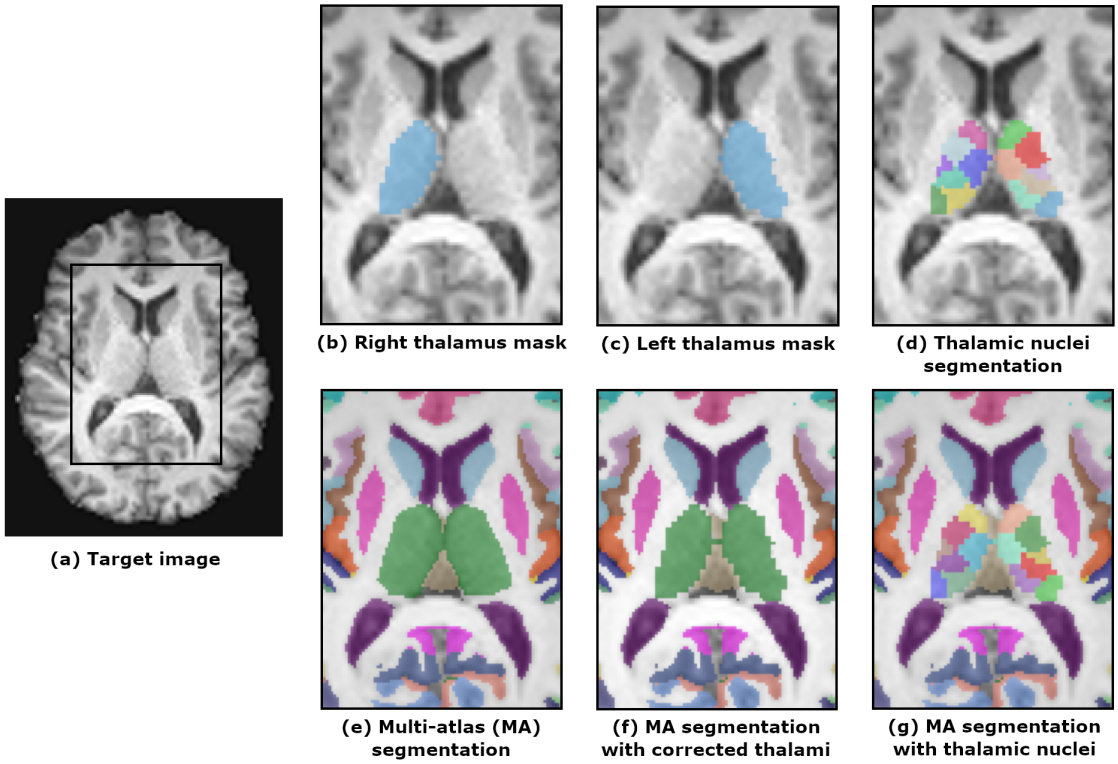


Figure 6.1: Example of the use of known label masks. The (a) target image for with both (b) the right and (c) the left thalamus segmentation is known beforehand. Furthermore, (d) the automatic segmentation [264] for fourteen anatomic thalamic subparts (seven for each hemisphere) is also available. The segmentation result obtained with (e) multi-atlas segmentation shows an over-segmentation of both thalami, which can be corrected if (f) the known thalamus masks are given to the fusion algorithm. Moreover, since the thalamic nuclei segmentation is also known, (g) this information can be included in the segmentation estimation process in such a way that the new labels, not available in the atlases, can be incorporated into the final target segmentation.

been estimated by means of an automatic segmentation algorithm [264], including labels that are not modeled in the atlases. Therefore, if we could combine the atlas label information with the one obtained from another source, which could be an automatic segmentation algorithm that parcellates a structure according to its sub-structures, new labels that are not included in the atlases could be added to the segmentation estimation. This approach, besides obtaining a more detailed segmentation of the target image, would also produce a better adjusted segmentation of the surrounding structures, as shown in Figure 6.1 (g) in comparison to Figure 6.1 (e).

Fortunately, the nature of multi-atlas strategies makes them very flexible to label *edits* integration, while they still allow us to deal with the abnormal lesion intensities. For this reason, in this chapter, we extend the theory of the strategies proposed in Chapter 4 to include a second mask, which we define as a "known mask", into the estimation process. This mask forces the voxel label assignation in case it is known beforehand and enables seamless integration of manual edits. Together, this innovation in combination with the ones proposed in Chapter 4 enable the inclusion of masks of abnormal anatomy and manually provided *edits* within modern statistical fusion approaches.

6.2 Incorporation of known label masks

In the following, we first extend the theory of m-NLSS to include a binary known label mask and then, we do the appropriate extention for the proposed m-JLF method to include a probabilistic known mask, which will allow us to embrace two different points of view.

6.2.1 Masked Non-local Spatial STAPLE (mk-NLSS)

In this section, we extend the theory of m-NLSS to include the so called known mask K. Let $K \in \{0,1\}^{N\times 1}$ be a binary mask specifying if for a given voxel i of the target image, the true label is known, hence $K_i = p\left(T_i = T_k \in \mathcal{L}'^{N\times 1}\right)$ and $T \in \mathcal{L}'^{N\times 1}$, the latent representation of the true target segmentation, where $\mathcal{L}' = \{0,\ldots,L'-1\}$ is the set of possible labels which can be assigned to a concrete voxel. Note that this mask is optional and can be neglected if all their voxels are set to 0.

Let's redefine the density function of the performance level parameters, θ, to include the known mask as follows:

$$\theta_{jis's} \equiv p\left(D_{i^*j} = s', A_j | T_i = s, I_i, M_i, K_i, \theta_{jis's}\right) \tag{6.1}$$

where i^* is the voxel on atlas j that corresponds to the target voxel i assigned by our correspondence model (see Equation 4.1).

Using an assumption of conditional independence between the labels, lesion mask and intensity, we can approximate the density function as:

$$\begin{aligned} \mathrm{p}(D_{i^*j} = s', A_j | T_i = s, I_i, M_i, K_i, \theta_{jis's}) &\approx E\left[p\left(D_j, A_j | T_i = s, I_i, M_i, K_i, \theta_{jis}\right)\right] \\ &= E\left[p\left(D_j | T_i = s, M_i, K_i, \theta_{jis}\right) \cdot p\left(A_j | I_i, M_i\right)\right] \\ = \sum_{i' \in \mathcal{N}(i)} p\left(D_{i^*j} = s' | T_i = s, M_i, K_i, \theta_{jis's}\right) \cdot p\left(A_{i'j} | I_i, M_i\right) &= \sum_{i' \in \mathcal{N}(i)} \theta_{jis's} \cdot \alpha_{ji'i} \end{aligned} \tag{6.2}$$

E-step

As previously seen in Section 4.2.2, $W_{si}^{(t)}$ represents the probability that the true label associated with voxel i is s at iteration t of the algorithm given the provided information and the performance level parameters.

$$W_{si}^{(t)} \equiv p\left(T_i = s | D, A, I, M, K, \theta^{(t)}\right) \tag{6.3}$$

If we assume independence among raters, using Bayes' rule we can rewrite this equation as follows:

$$\mathrm{W}_{si}^{(t)} \equiv \frac{(1-K_i)\cdot\left(p(T_i=s)\cdot\prod_j p\left(D_{i^*j}=s', A_j | T_i=s, I_i, M_i, K_i, \theta_{jis's}^{(t)}\right)\right)+K_i\cdot\delta(s'=s)}{(1-K_i)\cdot\left(\sum_n p(T_i=n)\cdot\prod_j p\left(D_{i^*j}=s', A_j | T_i=n, I_i, M_i, K_i, \theta_{jis's}^{(t)}\right)\right)+K_i\cdot\delta(s'=s)} \tag{6.4}$$

where $\delta\left(s' = s\right)$ is the Dirac delta function (probability that the known label for voxel i of the truth segmentation is s). Replacing the approximated density function, we obtain:

$$\mathrm{W}_{si}^{(t)} \equiv \frac{(1-K_i)\cdot\left(p(T_i=s)\cdot\prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{jis's}^{(t)}\cdot\alpha_{ji'i}\right)+K_i\cdot\delta(s'=s)}{(1-K_i)\cdot\left(\sum_n p(T_i=n)\cdot\prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{jis's}^{(t)}\cdot\alpha_{ji'i}\right)+K_i\cdot\delta(s'=s)} \tag{6.5}$$

M-step

In all the STAPLE family of methods, the calculated $W_{si}^{(t)}$ are used in this step to obtain the performance level parameters of the next iteration, i.e. $\theta_{ji}^{(t+1)}$, by maximizing the expectation of the complete data log likelihood. As seen before in Section 4.2.2, since complete data log likelihood is not observable, it can replaced by its conditional expectation given the observable data, which in this case are D, A, I, M and K, using the current estimate θ.

$$\theta_{ji}^{(t+1)} = \sum_{\theta_{ji} i' \in \mathcal{B}_i} E\left[\ln\left(p\left(D_j, A_j | T_{i'}, I_{i'}, M_{i'}, K_{i'}, \theta_{ji} | D, A, I, M, K, \theta^{(t)}\right)\right)\right]$$

$$= \sum_{\theta_{ji}} \sum_{i' \in \mathcal{B}_i} \sum_{s} p\left(T_{i'} = s | D, A, I, M, K, \theta^{(t)}\right) \cdot \ln\left(p\left(D_j, A_j | T_{i'}, I_{i'}, M_{i'}, K_{i'}, \theta_{ji}\right)\right)$$

$$= \sum_{\theta_{ji}} \sum_{i' \in \mathcal{B}_i} \sum_{s} W_{si'}^{(t)} \cdot \ln\left(p\left(D_{i^* j} = s', A_j | T_{i'}, I_{i'}, M_{i'}, K_{i'}, \theta_{ji}\right)\right) \quad (6.6)$$

$$= \sum_{\theta_{ji}} \sum_{i' \in \mathcal{B}_i} \sum_{s} W_{si'}^{(t)} \cdot \ln\left(\sum_{i'' \in \mathcal{N}(i'): D_{i'' j} = s'} \theta_{jis's} \cdot \alpha_{ji''i'}\right)$$

By solving this equation, we obtain

$$\theta_{jis's}^{(t+1)} = \frac{\sum_{i' \in \mathcal{B}_i} \left(\sum_{i'' \in \mathcal{N}(i'): D_{i'' j} = s'} \alpha_{ji''i'}\right) \cdot W_{si'}^{(t)}}{\sum_{i' \in \mathcal{B}_i} W_{si'}^{(t)}} \quad (6.7)$$

6.2.2 Masked Joint Label Fusion (mk-JLF)

In the case of m-JLF, we propose to include a probabilistic known mask instead of a binary one, in order to make the segmentation algorithm more flexible. Thus, $K \in [0,1]^{L' \times N}$, where K_{si} represents the probability that the true label associated with voxel i is s, i.e. $K_{si} = p\left(T_i = s\right)$, and L' is the number of known labels. Note that if all their probabilities are set to 1, the known mask will work as a binary mask. Furthermore, we have decided to include a confidence term, $mProb \in [0,1]$, that will represent how much confidence we have in the value K. In this way, if the mask comes from another automatic algorithm on which we have, let's say 70% confidence, we can specify it in our approach, which will trust the known mask in a 70% weighting and the atlases decision by 30%. Note also that if $mProb$ is set to 0, the known mask will be neglected and the algorithm will behave in the same way as our previous proposal, i.e. m-JLF.

In order to include the known mask into the estimation process, once the final weights W have been obtained following the steps described in Section 4.2.3, it is necesary to normalize them to the rank $[0,1]$ and make sure that $\sum_s W_{si} = 1$. This step is crucial, since the new mask probabilities $K_{si} \in [0,1]$. In the previous version of the algorithm it was not strictly necessary that the weights followed a valid probability mass function, since the final segmentation (S) could be obtained from the labels with the largest assigned weight, i.e. $S_i = \underset{s \in \mathcal{L}}{argmax}\left(W_{si}\right)$. However, since here we need to assign a confidence factor to the old weights, we have to make sure that they are normalized:

$$W_{si}{}^{norm} = \frac{W_{si} - \min_{s' \in \mathcal{L}} (W_{s'i})}{W_{si} - \left(\min_{s' \in \mathcal{L}} (W_{s'i}) \cdot L \right)} \tag{6.8}$$

where $\sum_s W_{si}{}^{norm} = 1$, L is the number of labels available in the atlases, i.e in the previous labelset $\mathcal{L}$, and $\min_{s' \in \mathcal{L}} (W_{s'i})$ is the smallest W_{si} for the target voxel i, corresponding to the label s with the lowest probability to be the true target segmentation T_i.

Once the old weights are normalized (W^{norm}), we update their values with the new information offered by the known mask K, i.e. W^k, as follows:

$$W^k{}_{s'i} = K_{s'i} \cdot mProb + W_{s'i}{}^{norm} \cdot (1 - mProb), \qquad \forall s' \in \mathcal{L}' \tag{6.9}$$

where $\mathcal{L}' = \{0, \ldots, L' - 1\}$ is the set of labels for which K_{si} is known (known labels).

Then, the remaining $W_{si}{}^{norm}$'s, for which their values have not been updated in the previous step, will be updated as follows:

$$W^k{}_{si} = \frac{W_{si}{}^{norm} \cdot \left(1 - \sum_{s' \in \mathcal{L}'} W_{s'i} \right)}{1 - \sum_{s' \in \mathcal{L}'} W_{si}{}^{norm}}, \qquad \forall s \in \mathcal{L} - \mathcal{L}' \tag{6.10}$$

where $\sum_{s' \in \mathcal{L}'} W_{s'i}$ is the sum of all the updated $W_{s'i}$ in the previous step for the target voxel i, and $\sum_{s' \in \mathcal{L}'} W_{si}{}^{norm}$ is the sum of the same terms previous being updated.

6.3 Results

In this section, we present some qualitative results obtained from the application of the proposed strategies to a set of MR images. First, we show how the methods perform when applied to MS patient images, and the effect of the incorporation of the known mask on their output. Then, we apply the methods to a small cohort of images with MRI visible lesions, from patients with diseases different from MS, and analyze the results.

Figure 6.2 depicts the segmentation results for the Non-local Spatial STAPLE family of methods. The image shows the result of segmenting the brain structures on the T1-w sequence from an MS patient (Figure 6.2 (a)) with all the NLSS variants presented in Chapters 4 and 6. When the target image is segmented with the

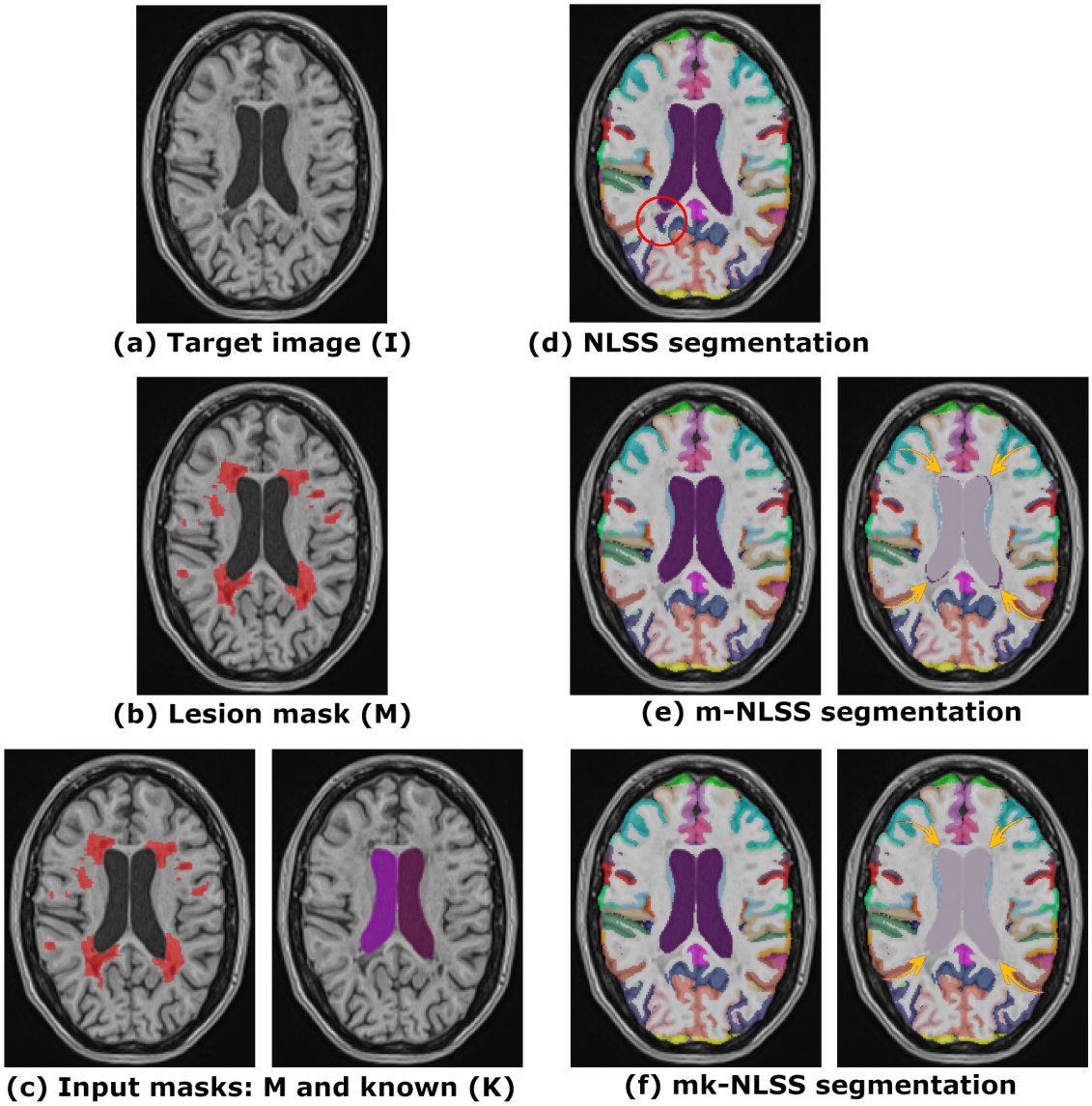


Figure 6.2: Non-local Spatial STAPLE family of methods. Example of the application of the M and K masks to an MS patient image. Segmentation of (a) the target image with (d) the original NLSS algorithm, (e) the proposed m-NLSS with (b) the lesion mask (M) as input parameter, and (f) the proposed mk-NLSS algorithm with (c) the lesion mask (M) and the known mask (K) correponding to the binary segmentation of the lateral ventricles (LV).

original NLSS method (Figure 6.2 (d)), we observe that some lesion voxels are mis-classified as part of the lateral ventricles (red circle). However, this mis-classification is corrected with the proposed m-NLSS (see Section 4.2.2), as depicted in Figure

6.2 (e). In that case, m-NLSS has been given the lesion mask shown in Figure 6.2 (b) as an input parameter, which allows the method to establish better voxel correspondences with the atlases, ignoring the lesion intensities, which are similar to those of the CSF. Nevertheless, even though the lateral ventricles (LV) segmentation has improved, it is still not perfect, as can be observed from the super-imposed segmentation on the right side of Figure 6.2 (e), where the yellow arrows show an over-segmentation of this structure. Since the real segmentation for the LV is known, we can use it as an input argument of the proposed mk-NLSS, combined with the lesion mask, as shown in Figure 6.2 (c). mk-NLSS (Figure 6.2 (f)) changes the label for the over-segmented voxels and correctly classifies them as WM, as indicated by the yellow arrows.

Figure 6.3 presents the segmentation results obtained for the Joint Label Fusion family of proposed methods. In this case the target image comes from another MS patient, Figure 6.3 (a), for which the probabilistic segmentation of both thalami is available. When segmenting this image with the original JLF method (Figure 6.3 (d)), we encounter the same mis-classification problem seen for NLSS. An atlas-target correspondence error, caused by the abnormal lesion intensities, produces an erroneous segmentation estimation, classifying some lesion voxels as part of the LV. As we previously saw for m-NLSS, our proposal for JLF, i.e. m-JLF, also corrects this mis-classification, as can be observed from Figure 6.3 (e). On the other hand, as we mentioned before, we have a probabilistic segmentation for both thalami, to which we give a confidence level of 100%. If we overlap this segmentation on the result obtained with m-JLF, as shown on the right of Figure 6.3 (e), we observe that both thalami are slightly over-segmented, as indicated by the yellow arrows. Since the probabilistic thalamus segmentation is known beforehand, by applying mk-JLF with both known masks (Figure 6.3 (c)) and $mProb = 1$, we correct the segmentation result for these structures, as shown in Figure 6.3 (f).

6.3.1 Application to other diseases

Even though the proposed methods having been designed with the goal of segmenting brains of MS patients, they can also be applied to patients of other diseases which present small MRI visible lesions, such as diabetes or lupus. Here, we show a couple of examples where the proposed strategies have been used to segment brain images with MRI visible lesions, which are not caused by MS.

Figure 6.4 depicts an example of a patient with diabetes, presenting WM lesions similar to MS. As expected, the behavior of the analyzed methods when segmenting this subject is similar to that of the previously seen examples. The image shows (a) the target image and the corresponding segmentation results of the original (d) Non-local Spatial STAPLE (NLSS) and (g) Joint Label Fusion (JLF) methods. As can

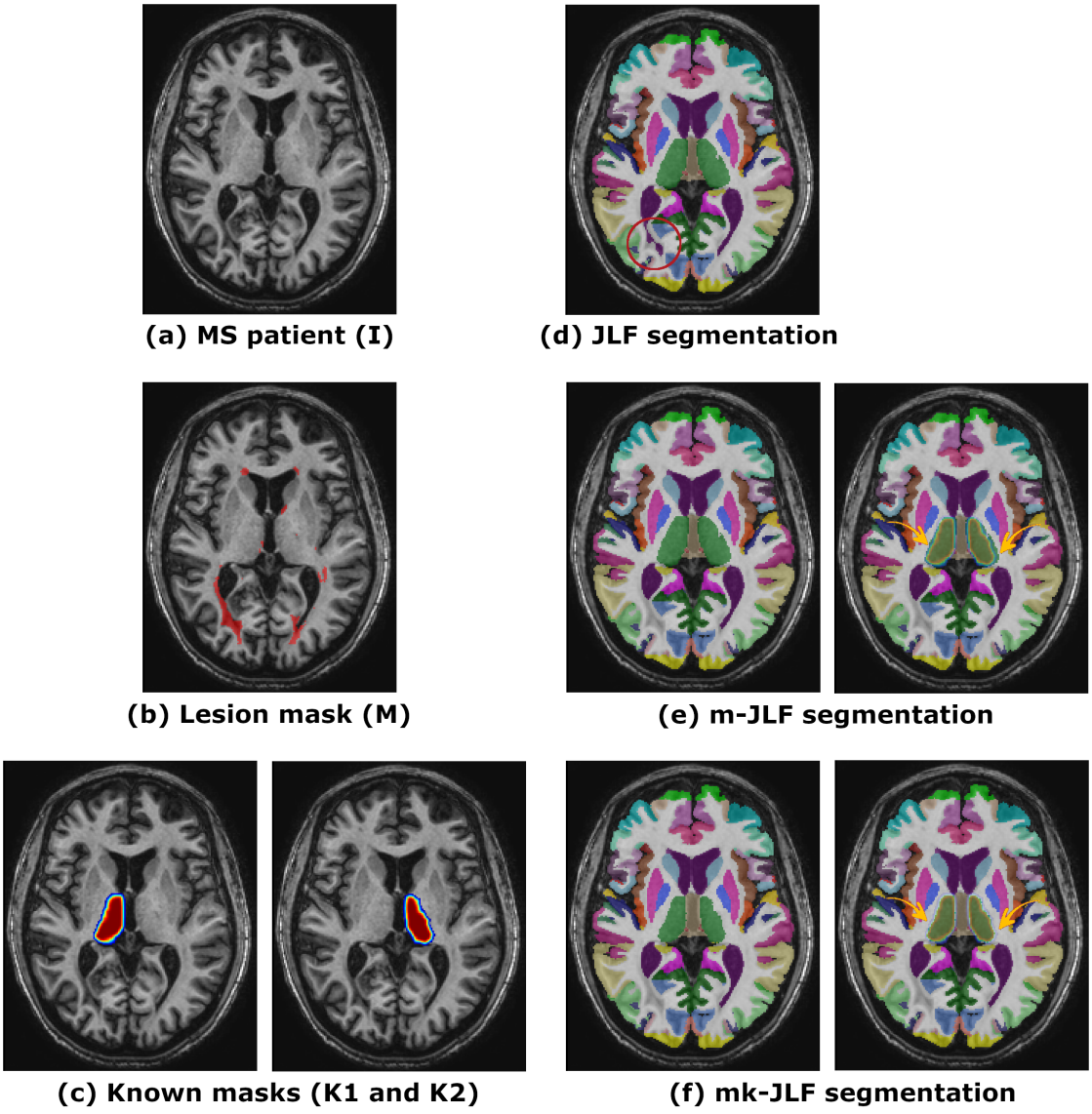


Figure 6.3: Joint Label Fusion family of methods. Example of the application of the M and K masks to an MS patient image. Segmentation of (a) the target image with (d) the original JLF algorithm, (e) the proposed m-JLF algorithm with (b) the lesion mask (M) as input parameter, and (f) the proposed mk-JLF algorithm with (b) the lesion mask (M) and (c) the known masks (K1 and K2), corresponding to the probabilistic segmentation of left and and right thalamus, as input arguments. Hotter colors in K indicate higher probabilities.

be observed from these images, the original methods mis-classify some lesion areas (yellow arrows) as either CSF (lateral ventricles) or GM (cerebral cortex). As seen

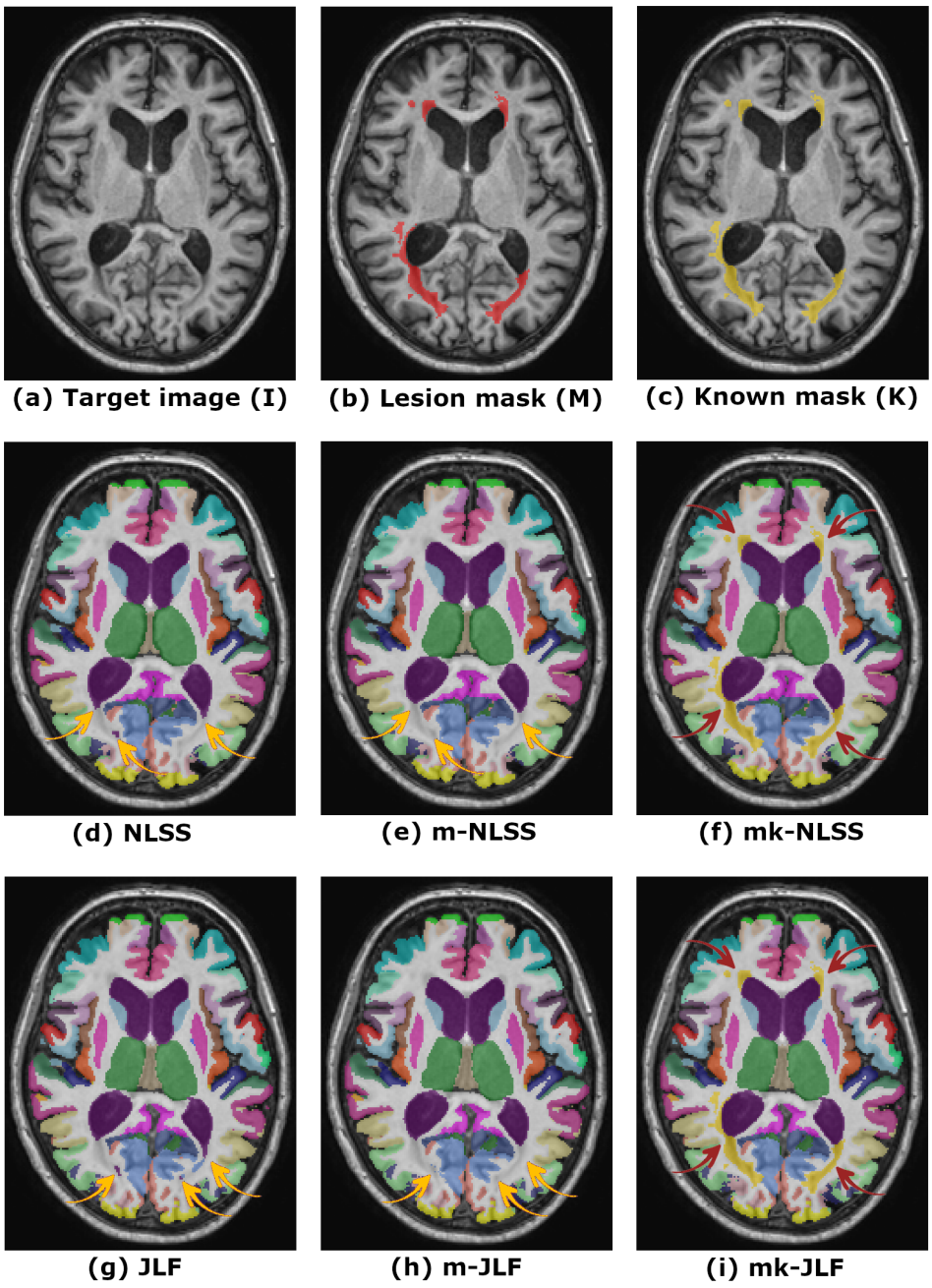


Figure 6.4: Example of application of the proposed strategies to a patient with diabetes. Results of segmenting (a) the target image with the original algorithms (d) NLSS and (g) JLF, the proposed strategies (e) m-NLSS and (h) m-JLF with (b) the lesion mask as input argument, and (f) mk-NLSS and (i) mk-JLF with the same mask as both (b) lesion mask and (c) known mask.

before for MS patients, the proposed strategies (e) m-NLSS and (h) m-JLF correct these labels, thanks to the new correspondence models, and correctly classify those regions as WM. Finally, the result of adding a new lesion class to the segmentation without the need of post-processing steps (highlighted by the red arrows) is accomplished by means of the (f) mk-NLSS and (i) mk-JLF proposals and using (b) the lesion mask also as a (c) known mask.

Until now, we have only evaluated how the analyzed strategies perform when demyelinating lesions are present. For this reason, in Figure 6.5 we show an example of a patient with an infarction, combined with several WM lesions. As depicted in this figure, when segmenting this image with the original methods, i.e. NLSS and JLF, shown in Figure 6.5 (d) and (g), some WM lesions are incorrectly classified as part of the lateral ventricles, as previously seen in other examples. Furthermore, since the infarction appears dark in the T1-w image, as we see in the lower part of Figure 6.5 (a), and also, it is close to the cortical area, these methods classify it as part of the cerebral cortex. On the other hand, when segmenting this image with the proposed strategies, i.e. m-NLSS and m-JLF, we observe from Figure 6.5 (e) and (f) that the previously mis-classified WM lesion voxels are correctly labelled. Furthermore, based on the atlas information instead of the lesion intensities, the proposed methods estimate that the infarction was produced on the WM, thus, both strategies classify this lesion as WM. However, if we want to be correct in the classification, an infarction is an area of necrotic tissue, and therefore, a new label, different from the ones existing in the atlases, should be incorporated to the estimation process. The result of using the infarction mask (Figure 6.5 (c)) as a known label mask is depicted in Figure 6.5 (f) and (i) for both mk-NLSS and mk-JLF (red arrow).

6.4 Discussion

In this chapter, we have extended the theory of the strategies proposed in Chapter 4, i.e. m-NLSS and m-JLF, to include the use of known label masks into the segmentation estimation process. Qualitative results of the application of all the proposed methods on a set of patients with different diseases, including MS, diabetes and infarction, have been analyzed. The results show that the proposed strategies tend to be more accurate in their segmentation estimations than their corresponding original methods.

The incorporation of known masks into the label fusion process allows us produce better segmentation estimations, since, if a label for a concrete structure from the atlases is known beforehand, the segmentation result for the surrounding structures will be better adjusted. Moreover, with the proposed confidence term $mProb$, the

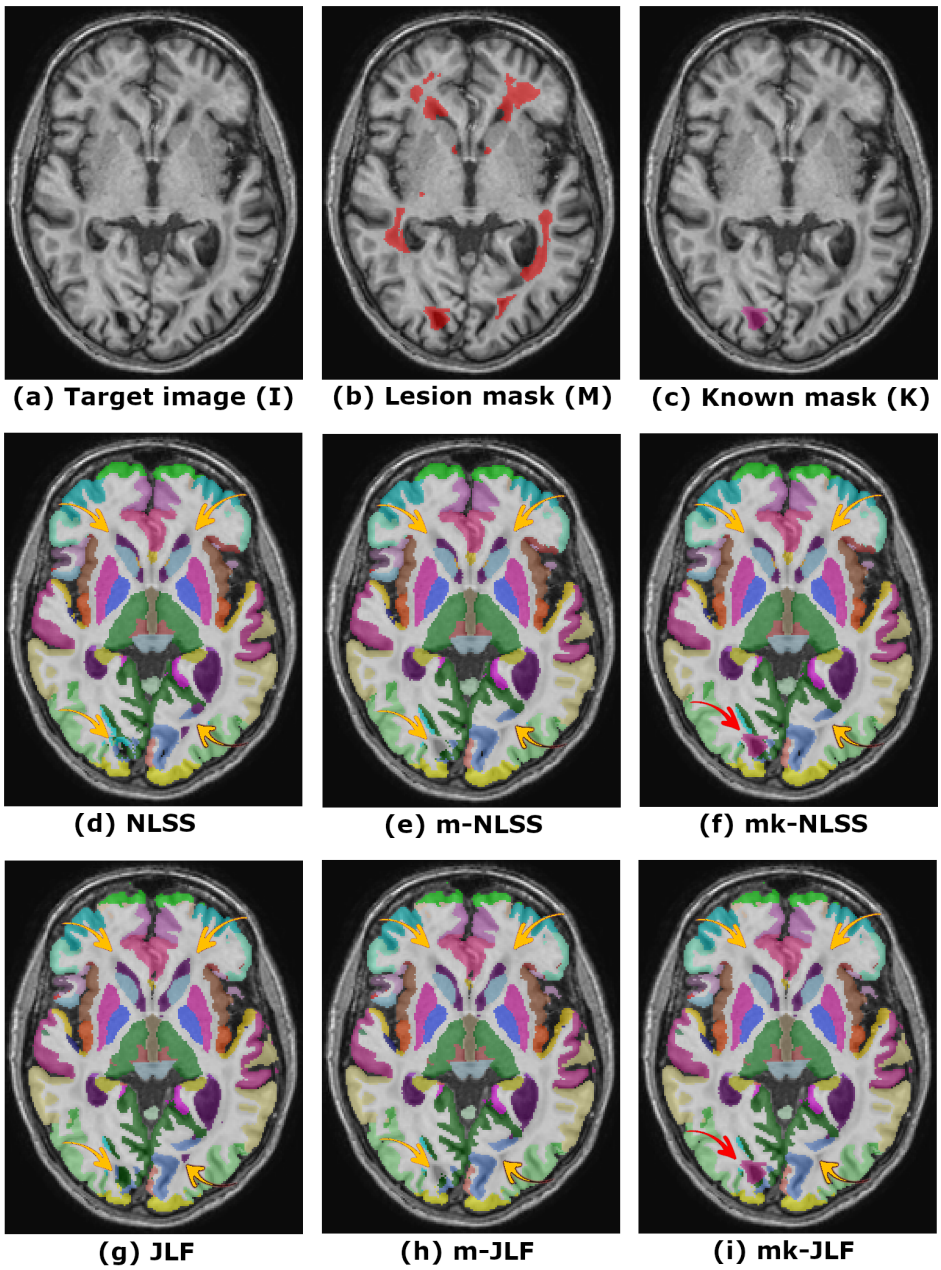


Figure 6.5: Example of application of the proposed strategies to a patient with an infaction and and white matter lesions. Results of segmenting (a) the target image with the original algorithms (d) NLSS and (g) JLF, the proposed strategies (e) m-NLSS and (h) m-JLF with (b) the lesion mask as input argument, and (f) mk-NLSS and (i) mk-JLF with (b) the lesion mask and (c) the infaction mask as input parameters.

weight that the K mask has in the final label assignation decision, as opposed to the atlases, can be easily controlled. This allows us to specify how reliable we think the provided known labels are, depending, for example, on the source of origin.

Additionally, the integration of these masks permits the incorporation of new labels, that are not included in the atlases, into the target segmentation. This enables the possibility of incorporating segmentation estimations from structure-specific algorithms [44, 45, 77, 126, 175] that provide labels for the internal sub-regions of the atlas structures, obtaining combined segmentations for the whole brain, with detailed structure information.

On the other hand, when the known masks are combined with the previously integrated lesion mask (M), these algorithms can be applied to brain images of patients with MRI visible lesions. As seen in the presented examples, combined algorithms, with both M and K masks, are useful in correcting mis-classified structure labels, but also allow the natural incorporation of new classes, such as lesion or necrotic tissue, into the estimation process without the need of post-processing steps.

In conclusion, we have presented the theory of how to include masks of abnormal anatomy, that model the patient-atlas correspondences, combined with masks of prior knowledge about the "real" voxel labels, into modern statistical fusion approaches. Together, these innovations improve the performance of the original fusion methods when segmenting brains with MRI irregularities, such as MS lesions, in addition to enabling the integration of manual and automatic *edits*.

Chapter 7

Conclusions and future work

7.1 Summary of the thesis

The aim of this thesis has been the proposal of new methods capable of classifying brain structures in brains of MS patients. Starting with an initial study of the current state of the art of brain structure segmentation on MRI, we realized that the vast majority of methods were intended to segment either healthy subject images or images of patients without brain lesions, and therefore were not oriented to MS patients. On the other hand, we also realized that most of the methods proposed were centered on a small group of structures or even on a single structure, which forced us to use more than one method if we wanted to segment the whole brain.

After this thorough study, we proposed a classification of the methods based on their segmentation strategy, which includes atlas-based, learning-based, deformable, region-based and hybrid approaches. We further analyzed the advantages and drawbacks of each segmentation strategy with the aim of better understanding their strengths and weaknesses, in order to obtain a global overview which has helped us in establishing the direction of our segmentation proposal.

A further analysis of the most commonly used databases and metrics for evaluation in brain structure segmentation was also presented, as well as a summary of the most relevant segmentation competitions, i.e. challenges, organized with regard to this topic. In addition, quantitative results of all the reviewed proposals, based on the results found in the literature, were also carefully analyzed, providing an overview of the state of the art from both the point of view of the segmentation strategies and the segmented target structures.

In order to unify the results and perform a fair comparison of the accuracy of the methods, a quantitative evaluation of three popular and publicly available software tools, i.e. FreeSurfer [43], FIRST [40] and MABMIS [150], each relying on a different category of the proposed classification, was performed on a dataset of healthy subjects that incorporated manual brain structure annotations, with the objective of establishing a baseline on the performance of the methods under ideal conditions.

Since the main goal here was to propose a brain structure segmentation approach able to deal with MS lesions, the next step of this thesis was to quantify to which extent those lesions affect different brain structure segmentation strategies. The aim of this step was to better understand the weaknesses of each strategy when segmenting MS patients, which we believed to be beneficial for a robust design of our segmentation proposal. For this reason, we performed an extensive analysis of the lesion effect on three well-known and publicly available segmentation methods, i.e. FreeSurfer [43], FIRST [40] and majority vote label fusion [139], which belonged to different categories of our classification. Such analysis was based on the segmentation strategy used, the lesion location, the affected structures and the total lesion load, and demonstrates that MS lesions indeed have a direct effect on the performance of the evaluated methods.

The main findings of that study revealed that the learning-based approach was the most affected by the presence of lesions, whereas the deformable method provided the most consistent segmentations, closely followed by the atlas-based strategy, in a comparison between healthy subjects and simulated MS patients. Furthermore, we observed that global registration procedures, such as the ones performed in the atlas-based and the learning-based approaches, allowed the errors produced by brain irregularities, i.e. lesions, to propagate to other parts of the brain and consequently affect the segmentation performance independently of the lesion location. On the other hand, local registration procedures, such as the one used in the deformable method, in which a mask that determined whether a voxel was included or not in the calculation of the similarity function, allowed the registration to concentrate only on the alignment of a region of the image. And therefore, if the lesions were located outside the mask, they did not interfere with the registration result.

Findings from those previous analyses, combined with the lack of labeled images for training, made us concentrate on an atlas-based strategy. In that case, we proposed a masked label fusion approach for multi-atlas segmentation in which lesions were excluded during the atlas registration procedure. As seen in the state-of-the-art analysis performed on the first part of this thesis, label fusion strategies that incorporated information on the structural image intensities, i.e. atlases and target, provided more accurate segmentation results on healthy subjects than the ones that simply used the propagated labels. In these strategies, the structural im-

ages were used to compensate for registration inaccuracies, thus establishing more accurate correspondences between the atlases and each voxel of the target image, based on their patch intensity similarity. Unfortunately, when the target image contained abnormal intensities, such as MS lesions, the correspondence search model used in those strategies, which relied on the target-atlas intensities, tended to fail on the lesion areas and consequently affected the segmentation result. In order to overcome this problem, we proposed a new correspondence search strategy that dynamically redefined the patch shape and size to prevent the abnormal lesion intensities from interfering in the search. Then, we adapted this strategy to fit two state-of-the-art statistical fusion algorithms, i.e. Non-local Spatial STAPLE [250] and Joint Label Fusion [251], defining a new correspondence model for each strategy and extending their corresponding theory to incorporate our models into their segmentation-estimation pipelines.

We further evaluated the proposed strategies on a set of simulated MS patients and analyzed their performance in comparison to the original algorithms. The results showed a significant improvement of the proposed methods on the lesion areas, which was also reflected in the whole brain segmentation performance, achieving significantly better results for the WM and the sub-cortical and cortical GM in both the proposed strategies. On the other hand, we analyzed to which extent the pre-processing technique of lesion filling [31], i.e. replace the lesions with normal-appearing white-matter intensities, improved the performance of the original label fusion strategies. For this purpose, we compared the results of segmenting the lesion-filled images using the original methods to the ones obtained with the proposed strategies when segmenting the un-pre-processed images. This analysis revealed a significant improvement with our methods around the lesion areas, whereas no significant performance differences were found on the lesion voxels themselves for any of the strategies. However, this improvement around the lesion areas was reflected globally, obtaining more accurate results with the proposed strategies for the whole brain, with a significant performance increase on both the WM and the GM. The results of these analyses demonstrate that with the proposed algorithms, we improve the segmentation results of their corresponding original methods, while eliminating the pre-processing steps required by lesion filling.

The lesion masks used in this experiment to test the proposed masked methods and to generate the lesion-filled images, besides performing the atlas registration, were segmented manually. However, manual lesion masks are not always available in practice besides requiring human expert interaction to obtain them. Alternatively, several automatic lesion segmentation algorithms have been proposed in recent years to accomplish this task [14, 15, 16, 17, 18, 19]. For this reason, and given that both lesion filling and the new strategies proposed rely on accurate lesion masks for their success, in a second experiment we analyzed how the use of these automatic lesion masks affected the segmentation result of the previously analyzed approaches. To do

so, we evaluated both approaches, i.e. original methods on the lesion-filled images and proposed methods on the non-pre-processed images, on a cohort of 15 MS patient images, that included images acquired with different scanners and magnetic field strengths. The 15 images were segmented twice with each brain parcellation strategy, first, using the consensus lesion mask from seven different trained experts and then, using the lesion mask obtained from the SLS method [16], an automatic lesion segmentation algorithm of the current state of the art. In a comparison of the results obtained with both masks, i.e. manual vs. automatic, we saw that the proposed methods were indifferent to the lesion mask used, showing no significant structure-volume differences for any of the two proposals. In contrast, when the original methods were used on the lesion-filled images, significant volume differences on the CSF, the cerebellum GM and the cerebral WM appeared.

Finally, we extended the theory of the proposed methods to incorporate known label masks into the segmentation estimation process. These known masks are labels which are known beforehand for a concrete brain structure or region, either because they come from an expert segmentation or from a automatic structure-specific algorithm. Such incorporation allows us to integrate new labels, that are not included in the atlases, into the segmentation estimation, without the need of post-processing steps, while dynamically correcting the estimated segmentation of the surrounding structures. Finally, qualitative evaluation summarizing the approaches analyzed in this thesis, on a set of patients with different diseases, including MS, diabetes and infarction, was performed.

7.1.1 Contributions

The goal of this thesis is to aid radiologists and neurologists in day-to-day practise by assisting them in the challenging and time-consuming task of segmenting brain structures.

From this point of view, the main contributions of this thesis for both the scientific and medical community are:

- An extensive review of the automatic and semi-automatic brain structure segmentation algorithms in the literature, and a corresponding proposal for their classification, based on the segmentation strategy used.
- A revision of the most commonly used databases and metrics for evaluation of brain structure segmentation algorithms, as well as an overview of the most relevant competitions, i.e. challenges, performed during the last years.
- A quantitative review of the state-of-the-art brain structure segmentation methods, based on the results found in the literature, followed by a quan-

titative evaluation of three popular software tools, i.e. FreeSurfer, FIRST and MABMIS, that belong to the most relevant categories of our classification.

- An analysis of the effect of MS lesions on several state-of-the-art automatic brain structure segmentation approaches, each relying on a different category of the proposed classification, separated by: lesion location, affected structures, total lesion load and segmentation strategy.
- A pipeline proposal for synthetic MS lesion generation, used to create the artificial patients in our experiments.
- A new correspondence model proposal for intensity-based multi-atlas segmentation, which is able to deal with MRI-visible lesions, and the corresponding theory extension to incorporate it into two well-known label fusion algorithms: Non-local Spatial STAPLE (NLSS) and Joint Label Fusion (JLF).
- The evaluation of the extended label fusion algorithms, i.e. m-NLSS and m-JLF, on a set of MS patients and their comparison to the original strategies, i.e. NLSS and JLF, when segmenting both the same patients and the patients after performing lesion filling.
- An analysis of the effect of automatically segmenting MS lesions, compared to the use of manual lesion masks, on the performance of the proposed strategies, and its comparison to the original strategies on the lesion-filled images.
- A theory extension of the proposed label fusion algorithms to incorporate a second mask, i.e. the 'known mask', that allows us to introduce manual and automatic label edits into the segmentation-estimation process, and its corresponding qualitative evaluation on a set of patients with different diseases.

7.1.2 International research stay

While carrying out the work presented in this PhD thesis, I had the opportunity of spending 5 months on a research stay at the Medical-image Analysis and Statistical Interpretation (MASI) lab, at the Vanderbilt University (Nashville, TN), under the supervision of Dr. Bennett Landman. During this time, I became better familiarized with the multi-atlas strategies and their corresponding pre-processing steps. I tested their performance on several MS patients, being aware of their limitations, which led to the development of the label fusion proposal for Non-local Spatial STAPLE presented in this thesis. This algorithm was first presented in a talk at the MICCAI Patch-MI Workshop 2018.

7.2 Future work

The analysis of brain MR images for MS patients is a complex topic involving several aspects and multiple research lines. This notion is exemplified in this PhD thesis by the several steps that are involved in the multi-atlas segmentation process of MS patients prior to the label fusion itself. Furthermore, some of the concepts applied to MS patients can be applied to other MR-imaging fields or can be studied further. Besides this, other interesting topics arise from the needs of current clinical practise for MS patients.

Hence, future directions are presented divided into two categories: those related to improve our proposal, and long term future research lines departing from this thesis.

7.2.1 Short-term proposal improvements

In this thesis, we presented two reformulated label fusion algorithms to segment brain structures on images of patients with abnormal brain anatomy. As seen through most of the chapters, multi-atlas strategies require two main steps that are strictly necessary to achieve a segmentation result. In the first, a set of altases are registered to the target image, and then, in a second step, their propagated labels are fused to obtain the final segmentation. The bigger disadvantage of multi-atlas segmentation is its computational cost, being the registration step highly time consuming in comparison to the second one, and increasing with the number of atlases used. Given the recent success of deep-learning registration proposals in the medical imaging community, and its high speed in comparison to traditional algorithms, the first possible improvement could be the substitution of the registration algorithms used, i.e. nifyreg and ANTs, by a deep-learning strategy [265, 266, 267]. As stated earlier, patch-based multi-atlas techniques do not require very accurate registration results, and thus, a fast algorithm with an acceptable precision would be enough for our needs. However, it would be essential that the method can perform masked registrations, since all the theory presented in this thesis is based on a registration that excludes the lesion voxels from the similarity-metric calculation.

Another short-term improvement could be to calculate several measures and statistics from the segmentation result obtained with the proposed methods to support doctors in the task of patient follow-up and integrate it with the fully automated pipeline presented in Chapter 5. We plan to calculate the voxel-based cortical thickness based on the estimated segmentation, as well as to automatically generate a report containing the most relevant results, such as lesion load, number of lesions, lesion location, and cortical thickness and structure volumes in comparison to normal

controls. Furthermore, when longitudinal studies are available, measures related to the basal image, such as structure volume changes, regional cortical thinning and lesion follow-up, may also be an interesting source of information to include. These measures may be helpful for the medical community, since strong correlations have been reported between the local atrophy of isolated structures, cortical thickness or lesion volume, and the progression of disability in MS [268, 269, 270].

Finally, we would like to analyze the effect of other automatic lesion segmentation strategies in the result of the proposed algorithms, as well as their robustness with respect to the different lesion masks. By performing such analysis, we would be able to select the strategy that generates the lesion masks which better fit our structure segmentation methods, thus facilitating the search for the best possible estimate, and with this aim in mind, we could integrate such a strategy into our fully automated pipeline.

7.2.2 Future research lines

In the long term, there are several new research lines departing from this thesis that could be studied by our research group. For example, the analysis of the involvement of individual brain structures in the MS disease. Here, we have addressed the structure segmentation problem for a single scan, however, the study of these segmentations at different time points would be of special interest to accomplish this goal. As previously noted in this thesis, thalamus atrophy has already been proved to be a clinically relevant biomarker of the neurodegenerative disease process, therefore the study of the thalamic nuclei regions across longitudinal data seems a good point to focus on.

Another possible study derived from this thesis is the use of the information gathered from the methods analyzed here in combination with the lesion analysis and other biological markers to predict the clinical evolution of MS, in terms of both clinical history and treatment response. Moreover, combining the structure segmentation obtained with the proposed methods with the lesion masks, provides us with an overview of the lesion location within the brain, which in conjunction with the previously mentioned information and the patient symptoms opens up a very attractive field of research.

As already seen during the development of this thesis, there is a lack of public databases of MS patients with both manual lesion annotations and brain structure ground truth on which automatic segmentation algorithms can be tested. Therefore, the construction of a standardized and publicly available dataset of patients with different diseases, including MS, with reliable annotations of lesions and structures, would not only be of special interest for testing all the methods presented in this thesis, but also very useful to the scientific community. Furthermore, with the

increasing success of the deep learning approaches for medical image analysis, which need large amounts of data for their training, the creation of the mentioned database would be an incentive for new method proposals able to achieve more accurate results.

Finally, the methods and concepts presented here could also be applied to the study of other diseases with similar properties. Although we have presented a couple of examples of the application of our methods to patient images of other diseases, a deeper study involving different diseases and lesion properties would be interesting. Patients with lupus, whose lesions appear hypo-intense in T1-w sequence and can develop near the lateral ventricles, or patients of stroke, whose lesions present very variable sizes, location and morphology, may be interesting groups to study.

Bibliography

[1] R. Milo and E. Kahana, "Multiple sclerosis: Geoepidemiology, genetics and the environment," *Autoimmunity Reviews*, vol. 9, no. 5, pp. A387–A394, 2010.

[2] S. G. Waxman, *Multiple sclerosis as a neuronal disease*. Academic Press, 2005.

[3] T. K. Vartanian, "MS as a neurodegenerative disease (a thought experiment): The clinical evidence," *Johns Hopkins Advanced Studies in Medicine*, vol. 8, no. 9, pp. 305–308, 2008.

[4] A. J. Thompson, S. E. Baranzini, J. Geurts, B. Hemmer, and O. Ciccarelli, "Multiple sclerosis," *The Lancet*, vol. 391, no. 10130, pp. 1622–1636, 2018.

[5] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, K. Fujihara, S. L. Galetta, H. P. Hartung, L. Kappos, F. D. Lublin, R. A. Marrie, A. E. Miller, D. H. Miller, X. Montalban, E. M. Mowry, P. S. Sorensen, M. Tintorè, A. L. Traboulsee, M. Trojano, B. M. J. Uitdehaag, S. Vukusic, E. Waubant, B. G. Weinshenker, S. C. Reingold, and J. A. Cohen, "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria," *The Lancet Neurology*, vol. 17, no. 2, pp. 162–173, 2018.

[6] W. J. Brownlee, T. A. Hardy, F. Fazekas, and D. H. Miller, "Diagnosis of multiple sclerosis: Progress and challenges," *The Lancet*, vol. 389, no. 10076, pp. 1336–1346, 2017.

[7] A. Compston and A. Coles, "Multiple sclerosis," *The Lancet*, vol. 372, no. 9648, pp. 1502–1517, 2008.

[8] H. F. Harbo, R. Gold, and M. Tintorè, "Sex and gender issues in multiple sclerosis," *Therapeutic advances in neurological disorders*, vol. 6, no. 4, pp. 237–248, 2013.

[9] W. I. McDonald, A. Compston, G. Edan, D. Goodkin, H.-P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. Van Den Noort, B. Y. Weinshenker, and J. S. Wolinsky, "Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis," *Annals of Neurology*, vol. 50, no. 1, pp. 121–127, 2001.

[10] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H.-P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, B. G. Weinshenker, and J. S. Wolinsky, "Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald criteria"," *Annals of Neurology*, vol. 58, no. 6, pp. 840–846, 2005.

[11] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. D. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, E. Waubant, B. Weinshenker, and J. S. Wolinsky, "Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria," *Annals of Neurology*, vol. 69, no. 2, pp. 292–302, 2011.

[12] I. D. Kilsdonk, L. E. Jonkman, R. Klaver, S. J. van Veluw, J. J. Zwanenburg, J. P. Kuijer, P. J. Pouwels, J. W. Twisk, M. P. Wattjes, P. R. Luijten, F. Barkhof, and J. J. Geurts, "Increased cortical grey matter lesion detection in multiple sclerosis with 7 T MRI: A post-mortem verification study," *Brain*, vol. 139, no. 5, pp. 1472–1481, 2016.

[13] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and A. Rovira, "Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches," *Information Sciences*, vol. 186, no. 1, pp. 164–185, 2012.

[14] N. Guizard, P. Coupé, V. Fonov, J. Manjón, D. Arnold, and D. Collins, "Rotation-invariant multi-contrast non-local means for MS lesion segmentation," *NeuroImage: Clinical*, vol. 8, pp. 376–389, 2015.

[15] H. Deshpande, P. Maurel, and C. Barillot, "Classification of multiple sclerosis lesions using adaptive dictionary learning," *Computerized Medical Imaging and Graphics*, vol. 46, pp. 2–10, 2015.

[16] E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó, "A toolbox for multiple sclerosis lesion segmentation," *Neuroradiology*, vol. 57, no. 10, pp. 1031–1043, 2015.

[17] X. Tomas-Fernandez and S. K. Warfield, "A Model Of Population and Subject (MOPS) intensities with application to multiple sclerosis lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 34, no. 6, pp. 1349–1361, 2015.

[18] R. Harmouche, N. Subbanna, D. Collins, D. Arnold, and T. Arbel, "Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood information," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1281–1292, 2015.

[19] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. Vilanova, L. Ramió-Torrentà, Á. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[20] M. Filippi, P. Preziosa, M. Copetti, G. Riccitelli, M. Horsfield, V. Martinelli, G. Comi, and M. Rocca, "Gray matter damage predicts the accumulation of disability 13 years later in MS," *Neurology*, vol. 81, no. 20, pp. 1759–1767, 2013.

[21] E. Fisher, J. Lee, K. Nakamura, and R. Rudick, "Gray matter atrophy in multiple sclerosis: A longitudinal study," *Annals of Neurology*, vol. 64, no. 3, pp. 255–265, 2008.

[22] K. Pohl, J. Fisher, W. Grimson, R. Kikinis, and W. Wells, "A Bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, no. 1, pp. 228–239, 2006.

[23] S. Roy, A. Carass, P. Bazin, S. Resnick, and J. Prince, "Consistent segmentation using a Rician classifier," *Med. Image Anal.*, vol. 16, no. 2, pp. 524–535, 2012.

[24] B. Caldairou, N. Passat, P. Habas, C. Studholme, and F. Rousseau, "A non-local fuzzy segmentation method: application to brain MRI," *Pattern Recog.*, vol. 44, no. 9, pp. 1916–1927, 2011.

[25] H. Vrooman, F. van der Lijn, and W. Niessen, "Auto-kNN: Brain tissue segmentation using automatically trained k-nearest-neighbor classification," in *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS'13)*, 2013.

[26] A. van Opbroek, F. van der Lijn, and M. de Bruijne, "Automated brain-tissue segmentation by multi-feature SVM classification," in *Proceedings of the MICCAI Workshops—The MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS'13)*, 2013.

[27] M. Rajchl, J. Baxter, A. McLeod, J. Yuan, W. Qiu, T. Peters, and A. Khan, "Hierarchical max-flow segmentation framework for multi-atlas segmentation with Kohonen self-organizing map based Gaussian mixture modeling," *Med. Image Anal.*, vol. 27, pp. 45–56, 2016.

[28] S. Valverde, A. Oliver, Y. Díez, M. Cabezas, J. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó, "Evaluating the effects of white matter multiple sclerosis lesions on the volume estimation of 6 brain tissue segmentation methods," *Am. J. Neuroradiol.*, vol. 36, no. 6, pp. 1109–1115, 2015.

[29] M. Battaglini, M. Jenkinson, and N. De Stefano, "Evaluating and reducing the impact of white matter lesions on brain volume measurements," *Hum. Brain Mapp.*, vol. 33, no. 9, pp. 2062–2071, 2012.

[30] D. T. Chard, J. S. Jackson, D. H. Miller, and C. A. Wheeler-Kingshott, "Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes," *J. Magn. Reson. Imaging*, vol. 32, no. 1, pp. 223–228, 2010.

[31] S. Valverde, A. Oliver, and X. Lladó, "A white matter lesion-filling approach to improve brain tissue volume measurements," *NeuroImage: Clinical*, vol. 6, pp. 86–92, 2014.

[32] V. Popescu, N. Ran, F. Barkhof, D. Chard, C. Wheeler-Kingshott, and H. Vrenken, "Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks," *NeuroImage: Clinical*, vol. 4, pp. 366–373, 2014.

[33] C. Jacobsen, J. Hagemeier, K.-M. Myhr, H. Nyland, K. Lode, N. Bergsland, D. P. Ramasamy, T. O. Dalaker, J. P. Larsen, E. Farbu, *et al.*, "Brain atrophy and disability progression in multiple sclerosis patients: A 10-year follow-up study," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 85, pp. 1109–1115, 2014.

[34] N. Bergsland, D. Horakova, M. Dwyer, O. Dolezal, Z. Seidl, M. Vaneckova, J. Krasensky, E. Havrdova, and R. Zivadinov, "Subcortical and cortical gray matter atrophy in a large sample of patients with clinically isolated syndrome and early relapsing-remitting multiple sclerosis," *Am. J. Neuroradiol.*, vol. 33, no. 8, pp. 1573–1578, 2012.

[35] C. A. Bishop, R. D. Newbould, J. S. Lee, L. Honeyfield, R. Quest, A. Colasanti, R. Ali, M. Mattoscio, A. Cortese, R. Nicholas, P. M. Matthews, P. A. Muraro, and A. D. Waldman, "Analysis of ageing-associated grey matter volume in patients with multiple sclerosis shows excess atrophy in subcortical regions," *NeuroImage: Clinical*, vol. 13, pp. 9–15, 2017.

[36] A. Minagar, M. H. Barnett, R. H. Benedict, D. Pelletier, I. Pirko, M. A. Sahraian, E. Frohman, and R. Zivadinov, "The thalamus and multiple sclerosis: Modern views on pathologic, imaging, and clinical aspects," *Neurology*, vol. 80, no. 4, pp. 210–219, 2013.

[37] M. Houtchens, R. Benedict, R. Killiany, J. Sharma, Z. Jaisani, B. Singh, B. Weinstock-Guttman, C. Guttmann, and R. Bakshi, "Thalamic atrophy and cognition in multiple sclerosis," *Neurology*, vol. 69, no. 12, pp. 1213–1223, 2007.

[38] M. Kim, G. Wu, and D. Shen, "Unsupervised deep learning for hippocampus segmentation in 7.0 Tesla MR images," *Machine Learning in Medical Imaging*, vol. 8184, pp. 1–8, 2013.

[39] K. O. Babalola, V. Petrovic, T. F. Cootes, C. J. Taylor, C. J. Twining, and A. Mills, "Automatic segmentation of the caudate nuclei using active appearance models," in *3D Segmentation in the clinic: A grand challenge*, pp. 57–64, 2007.

[40] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, "A Bayesian model of shape and appearance for subcortical brain segmentation," *NeuroImage*, vol. 56, no. 3, pp. 907–922, 2011.

[41] O. Benkarim, P. Radeva, and L. Igual, "Label consistent multiclass discriminative dictionary learning for MRI segmentation," *Articulated Motion and Deformable Objects*, vol. 8563, pp. 138–147, 2014.

[42] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.

[43] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[44] J. E. Iglesias, K. V. Leemput, P. Bhatt, C. Casillas, S. Dutt, N. Schuff, D. Truran-Sacrey, A. Boxer, and B. Fischl, "Bayesian segmentation of brainstem structures in MRI," *NeuroImage*, vol. 113, pp. 184–195, 2015.

[45] C. R. Traynor, G. J. Barker, W. R. Crum, S. C. Williams, and M. P. Richardson, "Segmentation of the thalamus in MRI based on T1 and T2," *NeuroImage*, vol. 56, no. 3, pp. 939–950, 2011.

[46] E. Roura, A. Oliver, M. Cabezas, J. Vilanova, A. Rovira, L. Ramió-Torrentà, and X. Lladó, "MARGA: Multispectral Adaptive Region Growing Algorithm for brain extraction on axial MRI," *Comput. Meth. Prog. Biomed.*, vol. 113, no. 2, pp. 655–673, 2014.

[47] M. Cabezas, A. Oliver, E. Roura, J. Freixenet, J. Vilanova, L. Ramió-Torrentà, A. Rovira, and X. Lladó, "Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding," *Comput. Meth. Prog. Biomed.*, vol. 115, no. 3, pp. 147–161, 2014.

[48] M. Cabezas, A. Oliver, S. Valverde, B. Beltran, J. Freixenet, J. Vilanova, L. Ramió-Torrentà, A. Rovira, and X. Lladó, "BOOST: A supervised approach for multiple sclerosis lesion segmentation," *Journal of Neuroscience Methods*, vol. 237, pp. 108–117, 2014.

[49] E. Roura, N. Sarbu, A. Oliver, S. Valverde, S. González-Villà, R. Cervera, N. Bargalló, and X. Lladó, "Automated detection of lupus white matter lesions in MRI," *Frontiers in Neuroinformatics*, vol. 10, p. 33, 2016.

[50] O. Ganiler, A. Oliver, Y. Diez, J. Freixenet, J. Vilanova, B. Beltran, L. Ramió-Torrentà, A. Rovira, and X. Lladó, "A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies," *Neuroradiology*, vol. 56, no. 5, pp. 363–374, 2014.

[51] X. Lladó, O. Ganiler, A. Oliver, R. Martí, J. Freixenet, L. Valls, J. Vilanova, L. Ramió-Torrentà, and A. Rovira, "Automated detection of multiple sclerosis lesions in serial brain MRI," *Neuroradiology*, vol. 54, pp. 787–807, 2012.

[52] Y. Diez, A. Oliver, M. Cabezas, S. Valverde, R. Martí, J. Vilanova, L. Ramió-Torrentà, A. Rovira, and X. Lladó, "Intensity based methods for brain MRI longitudinal registration. A study on multiple sclerosis patients," *Neuroinformatics*, vol. 12, no. 3, pp. 365–379, 2014.

[53] E. Roura, T. Schneider, M. Modat, P. Daga, N. Muhlert, D. Chard, S. Ourselin, X. Lladó, and C. Wheeler-Kingshott, "Multi-channel registration of FA and T1-w images in the presence of atrophy: Application to multiple sclerosis," *Functional Neurology*, vol. 30, no. 4, pp. 245–256, 2015.

[54] S. Valverde, A. Oliver, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, Àlex Rovira, and X. Lladó, "Automated tissue segmentation of MR brain images in the presence of white matter lesions," *Med. Image Anal.*, vol. 35, pp. 446–457, 2017.

[55] S. González-Villà, A. Oliver, S. Valverde, L. Wang, R. Zwiggelaar, and X. Lladó, "A review on brain structures segmentation in magnetic resonance imaging," *Artificial Intelligence in Medicine*, vol. 73, pp. 45–69, 2016.

[56] S. González-Villà, S. Valverde, M. Cabezas, D. Pareto, J. Vilanova, L. Ramió-Torrentà, Á. Rovira, A. Oliver, and X. Lladó, "Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation," *NeuroImage: Clinical*, vol. 15, pp. 228–238, 2017.

[57] S. González-Villà, S. Valverde, M. Cabezas, D. Pareto, J. Vilanova, L. Ramió-Torrentà, Á. Rovira, A. Oliver, and X. Lladó, "Do multiple sclerosis lesions affect automatic brain structure segmentation?," *Multiple Sclerosis*, vol. 23, no. S3, pp. 257–258, 2017.

[58] S. González-Villà, A. Oliver, Y. Huo, X. Lladó, and B. A. Landman, "Brain structure segmentation in the presence of multiple sclerosis lesions," *NeuroImage: Clinical*, vol. 22, p. 101709, 2019.

[59] S. González-Villà, Y. Huo, A. Oliver, X. Lladó, and B. A. Landman, "Multi-atlas parcellation in the presence of lesions: Application to multiple sclerosis," in *International Workshop on Patch-based Techniques in Medical Imaging*, pp. 104–113, 2018.

[60] M. Filippi, M. A. Rocca, O. Ciccarelli, N. DeStefano, N. Evangelou, L. Kappos, A. Rovira, J. Sastre-Garriga, M. Tintorè, J. L. Frederiksen, C. Gasperini, J. Palace, D. S. Reich, B. Banwell, X. Montalban, and F. Barkhof, "MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines," *The Lancet Neurology*, vol. 15, no. 3, pp. 292–303, 2016.

[61] C. Jacobsen, J. Hagemeier, K.-M. Myhr, H. Nyland, K. Lode, N. Bergsland, D. P. Ramasamy, T. O. Dalaker, J. P. Larsen, E. Farbu, and R. Zivadinov, "Brain atrophy and disability progression in multiple sclerosis patients: A 10-year follow-up study," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 85, no. 10, pp. 1109–1115, 2014.

[62] N. C. Andreasen, D. Liu, S. Ziebell, A. Vora, and B.-C. Ho, "Relapse duration, treatment intensity, and brain tissue loss in schizophrenia: A prospective longitudinal MRI study," *The American Journal of Psychiatry*, vol. 170, no. 6, pp. 609–615, 2013.

[63] R. C. Knickmeyer, S. Gouttard, C. Kang, D. Evans, K. Wilber, J. K. Smith, R. M. Hamer, W. Lin, G. Gerig, and J. H. Gilmore, "A structural MRI study of human brain development from birth to 2 years," *The Journal of Neuroscience*, vol. 28, no. 47, pp. 12176–12182, 2008.

[64] D. L. Collins and A. C. Evans, "ANIMAL: Validation and applications of non-linear registration-based segmentation," *Int. J. Pattern Recogn. Artif. Intell.*, vol. 11, no. 8, pp. 1271–1294, 1997.

[65] S. Warfield, K. Zou, and W. Wells, "Simultaneous Truth and Performance Level Estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, 2004.

[66] K. Pohl, S. Bouix, M. Nakamura, T. Rohlfing, R. McCarley, R. Kikinis, W. Grimson, M. Shenton, and W. Wells, "A hierarchical algorithm for MR brain image parcellation," *IEEE Trans. Med. Imag.*, vol. 26, no. 9, pp. 1201–1212, 2007.

[67] Y. Hao, T. Wang, X. Zhang, Y. Duan, C. Yu, T. Jiang, Y. Fan, and A. D. N. Initiative, "Local Label Learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation," *Hum. Brain Mapp.*, vol. 35, no. 6, pp. 2674–2697, 2014.

[68] V. Dill, A. R. Franco, and M. S. Pinho, "Automated methods for hippocampus segmentation: The evolution and a review of the state of the art," *Neuroinformatics*, vol. 13, no. 2, pp. 133–150, 2015.

[69] M.-P. Hosseini, M. Nazem-Zadeh, D. Pompili, and H. Soltanian-Zadeh, "Statistical validation of automatic methods for hippocampus segmentation in MR images of epileptic patients," in *Engineering in Medicine and Biology Society, EMBC, Annual International Conference of the IEEE*, pp. 4707–4710, 2014.

[70] J. Zhou and J. C. Rajapakse, "Segmentation of subcortical brain structures using fuzzy templates," *NeuroImage*, vol. 28, no. 4, pp. 915–924, 2005.

[71] X. Lin, T. Qiu, F. Morain-Nicolier, and S. Ruan, "A topology preserving non-rigid registration algorithm with integration shape knowledge to segment brain subcortical structures from MRI images," *Pattern Recog.*, vol. 43, no. 7, pp. 2418–2427, 2010.

[72] J. Dolz, L. Massoptier, and M. Vermandel, "Segmentation algorithms of subcortical brain structures on MRI for radiotherapy and radiosurgery: A survey," *IRBM*, vol. 36, no. 4, pp. 200–212, 2015.

[73] K. O. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. Cootes, M. Jenkinson, and D. Rueckert, "An evaluation of four automatic methods of segmenting the subcortical structures in the brain," *NeuroImage*, vol. 47, no. 4, pp. 1435–1447, 2009.

[74] C. N. Devi, A. Chandrasekharan, V. Sundararaman, and Z. C. Alex, "Neonatal brain MRI segmentation: A review," *Comput. Biol. Med.*, vol. 64, pp. 163–178, 2015.

[75] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[76] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[77] L. Igual, J. Soliva, R. Gimeno, S. Escalera, O. Vilarroya, and P. Radeva, "Automatic internal segmentation of caudate nucleus for diagnosis of attention-deficit/hyperactivity disorder," *Image Analysis and Recognition*, vol. 7325, pp. 222–229, 2012.

[78] R. Kikinis, M. E. Shenton, D. V. Iosifescu, R. W. McCarley, P. Saiviroonporn, H. H. Hokama, A. Robatino, D. Metcalf, C. G. Wible, C. M. Portas, R. M. Donnino, and F. A. Jolesz, "A digital brain atlas for surgical planning, model-driven segmentation, and teaching," *IEEE Trans. Visual. Comput. Graphics*, vol. 2, no. 3, pp. 232–241, 1996.

[79] F. Castro, C. Pollo, R. Meuli, P. Maeder, O. Cuisenaire, M. Cuadra, J.-G. Villemure, and J.-P. Thiran, "A cross validation study of deep brain stimulation targeting: From experts to atlas-based, segmentation-based and automatic registration algorithms," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1440–1450, 2006.

[80] J. L. Phillips, L. A. Batten, P. Tremblay, F. Aldosary, and P. Blier, "A prospective, longitudinal study of the effect of remission on cortical thickness and hippocampal volume in patients with treatment-resistant depression," *International Journal of Neuropsychopharmacology*, vol. 18, no. 8, 2015.

[81] A. Pitiot, H. Delingette, P. M. Thompson, and N. Ayache, "Expert knowledge-guided segmentation system for brain MRI," *NeuroImage*, vol. 23, no. Supplement 1, pp. S85–S96, 2004.

[82] N. Fox, E. Warrington, P. Freeborough, P. Hartikainen, A. Kennedy, J. Stevens, and M. N. Rossor, "Presymptomatic hippocampal atrophy in Alzheimer's disease," *Brain*, vol. 119, no. 6, pp. 2001–2007, 1996.

[83] E. Geuze, E. Vermetten, and J. Bremner, "MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed," *Molecular Psychiatry*, vol. 10, no. 2, pp. 147–159, 2005.

[84] L. L. Altshuler, G. Bartzokis, T. Grieder, J. Curran, and J. Mintz, "Amygdala enlargement in bipolar disorder and hippocampal reduction in schizophrenia: An MRI study demonstrating neuroanatomic specificity," *Archives of General Psychiatry*, vol. 55, no. 7, pp. 663–664, 1998.

[85] N. Bernasconi, S. Duchesne, A. Janke, J. Lerch, D. Collins, and A. Bernasconi, "Whole-brain voxel-based statistical analysis of gray matter and white matter in temporal lobe epilepsy," *NeuroImage*, vol. 23, no. 2, pp. 717–723, 2004.

[86] G. Villarreal, D. A. Hamilton, H. Petropoulos, I. Driscoll, L. M. Rowland, J. A. Griego, P. W. Kodituwakku, B. L. Hart, R. Escalona, and W. M. Brooks, "Reduced hippocampal volume and total white matter volume in posttraumatic stress disorder," *Biological Psychiatry*, vol. 52, no. 2, pp. 119–125, 2002.

[87] N. Kitayama, V. Vaccarino, M. Kutner, P. Weiss, and J. D. Bremner, "Magnetic resonance imaging (MRI) measurement of hippocampal volume in posttraumatic stress disorder: A meta-analysis," *Journal of Affective Disorders*, vol. 88, no. 1, pp. 79–86, 2005.

[88] J. D. Bremner, M. Narayan, E. R. Anderson, L. H. Staib, H. L. Miller, and D. S. Charney, "Hippocampal volume reduction in major depression," *The American Journal of Psychiatry*, vol. 157, no. 1, pp. 115–118, 2000.

[89] L. L. Altshuler, G. Bartzokis, T. Grieder, J. Curran, and J. Mintz, "Amygdala enlargement in bipolar disorder and hippocampal reduction in schizophrenia: An MRI study demonstrating neuroanatomic specificity," *Archives of General Psychiatry*, vol. 55, no. 7, pp. 663–664, 1998.

[90] S. Strakowski, M. Delbello, and C. Adler, "The functional neuroanatomy of bipolar disorder: A review of neuroimaging findings," *Molecular Psychiatry*, vol. 10, no. 1, pp. 105–116, 2005.

[91] H. P. Blumberg, J. Kaufman, A. Martin, R. Whiteman, J. H. Zhang, J. C. Gore, D. S. Charney, J. H. Krystal, and B. S. Peterson, "Amygdala and hippocampal volumes in adolescents and adults with bipolar disorder," *Archives of General Psychiatry*, vol. 60, no. 12, pp. 1201–1208, 2003.

[92] A. L. Boxer, M. D. Geschwind, N. Belfor, M. L. Gorno-Tempini, G. F. Schauer, B. L. Miller, M. W. Weiner, and H. J. Rosen, "Patterns of brain atrophy that differentiate corticobasal degeneration syndrome from progressive supranuclear palsy," *Archives of Neurology*, vol. 63, no. 1, pp. 81–86, 2006.

[93] R. P. Bote and M. Fernández-Gil, "Degeneration of the brainstem," in *Seminars in Ultrasound, CT and MRI*, vol. 34, pp. 142–152, 2013.

[94] K. A. Josephs, "Key emerging issues in progressive supranuclear palsy and corticobasal degeneration," *Journal of Neurology*, vol. 262, no. 3, pp. 783–788, 2015.

[95] D. R. Williams and A. J. Lees, "Progressive supranuclear palsy: Clinicopathological concepts and diagnostic challenges," *The Lancet Neurology*, vol. 8, pp. 270–79, 2009.

[96] L. T. Grinberg, U. Rüb, R. E. L. Ferretti, R. Nitrini, J. M. Farfel, L. Polichiso, K. Gierga, W. Jacob-Filho, H. Heinsen, and B. B. B. S. Group, "The dorsal raphe nucleus shows phospho-tau neurofibrillary changes before the transentorhinal region in Alzheimer's disease. A precocious onset?," *Neuropathology and Applied Neurobiology*, vol. 35, no. 4, pp. 406–416, 2009.

[97] W. Sako, N. Murakami, Y. Izumi, and R. Kaji, "MRI can detect nigral volume loss in patients with Parkinson's disease: Evidence from a meta-analysis," *Journal of Parkinson's disease*, vol. 4, no. 3, pp. 405–411, 2014.

[98] J. H. Lee, J. Ryan, C. Andreescu, H. Aizenstein, and H. K. Lim, "Brainstem morphological changes in Alzheimer's disease," *NeuroReport*, vol. 26, no. 7, pp. 411–415, 2015.

[99] E. H. Aylward, A. M. Codori, A. Rosenblatt, M. Sherr, J. Brandt, O. C. Stine, P. E. Barta, G. D. Pearlson, and C. A. Ross, "Rate of caudate atrophy in presymptomatic and symptomatic stages of Huntington's disease," *Movement Disorders*, vol. 15, no. 3, pp. 552–560, 2000.

[100] A. Peinemann, S. Schuller, C. Pohl, T. Jahn, A. Weindl, and J. Kassubek, "Executive dysfunction in early stages of Huntington's disease is associated with striatal and insular atrophy: A neuropsychological and voxel-based morphometric study," *Journal of the Neurological Sciences*, vol. 239, no. 1, pp. 11–19, 2005.

[101] J. W. Mink, "Neurobiology of basal ganglia and Tourette syndrome: basal ganglia circuits and thalamocortical outputs," *Advances in Neurology*, vol. 99, pp. 89–98, 2006.

[102] E. Hollander, E. Anagnostou, W. Chaplin, K. Esposito, M. M. Haznedar, E. Licalzi, S. Wasserman, L. Soorya, and M. Buchsbaum, "Striatal volume on magnetic resonance imaging and repetitive behaviors in autism," *Biological Psychiatry*, vol. 58, no. 3, pp. 226–232, 2005.

[103] M. H. Bloch, J. F. Leckman, H. Zhu, and B. S. Peterson, "Caudate volumes in childhood predict symptom severity in adults with Tourette syndrome," *Neurology*, vol. 65, no. 8, pp. 1253–1258, 2005.

[104] V. Tremols, A. Bielsa, J.-C. Soliva, C. Raheb, S. Carmona, J. Tomas, J.-D. Gispert, M. Rovira, J. Fauquet, A. Tobeña, A. Bulbena, and O. Vilarroya, "Differential abnormalities of the head and body of the caudate nucleus in attention deficit-hyperactivity disorder," *Psychiatry Research: Neuroimaging*, vol. 163, no. 3, pp. 270–278, 2008.

[105] S. Eliez, C. M. Blasey, L. S. Freund, T. Hastie, and A. L. Reiss, "Brain anatomy, gender and IQ in children and adolescents with fragile X syndrome," *Brain*, vol. 124, no. 8, pp. 1610–1618, 2001.

[106] N. C. Andreasen, S. Arndt, V. Swayze, T. Cizadlo, M. Flaum, D. O'Leary, J. C. Ehrhardt, and W. Yuh, "Thalamic abnormalities in schizophrenia visualized through magnetic resonance image averaging," *Science*, vol. 266, no. 5183, pp. 294–298, 1994.

[107] L. De Jong, K. Van der Hiele, I. Veer, J. Houwing, R. Westendorp, E. Bollen, P. De Bruin, H. Middelkoop, M. Van Buchem, and J. Van Der Grond, "Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study," *Brain*, vol. 131, no. 12, pp. 3277–3285, 2008.

[108] W. Byne, E. A. Hazlett, M. S. Buchsbaum, and E. Kemether, "The thalamus and schizophrenia: Current status of research," *Acta neuropathologica*, vol. 117, no. 4, pp. 347–368, 2009.

[109] S. Lee, S. Kim, W. Tae, S. Lee, J. Choi, S. Koh, and D. Kwon, "Regional volume analysis of the Parkinson disease brain in early disease stage: Gray matter, white matter, striatum, and thalamus," *Am. J. Neuroradiol.*, vol. 32, no. 4, pp. 682–687, 2011.

[110] A. Z. Kazi, P. C. Joshi, A. B. Kelkar, M. S. Mahajan, and A. S. Ghawate, "MRI evaluation of pathologies affecting the corpus callosum: A pictorial essay," *The Indian journal of radiology & imaging*, vol. 23, no. 4, pp. 321–332, 2013.

[111] P. M. Thompson, K. L. Narr, R. E. Blanton, and A. W. Toga, "Mapping structural alterations of the corpus callosum during brain development and degeneration," *The parallel brain: The cognitive neuroscience of the corpus callosum*, 2003.

[112] N. Garg, S. Reddel, D. Miller, J. Chataway, D. Riminton, Y. Barnett, L. Masters, M. Barnett, and T. Hardy, "The corpus callosum in the diagnosis of multiple sclerosis and other CNS demyelinating and inflammatory diseases," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 86, no. 12, pp. 1374–1382, 2015.

[113] J. E. Downhill, M. S. Buchsbaum, T. Wei, J. Spiegel-Cohen, E. A. Hazlett, M. M. Haznedar, J. Silverman, and L. J. Siever, "Shape and size of the corpus callosum in schizophrenia and schizotypal personality disorder," *Schizophrenia research*, vol. 42, no. 3, pp. 193–208, 2000.

[114] I. K. Lyoo, A. Satlin, C. K. Lee, and P. F. Renshaw, "Regional atrophy of the corpus callosum in subjects with Alzheimer's disease and multi-infarct dementia," *Psychiatry Research: Neuroimaging*, vol. 74, no. 2, pp. 63–72, 1997.

[115] S. M. Lawrie, H. C. Whalley, D. E. Job, and E. C. Johnstone, "Structural and functional abnormalities of the amygdala in schizophrenia," *Annals of the New York Academy of Sciences*, vol. 985, no. 1, pp. 445–460, 2003.

[116] M. P. DelBello, M. E. Zimmerman, N. P. Mills, G. E. Getz, and S. M. Strakowski, "Magnetic resonance imaging analysis of amygdala and other subcortical brain regions in adolescents with bipolar disorder," *Bipolar Disorders*, vol. 6, no. 1, pp. 43–52, 2004.

[117] M. P. Milham, A. C. Nugent, W. C. Drevets, D. S. Dickstein, E. Leibenluft, M. Ernst, D. Charney, and D. S. Pine, "Selective reduction in amygdala volume in pediatric anxiety disorders: A voxel-based morphometry investigation," *Biological Psychiatry*, vol. 57, no. 9, pp. 961–966, 2005.

[118] L. L. Altshuler, G. Bartzokis, T. Grieder, J. Curran, T. Jimenez, K. Leight, J. Wilkins, R. Gerner, and J. Mintz, "An MRI study of temporal lobe structures in men with bipolar disorder or schizophrenia," *Biological Psychiatry*, vol. 48, no. 2, pp. 147–162, 2000.

[119] S. M. Strakowski, M. P. DelBello, K. W. Sax, M. E. Zimmerman, P. K. Shear, J. M. Hawkins, and E. R. Larson, "Brain magnetic resonance imaging of structural abnormalities in bipolar disorder," *Archives of General Psychiatry*, vol. 56, no. 3, pp. 254–260, 1999.

[120] A. Sakalauskas, A. Lukoševičius, and K. Laučkaitė, "Transcranial echoscopy for diagnostic of Parkinson disease: Technical constraints and possibilities," *Ultragarsas*, vol. 65, pp. 47–50, 2010.

[121] J. Olveres, R. Nava, B. Escalante-Ramírez, G. Cristóbal, and C. M. García-Moreno, "Midbrain volume segmentation using active shape models and LBPs," *Proc. SPIE*, vol. 8856, 2013.

[122] A. Nayate, J. L. Bradshaw, and N. J. Rinehart, "Autism and Asperger's disorder: Are they movement disorders involving the cerebellum and/or basal ganglia?," *Brain Research Bulletin*, vol. 67, no. 4, pp. 327–334, 2005.

[123] L. L. Sears, C. Vest, S. Mohamed, J. Bailey, B. J. Ranson, and J. Piven, "An MRI study of the basal ganglia in autism," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 23, no. 4, pp. 613–624, 1999.

[124] L. J. Seidman, E. M. Valera, and N. Makris, "Structural brain imaging of attention-deficit/hyperactivity disorder," *Biological Psychiatry*, vol. 57, no. 11, pp. 1263–1272, 2005.

[125] Y. Xia, K. Bettinger, L. Shen, and A. Reiss, "Automatic segmentation of the caudate nucleus from human brain MR images," *IEEE Trans. Med. Imag.*, vol. 26, no. 4, pp. 509–517, 2007.

[126] M. Iacono, N. Makris, L. Mainardi, J. Gale, A. van der Kouwe, A. Mareyam, J. Polimeni, L. Wald, B. Fischl, E. Eskandar, and G. Bonmassar, "Atlas-based segmentation for globus pallidus internus targeting on low-resolution MRI," in *Engineering in Medicine and Biology Society, EMBC, Annual International Conference of the IEEE*, pp. 5706–5709, 2011.

[127] F. Seixas, A. de Souza, A. dos Santos, and D. Saade, "Automated segmentation of the corpus callosum midsagittal surface area," in *XX Brazilian Symposium on Computer Graphics and Image Processing*, pp. 287–293, 2007.

[128] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. Bach-Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images," *Comput. Meth. Prog. Biomed.*, vol. 104, no. 3, pp. e158–e177, 2011.

[129] D. L. Collins and J. C. Pruessner, "Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion," *NeuroImage*, vol. 52, no. 4, pp. 1355–1366, 2010.

[130] S. Duchesne, J. Pruessner, and D. Collins, "Appearance-based segmentation of medial temporal lobe structures," *NeuroImage*, vol. 17, no. 2, pp. 515–531, 2002.

[131] D. Shen and C. Davatzikos, "HAMMER: Hierarchical Attribute Matching Mechanism for Elastic Registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, 2002.

[132] G. Postelnicu, L. Zollei, and B. Fischl, "Combined volumetric and surface registration," *IEEE Trans. Med. Imag.*, vol. 28, no. 4, pp. 508–522, 2009.

[133] D. W. Shattuck and R. M. Leahy, "BrainSuite: An automated cortical surface identification tool," *Med. Image Anal.*, vol. 6, no. 2, pp. 129–142, 2002.

[134] F. Wang and B. Vemuri, "Simultaneous registration and segmentation of anatomical structures from brain MRI," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 3749, pp. 17–25, 2005.

[135] Y. Luo and A. Chung, "An atlas-based deep brain structure segmentation method: From coarse positioning to fine shaping," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1085–1088, 2011.

[136] S. Yousefi, N. Kehtarnavaz, and A. Gholipour, "Improved labeling of subcortical brain structures in atlas-based segmentation of magnetic resonance images," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 7, pp. 1808–1817, 2012.

[137] A. Joshi, D. Shattuck, and R. Leahy, "A method for automated cortical surface registration and labeling," *Biomedical Image Registration*, vol. 7359, pp. 180–189, 2012.

[138] M. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1714–1729, 2010.

[139] X. Artaechevarría, A. Muñoz-Barrutia, and C. Ortiz-de Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1266–1277, 2009.

[140] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[141] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Classifier selection strategies for label fusion using large atlas databases," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 4791, pp. 523–531, 2007.

[142] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.

[143] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation," *Med. Image Anal.*, vol. 17, no. 6, pp. 671–684, 2013.

[144] F. Rousseau, P. Habas, and C. Studholme, "A supervised patch-based approach for human brain labeling," *IEEE Trans. Med. Imag.*, vol. 30, no. 10, pp. 1852–1862, 2011.

[145] J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.

[146] R. A. Heckemann, S. Keihaninejad, P. Aljabar, D. Rueckert, J. V. Hajnal, and A. Hammers, "Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation," *NeuroImage*, vol. 51, no. 1, pp. 221–227, 2010.

[147] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert, "LEAP: Learning Embeddings for Atlas Propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, 2010.

[148] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, and D. Rueckert, "Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling," *NeuroImage*, vol. 76, pp. 11–23, 2013.

[149] D. Zhang, Q. Guo, G. Wu, and D. Shen, "Sparse patch-based label fusion for multi-atlas segmentation," *Multimodal Brain Image Analysis*, vol. 7509, pp. 94–102, 2012.

[150] H. Jia, P.-T. Yap, and D. Shen, "Iterative multi-atlas-based multi-image segmentation with tree-based registration," *NeuroImage*, vol. 59, no. 1, pp. 422–430, 2012.

[151] H. Wang and P. Yushkevich, "Multi-atlas segmentation with joint label fusion and corrective learning - An open source implementation," *Frontiers in Neuroinformatics*, vol. 7, no. 27, pp. 1–12, 2013.

[152] A. J. Asman and B. A. Landman, "Non-local statistical label fusion for multi-atlas segmentation," *Med. Image Anal.*, vol. 17, no. 2, pp. 194–208, 2013.

[153] J. Wang, C. Vachet, A. Rumple, S. Gouttard, C. Ouziel, E. Perrot, G. Du, X. Huang, G. Gerig, and M. A. Styner, "Multi-atlas segmentation of subcortical brain structures via the AutoSeg software pipeline," *Frontiers in Neuroinformatics*, vol. 8, no. 7, pp. 1–11, 2014.

[154] J. Pipitone, M. T. M. Park, J. Winterburn, T. A. Lett, J. P. Lerch, J. C. Pruessner, M. Lepage, A. N. Voineskos, and M. M. Chakravarty, "Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates," *NeuroImage*, vol. 101, pp. 494–512, 2014.

[155] G. Wu, M. Kim, G. Sanroma, Q. Wang, B. C. Munsell, and D. Shen, "Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition," *NeuroImage*, vol. 106, pp. 34–46, 2015.

[156] A. Pitiot, A. Toga, N. Ayache, and P. Thompson, "Texture based MRI segmentation with a two-stage hybrid neural classifier," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 2053–2058, 2002.

[157] S. C. Deoni, B. K. Rutt, A. G. Parrent, and T. M. Peters, "Segmentation of thalamic nuclei using a modified k-means clustering algorithm and high-resolution quantitative magnetic resonance imaging at 1.5 T," *NeuroImage*, vol. 34, no. 1, pp. 117–126, 2007.

[158] S. Bao and A. C. Chung, "Multi-scale structured CNN with label consistency for brain MR image segmentation," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 1, pp. 113–117, 2018.

[159] R. Mehta, A. Majumdar, and J. Sivaswamy, "BrainSegNet: A convolutional neural network architecture for automated segmentation of human brain structures," *Journal of Medical Imaging*, vol. 4, no. 2, pp. 024003–1–024003–11, 2017.

[160] M. Shakeri, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, and I. Kokkinos, "Sub-cortical brain structure segmentation using F-CNN's," in *IEEE Int. Symp. Biomed. Imag.*, pp. 269–272, 2016.

[161] J. Dolz, C. Desrosiers, and I. B. Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study," *NeuroImage*, vol. 170, pp. 456–470, 2018.

[162] C. Wachinger, M. Reuter, and T. Klein, "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, vol. 170, pp. 434–445, 2018.

[163] K. Kushibar, S. Valverde, S. González-Villà, J. Bernal, M. Cabezas, A. Oliver, and X. Lladó, "Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features," *Med. Image Anal.*, vol. 48, pp. 177–186, 2018.

[164] J. Morra, Z. Tu, L. Apostolova, A. Green, A. Toga, and P. Thompson, "Automatic subcortical segmentation using a contextual model," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 5241, pp. 194–201, 2008.

[165] M. Jabarouti Moghaddam and H. Soltanian-Zadeh, "Automatic segmentation of brain structures using geometric moment invariants and artificial neural networks," *Information Processing in Medical Imaging*, vol. 5636, pp. 326–337, 2009.

[166] M. Wels, Y. Zheng, G. Carneiro, M. Huber, J. Hornegger, and D. Comaniciu, "Fast and robust 3-D MRI brain structure segmentation," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 5762, pp. 575–583, 2009.

[167] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 10, pp. 1744–1757, 2010.

[168] T. Riklin-Raviv, K. V. Leemput, B. H. Menze, W. M. W. III, and P. Golland, "Segmentation of image ensembles via latent atlases," *Med. Image Anal.*, vol. 14, no. 5, pp. 654–665, 2010.

[169] M. R. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "Supervised nonparametric image parcellation," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 5762, pp. 1075–1083, 2009.

[170] C.-Y. Liu, J. Iglesias, and Z. Tu, "Deformable templates guided discriminative models for robust 3D brain MRI segmentation," *Neuroinformatics*, vol. 11, no. 4, pp. 447–468, 2013.

[171] B. Scherrer, M. Dojat, F. Forbes, and C. Garbay, "LOCUS: LOcal Cooperative Unified Segmentation of MRI brain scans," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 4791, pp. 219–227, 2007.

[172] B. Scherrer, M. Dojat, F. Forbes, and C. Garbay, "MRF agent based segmentation: Application to MRI brain scans," *Artificial Intelligence in Medicine*, vol. 4594, pp. 13–23, 2007.

[173] A. Akselrod-Ballin, M. Galun, J. M. Gomori, A. Brandt, and R. Basri, "Prior knowledge driven multiscale segmentation of brain MRI," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 118–126, 2007.

[174] B. Scherrer, F. Forbes, C. Garbay, and M. Dojat, "Distributed local MRF models for tissue and structure brain segmentation," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1278–1295, 2009.

[175] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl, "Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI," *Hippocampus*, vol. 19, no. 6, pp. 549–557, 2009.

[176] Q. Razlighi, A. Orekhov, A. Laine, and Y. Stern, "Causal Markov random field for brain MR image segmentation," in *Engineering in Medicine and Biology Society, EMBC, Annual International Conference of the IEEE*, pp. 3203–3206, 2012.

[177] J. E. Iglesias, M. R. Sabuncu, and K. V. Leemput, "Improved inference in Bayesian segmentation using Monte Carlo sampling: Application to hippocampal subfield volumetry," *Med. Image Anal.*, vol. 17, no. 7, pp. 766–778, 2013.

[178] A. Makropoulos, I. Gousias, C. Ledig, P. Aljabar, A. Serag, J. Hajnal, A. Edwards, S. Counsell, and D. Rueckert, "Automatic whole brain MRI segmentation of the developing neonatal brain," *IEEE Trans. Med. Imag.*, vol. 33, no. 9, pp. 1818–1831, 2014.

[179] A. Ghanei, H. Soltanian-Zadeh, and J. P. Windham, "Segmentation of the hippocampus from brain MRI using deformable contours," *Computerized Medical Imaging and Graphics*, vol. 22, no. 3, pp. 203–216, 1998.

[180] A. Kelemen, G. Szekely, and G. Gerig, "Elastic model-based segmentation of 3-D neuroradiological data sets," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 828–839, 1999.

[181] E. A. Ashton, J. K. Riek, L. Molinelli, M. J. Berg, and K. J. Parker, "A method for fully automated measurement of neurological structures in MRI," *Proc. SPIE*, vol. 5032, pp. 1125–1134, 2003.

[182] H. Shariatpanahi, N. Batmanghelich, A. Kermani, M. Ahmadabadi, and H. Soltanian-Zadeh, "Distributed behavior-based multi-agent system for automatic segmentation of brain MR images," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 4535–4542, 2006.

[183] O. Colliot, O. Camara, and I. Bloch, "Integration of fuzzy spatial relations in deformable models: Application to brain MRI segmentation," *Pattern Recog.*, vol. 39, no. 8, pp. 1401–1414, 2006.

[184] D. Zarpalas, A. Zafeiropoulos, P. Daras, N. Maglaveras, and M. G. Strintzis, "Brain structures segmentation using optimum global and local weights on mixing active contours and neighboring constraints," in *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies*, no. 127, pp. 1–5, 2011.

[185] J. Cerrolaza, A. Villanueva, and R. Cabeza, "Hierarchical statistical shape models of multiobject anatomical structures: Application to brain MRI," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 713–724, 2012.

[186] Y. Gao, B. Corn, D. Schifter, and A. Tannenbaum, "Multiscale 3D shape representation and segmentation with applications to hippocampal/caudate extraction from brain MRI," *Med. Image Anal.*, vol. 16, no. 2, pp. 374–385, 2012.

[187] G. Fouquier, J. Atif, and I. Bloch, "Sequential model-based segmentation and recognition of image structures driven by visual features and spatial relations," *Comput. Vis. Image Underst.*, vol. 116, no. 1, pp. 146–165, 2012.

[188] H. García, M. Álvarez, and Á. Orozco, "Bayesian shape models with shape priors for MRI brain segmentation," *Advances in Visual Computing*, vol. 8888, pp. 851–860, 2014.

[189] S. D. S. Al-Shaikhli, M. Y. Yang, and B. Rosenhahn, "Multi-region labeling and segmentation using a graph topology prior and atlas information in brain images," *Computerized Medical Imaging and Graphics*, vol. 38, no. 8, pp. 725–734, 2014.

[190] S. Ettaieb, K. Hamrouni, and S. Ruan, "Statistical models of shape and spatial relation-application to hippocampus segmentation," in *International Conference on Computer Vision Theory and Applications*, vol. 1, pp. 448–455, 2014.

[191] J.-H. Xue, S. Ruan, B. Moretti, M. Revenu, D. Bloyet, and W. Philips, "Fuzzy modeling of knowledge for MRI brain structure segmentation," in *International Conference on Image Processing*, vol. 1, pp. 617–620, 2000.

[192] J.-H. Xue, S. Ruan, B. Moretti, M. Revenu, and D. Bloyet, "Knowledge-based segmentation and labeling of brain structures from MRI images," *Pattern Recognit. Lett.*, vol. 22, pp. 395–405, 2001.

[193] L. Gui, R. Lisowski, T. Faundez, P. S. Hüppi, F. Lazeyras, and M. Kocher, "Morphology-driven automatic segmentation of MR images of the neonatal brain," *Med. Image Anal.*, vol. 16, no. 8, pp. 1565–1579, 2012.

[194] Z. Tu, K. Narr, P. Dollar, I. Dinov, P. Thompson, and A. Toga, "Brain anatomical structure segmentation by hybrid discriminative/generative models," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 495–508, 2008.

[195] K. Karsch, Q. He, and Y. Duan, "A fast, semi-automatic brain structure segmentation algorithm for magnetic resonance imaging," in *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 297–302, 2009.

[196] Q. He, K. Karsch, and Y. Duan, "Semi-automatic 3D segmentation of brain structures from MRI," *Int. J. Data Mining and Bioinformatics*, vol. 5, no. 2, pp. 158–173, 2011.

[197] N. I. Weisenfeld and S. K. Warfield, "Learning likelihoods for labeling (L3): A general multi-classifier segmentation algorithm," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 322–329, 2011.

[198] J. Iglesias, M. Sabuncu, and K. Van Leemput, "A generative model for probabilistic label fusion of multimodal data," *Multimodal Brain Image Analysis*, vol. 7509, pp. 115–133, 2012.

[199] F. van der Lijn, M. de Bruijne, S. Klein, T. den Heijer, Y. Hoogendam, A. van der Lugt, M. Breteler, and W. Niessen, "Automated brain structure segmentation based on atlas registration and appearance models," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 276–286, 2012.

[200] J. Iglesias, M. Sabuncu, and K. Van Leemput, "A probabilistic, nonparametric framework for inter-modality label fusion," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, vol. 8151, pp. 576–583, 2013.

[201] A. Valsecchi, S. Damas, J. Santamaría, and L. Marrakchi-Kacem, "Intensity-based image registration using scatter search," *Artificial Intelligence in Medicine*, vol. 60, no. 3, pp. 151–163, 2014.

[202] F. Wang, B. C. Vemuri, M. Rao, and Y. Chen, "A new & robust information theoretic measure and its application to image alignment," *Information Processing in Medical Imaging*, pp. 388–400, 2003.

[203] W. R. Crum, T. Hartkens, and D. L. G. Hill, "Non-rigid image registration: Theory and practice," *Brit. J. Radiol.*, vol. 77, no. Supplement 2, pp. 140–153, 2004.

[204] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. Collins, A. C. Evans, G. Malandain, N. Ayache, G. Christensen, and H. Johnson, "Retrospective evaluation of intersubject brain registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1120–1130, 2003.

[205] O. T. Carmichael, H. A. Aizenstein, S. W. Davis, J. T. Becker, P. M. Thompson, C. C. Meltzer, and Y. Liu, "Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 27, no. 4, pp. 979–990, 2005.

[206] R. Woods, S. Grafton, C. Holmes, S. Cherry, and J. Mazziotta, "Automated image registration: I. General methods and intrasubject, intramodality validation," *J. Comp. Assist. Tomo.*, vol. 22, no. 1, pp. 139–152, 1998.

[207] A. Klein, B. Mensh, S. Ghosh, J. Tourville, and J. Hirsch, "Mindboggle: Automated brain labeling with multiple atlases," *BMC Med. Imag.*, vol. 5, no. 7, pp. 1–14, 2005.

[208] N. Robitaille and S. Duchesne, "Label fusion strategy selection," *International Journal of Biomedical Imaging*, vol. 2012, pp. 1–13, 2011.

[209] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, "Generative or discriminative? Getting the best of both worlds," *Bayesian Statistics*, no. 8, pp. 3–23, 2007.

[210] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review," *Artificial Intelligence in Medicine*, 2018.

[211] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *IEEE Conf. Comput. Vision Pattern Recog. Workshops*, pp. 20–28, 2015.

[212] L. Igual, J. C. Soliva, A. Hernández-Vela, S. Escalera, X. Jiménez, O. Vilarroya, and P. Radeva, "A fully-automatic caudate nucleus segmentation of brain MRI: Application in volumetric analysis of pediatric attention-deficit/hyperactivity disorder," *Biomedical Engineering Online*, vol. 10, no. 1, p. 105, 2011.

[213] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[214] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[215] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - Their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, 1995.

[216] T. Cootes and C. Taylor, "Active shape models - 'Smart snakes'," in *Proc. British Machine Vision Conference*, pp. 266–275, 1992.

[217] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 681–685, 2001.

[218] J. Ashburner and K. Friston, "Multimodal image coregistration and partitioning - A unified framework," *NeuroImage*, vol. 6, no. 3, pp. 209–217, 1997.

[219] S. Valverde, A. Oliver, E. Roura, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, J. Sastre-Garriga, X. Montalban, À. Rovira, and X. Lladó, "Quantifying brain tissue volume in multiple sclerosis with automated lesion segmentation and filling," *NeuroImage: Clinical*, vol. 9, pp. 640–647, 2015.

[220] K. Nakamura, N. Guizard, V. S. Fonov, S. Narayanan, D. L. Collins, and D. L. Arnold, "Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis," *NeuroImage: Clinical*, vol. 4, pp. 10–17, 2014.

[221] K. A. Johnson, J. A. Becker, and L. Williams, "The whole brain atlas," 1999.

[222] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[223] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[224] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[225] A. Hammers, R. Allom, M. J. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks, and J. S. Duncan, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Hum. Brain Mapp.*, vol. 19, no. 4, pp. 224–247, 2003.

[226] I. S. Gousias, D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers, "Automatic segmentation of brain MRIs of

2-year-olds into 83 regions of interest," *NeuroImage*, vol. 40, no. 2, pp. 672–684, 2008.

[227] A. Hammers, C.-H. Chen, L. Lemieux, R. Allom, S. Vossos, S. L. Free, R. Myers, D. J. Brooks, J. S. Duncan, and M. J. Koepp, "Statistical neuroanatomy of the human inferior frontal gyrus and probabilistic atlas in a standard stereotaxic space," *Hum. Brain Mapp.*, vol. 28, no. 1, pp. 34–48, 2007.

[228] B. Van Ginneken, T. Heimann, and M. Styner, "3D segmentation in the clinic: A grand challenge," in *3D Segmentation in the clinic: A grand challenge*, pp. 7–15, 2007.

[229] B. Landman and S. Warfield, "MICCAI 2012 workshop on multi-atlas labeling," in *MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling*, 2012.

[230] A. Asman, A. Akhondi-Asl, H. Wang, N. Tustison, B. Avants, S. Warfield, and B. Landman, "MICCAI 2013 segmentation algorithms, theory and applications (SATA) challenge results summary," in *MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (SATA)*, 2013.

[231] E. M. van Rikxoort, I. Isgum, M. Staring, S. Klein, and B. van Ginneken, "Adaptive local multi-atlas segmentation: Application to heart segmentation in chest CT scans," *Medical Imaging*, vol. 6914, 2008.

[232] A. R. Khan, L. Wang, and M. F. Beg, "FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping," *NeuroImage*, vol. 41, no. 3, pp. 735–746, 2008.

[233] C. Ledig, R. A. Heckemann, P. Aljabar, R. Wolz, J. V. Hajnal, A. Hammers, and D. Rueckert, "Segmentation of MRI brain scans using MALP-EM," in *MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling*, pp. 79–82, 2012.

[234] J. Doshi, G. Erus, Y. Ou, and C. Davatzikos, "Ensemble-based medical image labeling via sampling morphological appearance manifolds," in *MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (SATA)*, 2013.

[235] L. Debernard, T. R. Melzer, S. Alla, J. Eagle, S. V. Stockum, C. Graham, J. R. Osborne, J. C. Dalrymple-Alford, D. H. Miller, and D. F. Mason, "Deep grey matter MRI abnormalities and cognitive function in relapsing-remitting multiple sclerosis," *Psychiatry Research: Neuroimaging*, vol. 234, no. 3, pp. 352–361, 2015.

[236] E. Mak, N. Bergsland, M. Dwyer, R. Zivadinov, and N. Kandiah, "Subcortical atrophy is associated with cognitive impairment in mild Parkinson disease: A combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis," *Am. J. Neuroradiol.*, vol. 35, no. 12, pp. 2257–2264, 2014.

[237] B. Audoin, W. Zaaraoui, F. Reuter, A. Rico, I. Malikova, S. Confort-Gouny, P. J. Cozzone, J. Pelletier, and J.-P. Ranjeva, "Atrophy mainly affects the limbic system and the deep grey matter at the first stage of multiple sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 81, no. 6, pp. 690–695, 2010.

[238] M. Calabrese, F. Rinaldi, I. Mattisi, V. Bernardi, A. Favaretto, P. Perini, and P. Gallo, "The predictive value of gray matter atrophy in clinically isolated syndromes," *Neurology*, vol. 77, no. 3, pp. 257–263, 2011.

[239] M. M. Schoonheim, V. Popescu, F. C. R. Lopes, O. T. Wiebenga, H. Vrenken, L. Douw, C. H. Polman, J. J. Geurts, and F. Barkhof, "Subcortical atrophy and cognition sex effects in multiple sclerosis," *Neurology*, vol. 79, no. 17, pp. 1754–1761, 2012.

[240] T. Štecková, P. Hluštík, V. Sládková, F. Odstrčil, J. Mareš, and P. Kaňovskỳ, "Thalamic atrophy and cognitive impairment in clinically isolated syndrome and multiple sclerosis," *Journal of the Neurological Sciences*, vol. 342, no. 1, pp. 62–68, 2014.

[241] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, 2010.

[242] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, 2001.

[243] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, 2000.

[244] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Améli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M. M. Santos, W. P. Santos, A. G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F. Vera-Olmos, N. Malpica,

C. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S. K. Warfield, F. Cotton, and C. Barillot, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," vol. 8, no. 1, pp. 1–17, 2018.

[245] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield, "3D segmentation in the clinic: A grand challenge II: MS lesion segmentation," *The MIDAS Journal*, pp. 1–6, 2008.

[246] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. J. Cardoso, N. Cawley, O. Ciccarelli, C. A. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham, "Longitudinal multiple sclerosis lesion segmentation: Resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, 2017.

[247] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[248] R. Gelineau-Morel, V. Tomassini, M. Jenkinson, H. Johansen-Berg, P. M. Matthews, and J. Palace, "The effect of hypointense white matter lesions on automated gray matter segmentation in multiple sclerosis," *Hum. Brain Mapp.*, vol. 33, no. 12, pp. 2802–2814, 2012.

[249] K. Nakamura and E. Fisher, "Segmentation of brain magnetic resonance images for measurement of gray matter atrophy in multiple sclerosis patients," *NeuroImage*, vol. 44, no. 3, pp. 769–776, 2009.

[250] Y. Huo, A. Asman, A. Plassard, and B.A.Landman, "Simultaneous total intracranial volume and posterior fossa volume estimation using multi-atlas label fusion," *Hum. Brain Mapp.*, vol. 38, no. 2, pp. 599–616, 2017.

[251] H. Wang, J. W. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 3, pp. 611–623, 2013.

[252] F. Prados, M. J. Cardoso, B. Kanber, O. Ciccarelli, R. Kapoor, C. A. G. Wheeler-Kingshott, and S. Ourselin, "A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis," *NeuroImage*, vol. 139, pp. 376–384, 2016.

[253] S. Batista, R. Zivadinov, M. Hoogs, N. Bergsland, M. Heininen-Brown, M. G. Dwyer, B. Weinstock-Guttman, and R. H. B. Benedict, "Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis," *Journal of Neurology*, vol. 259, no. 1, pp. 139–146, 2012.

[254] A. Asman and B. Landman, "Formulating spatially varying performance in the statistical fusion framework," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1326–1336, 2012.

[255] B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008.

[256] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache, "Reconstructing a 3D structure from serial histological sections," *Image and Vision Computing*, vol. 19, no. 1, pp. 25–31, 2001.

[257] A. Klein, T. D. Canton, S. Ghosh, B. Landman, J. Lee, and A. Worth, "Open labels: Online feedback for a public resource of manually labeled brain images," in *16th Annual Meeting for the Organization of Hum. Brain Mapp.*

[258] "MICCAI grand challenge on MR Brain Segmentation (MRBrainS18)," 2018.

[259] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, 2010.

[260] "MICCAI white matter hyperintensities segmentation challenge," 2017.

[261] A. J. Asman, Y. Huo, A. J. Plassard, and B. A. Landman, "Multi-atlas learner fusion: An efficient segmentation approach for large-scale data," *Med. Image Anal.*, vol. 26, no. 1, pp. 82–91, 2015.

[262] A. Akhondi-Asl, L. Hoyte, M. Lockhart, and S. Warfield, "A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 1997–2009, 2014.

[263] H. Amiri, A. de Sitter, K. Bendfeldt, M. Battaglini, C. A. G. Wheeler-Kingshott, M. Calabrese, J. J. Geurts, M. A. Rocca, J. Sastre-Garriga, C. Enzinger, N. de Stefano, M. Filippi, Á. Rovira, F. Barkhof, and H. Vrenken, "Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI," *NeuroImage: Clinical*, vol. 19, pp. 466–475, 2018.

[264] G. Battistella, E. Najdenovska, P. Maeder, N. Ghazaleh, A. Daducci, J.-P. Thiran, S. Jacquemont, C. Tuleasca, M. Levivier, M. Bach Cuadra, and E. Fornari, "Robust thalamic nuclei segmentation method based on local diffusion magnetic resonance properties," *Brain Structure and Function*, vol. 222, no. 5, pp. 2203–2216, 2017.

[265] H. Sokooti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 232–239, 2017.

[266] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, 2019.

[267] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration - A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.

[268] M. A. Rocca, S. Mesaros, E. Pagani, M. P. Sormani, G. Comi, and M. Filippi, "Thalamic damage and long-term progression of disability in multiple sclerosis," *Radiology*, vol. 257, no. 2, pp. 463–469, 2010.

[269] M. Calabrese, V. Poretto, A. Favaretto, S. Alessio, V. Bernardi, C. Romualdi, F. Rinaldi, P. Perini, and P. Gallo, "Cortical lesion load associates with progression of disability in multiple sclerosis," *Brain*, vol. 135, no. 10, pp. 2952–2961, 2012.

[270] A. Charil, A. Dagher, J. P. Lerch, A. P. Zijdenbos, K. J. Worsley, and A. C. Evans, "Focal cortical atrophy in multiple sclerosis: Relation to lesion load and disability," *NeuroImage*, vol. 34, no. 2, pp. 509–517, 2007.