Robust Segmentation of Focal Lesions on Multi-Sequence MRI in Multiple Sclerosis

Daniel García-Lorenzo

N° d'ordre : 4025 ANNÉE 2010

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de
DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique
Ecole doctorale MATISSE

présentée par

Daniel García Lorenzo

préparée à l'unité de recherche UMR CNRS 6074 / CRIRBA
Équipe d'accueil : VisAGeS - INSERM U746
Composante universitaire: IFSIC

Robust Segmentation of Focal Lesions on Multi-Sequence MRI in Multiple Sclerosis

Thèse à soutenir à Rennes le 4 mars 2010

devant le jury composé de :

M. Patrick BOUTHEMY
Directeur de recherche à l'INRIA/ président

M. Gregoire MALANDAIN
Directeur de recherche à l'INRIA / rapporteur

M. Jean-Philippe RANJEVA
Professeur des universités à l'université de Marseille / rapporteur

M. Gilles EDAN
Professeur des universités - practicien hospitalier au CHU Pontchaillou / examinateur

M. Ponnada A. NARAYANA
Professeur à University Texas Medical School at Houston / examinateur

M. Christian BARILLOT
Directeur de recherche à l'INRIA / directeur de thèse

I would like to thank the Assotiation pour la Recherche sur la Sclérose en Plaques (ARSEP) and the Université Européenne de Bretagne (UEB) for partially funding my thesis. I would also thank my collaborators, supervisors and the jury members for making possible this work.

I would not forget the support received from my family and the help received and good times passed with my colleges and friends during this period of my life.

Muchas gracias a todos

Contents

Segmentation robuste de lésions focales de sclérose en plaques d'IRM multi séquence

Introduction

L'objectif de cette thèse est la création d'une application qui permettra la segmentation automatique des lésions focales que l'on peut observer sur les images par résonances magnétiques des patients atteints de sclérose en plaques (SEP). Dans cette section, nous réalisons un résumé en français du manuscrit qui a été écrit en anglais. Nous conservons la même structure et les mêmes acronymes que dans le document en anglais.

1. La sclérose en plaques et l'imagerie par résonance magnétique

La sclérose en plaques est une des causes les plus importantes de handicap chez le jeune adulte : elle touche plus de 80.000 personnes en France. L'imagerie par résonance magnétique permet la visualisation du cerveau et est largement utilisée dans le diagnostic et le suivi de patients atteints de SEP.

La sclérose en plaques

La sclérose en plaques est une maladie inflammatoire et démyélinisante du système nerveux central qui a été décrite par Charcot [Charcot 1868].

Du point de vue clinique, la SEP peut être décrite comme une maladie chronique qui comporte une grande variété de symptômes différents. En fonction de l'évolution de la maladie, on parle de quatre formes cliniques de la maladie (Figure 1):

Récurrente-rémittente (RRMS) : 55% des patients ont ce type d'évolution qui est caractérisé par des poussées suivies de rémissions de symptômes après une courte période en laissant peu ou pas de handicap.

Secondaire progressive (SPMS) : 30% des patients sont SPMS ce qui se caractérise par une évolution progressive de la maladie qui peut être accompagnée de poussées.

Primaire-progressive (PPMS) : l'évolution est progressive et sans poussée et elle concerne 10% des patients de SEP.

Progressive-rémittente(RRMS) : se caractérise par une évolution progressive de la maladie qui peut être accompagnée de poussées et elle concerne 5% des patients de SEP.

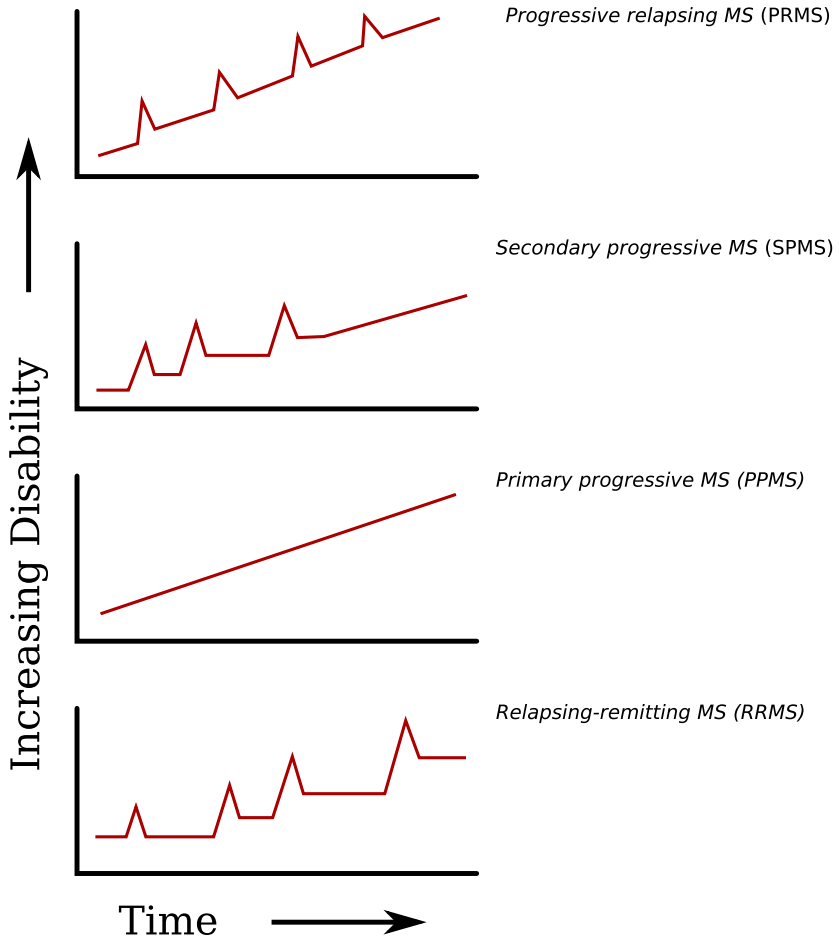


Figure 1: Évolution des quatre types cliniques de SEP. (Source: http://commons.wikimedia.org/)

2. L'imagerie par résonance magnétique

L'imagerie par résonance magnétique (IRM) est un type d'imagerie médicale basé sur les propriétés magnétiques de quelques noyaux atomiques comme l'hydrogène, qui sont présents dans les tissus du corps humain.

Lésions focales

L'IRM montre des anormalités dans la matière blanche chez la majorité des patients de SEP [Paty and Ebers 1997]. Ces anormalités ont été associées à des lésions dans

des études qui combinent des analyses histopatologiques avec l'IRM [Groot et al. 2001].

Il y a plusieurs types de lésions focales et pour les classifier nous avons besoin de plusieurs séquences IRM. Selon leur intensité, nous pouvons trouver trois types de lésions (Figure 2):

Lésions T2-w : Ces lésions sont hyper-intenses comparées à l'intensité de la matière blanche en T2-w, PD et FLAIR. Elles ne sont pas spécifiques d'un processus particulier et sont produites par inflammation, oedème, demyelinitiation ou perte axonale.

Lésions avec prise de contraste en Gadolinium : Une augmentation dans le contraste peut être observée dans les images avant et après l'injection du Gadolinium dans T1-w. Ces lésions sont moins nombreuses que les lésions T2-w et sont associées au processus inflammatoire.

"Trous noirs" : Ces lésions sont hypo-intenses par rapport à la matière blanche dans les images T1-w. Il y a quelques hypo-intensités qui disparaissent après un mois ou deux, donc l'hypo-intensité doit être présente pendant au moins trois mois pour être considérée comme un trou noir. Les trous noirs sont souvent associés à la perte axonale.

3. La segmentation des lésions de SEP

Nous pouvons classifier les méthodes de segmentation de lésion de SEP selon trois facteurs: le degré d'intervention humaine, le nombre de dimensions et la stratégie de fusion des informations des différentes séquences.

Selon le degré d'intervention humaine, nous pouvons différencier:

Interactives : Ces méthodes nécessite l'intervention d'un utilisateur humain pour procéder à la segmentation. Nous pouvons différencier la segmentation semi-automatique, où la méthode aide l'utilisateur dans le processus de segmentation, et la segmentation complètement manuelle, qui est normalement utilisée comme méthode de référence.

Automatiques : Les méthodes automatiques produisent une segmentation sans interaction humaine. Nous pouvons classifier les méthodes entre supervisées et les méthodes non-supervisées. Les premières ont besoin des données segmentées pour "apprendre" comment segmenter les lésions tandis que les méthodes non-supervisées extraient tout l'information exclusivement de l'image.

Souvent, plusieurs images du même patient sont prises au cours du temps pour suivre l'évolution de la maladie. Nous classifions les méthodes selon le nombre de dimensions utilisées :

2D : Chaque coupe de l'image est segmentée indépendamment [Grimaud et al. 1996].

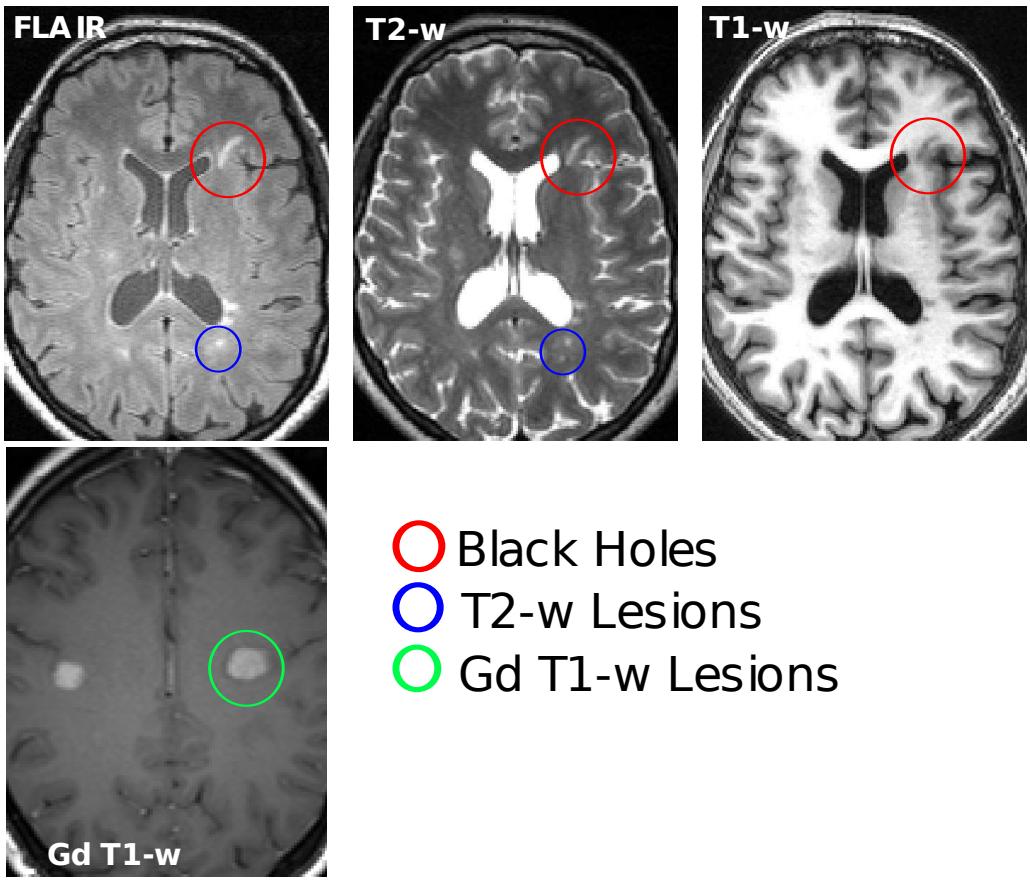


Figure 2: Exemples des lésions de SEP en IRM: T1-w, T2-w, FLAIR and T1-w avec contraste en Gadolinium.

3D : Toute l'image est traitée comme une image 3D pour utiliser l'information anatomique et spatiale [Van Leemput et al. 2001].

3D+t : Quelques méthodes utilisent des informations temporales pour créer un modèle temporal des lésions [Shahar and Greenspan 2004] ou pour utiliser l'information longitudinale redondante [Aït-Ali et al. 2005].

Finalement, les méthodes utilisent différentes stratégies pour combiner l'information des séquences:

Mono-séquence : Quelque méthodes utilisent exclusivement une séquence pour la segmentation des lésions, notamment FLAIR [Khayati et al. 2008, Anbeek et al. 2008].

Multi-śequence-Parallel : Chaque séquence peut être segmentée individuellement pour ensuite fusionner les résultats pour améliorer la segmentation [Johnston et al. 1996, Ardizzone et al. 2002].

Multi-séquence-Joint : Cees méthodes utilisent toutes les séquences en même temps afin d'améliorer la segmentation [Van Leemput et al. 2001, Zijdenbos et al. 1994].

Multi-séquence-Hierarchical : Dans ce cas, chaque séquence est utilisé pour une tâche spécifique, et le résultat de'une tâche est utilisée pour simplifier la tâche suivante [Sajja et al. 2006, Dugas-Phocion et al. 2004a].

3. La chaîne de traitement pour la segmentation des lésions de SEP

La segmentation est la dernière étape d'une chaîne de traitements qui comprend : débruitage, correction des inhomogénéités, recalage et extraction du cerveau. Tout ces traitements sont des prérequis nécessaires avant l'utilisation de la méthode de segmentation.

Débruitage

Les images sont corrompues par le bruit lié au processus d'acquisition des images. De nombreuses méthodes de correction du bruit ont été proposées pour les résonances magnétiques. Dans notre cas, nous utilisons la méthode de moyennes non-locales (NLM) [Coupé et al. 2008] qui a montré de meilleurs résultats que des approches classiques comme la diffusion anisotropique et la variation totale.

Correction des inhomogénéités d'intensité

La non-uniformité du champ magnétique dans un scanner de résonance magnétique est une des causes de l'inhomogénéité de l'intensité dans l'image. A cause de cet effet, l'intensité d'un même tissu dans l'image varie selon sa position dans l'image. Cet effet n'affecte pas la lecture effectuée par un radiologue mais il réduit la performance des méthodes de segmentation et de recalage qui nécessitent que l'intensité des tissus soit homogène. Nous avons réalisé une comparaison des différentes méthodes de correction d'inhomogénéité d'intensité pour des images de patients atteints de SEP. Ce travail nous a permis d'étudier la réduction obtenue avec chacune des méthodes et son comportement autour des lésions. Nous avons retenu la méthode du logiciel BrainVISA [Mangin 2000].

Recalage

Pendant l'acquisition de plusieurs séquences le patient peut bouger, ce qui provoque que les images ne soient pas toujours spatialement alignées. Le processus d'alignement des images s'appelle recalage et est nécessaire avant l'utilisation des images dans la segmentation. Nous utilisons une méthode de recalage qui se sert de l'information mutuelle entre les deux images qui a été largement employée dans le traitement d'images médicales. Nous utilisons, en particulier, une nouvelle méthode d'optimisation, plus rapide que l'originale [Wiest-Daesslé et al. 2007].

Extraction du cerveau

Pour simplifier la segmentation des lésions de SEP, dans la plupart des méthodes, il faut extraire le cerveau et éliminer les restes de voxels. Cette étape est très compliquée parce que le cerveau est très hétérogène et les structures voisines peuvent être confondues avec lui. Une comparaison des méthodes a été réalisée récemment [Fennema-Notestine et al. 2006]. Nous utilisons une méthode qui se base sur un modèle déformable qui est adapté au bord du cerveau en utilisant des forces locales [Smith 2002]; et qui a été utilisé dans la segmentation du cerveau de patients de SEP [Akselrod-Ballin et al. 2006, Lao et al. 2008].

Chaîne de traitement proposée

L'objectif de cette chaîne de traitement est de traiter les images originales provenant du scanner et d'obtenir une image de sortie où chaque voxel est identifié comme matière blanche (WM), matière grise (GM), liquide céphalo-rachidien (CSF) ou autre. Nous proposons la chaîne de traitement décrite dans la Figure 3. Les images sont débruitées (Denoise) et ensuite l'inhomogénéité d'intensité est corrigée (IIH correction) sauf pour l'image T2-w. En utilisant FLAIR comme référence, les images sont recalées rigidement (Registration) et le cerveau est extrait grâce à l'image T1-w (Skull Stripping). Nous allons par la suite décrire trois méthodes de segmentation qui pourraient être utilisées comme dernière partie de cette chaîne.

4. STREM: Estimation robuste pour la segmentation des lésions de SEP

Dans la dernière section, nous avons proposé une chaîne de traitement pour la segmentation. Dans cette section, nous présentons STREM [Aït-Ali 2006], une approche paramétrique qui servira de base pour les prochaines méthodes. En plus, nous décrirons quelques modifications pour améliorer la méthode.

Méthode

La méthode classifique chaque voxel du cerveau comme une de quatre classes suivantes: WM, GM, CSF, or MS lésions. Nous utilisons un protocole IRM classique (T1-w, T2-w and PD) comme entrée de la méthode mais d'autres séquences, comme FLAIR, peuvent être intégrées avec quelques modifications. La Figure 4 montre la chaîne de traitement proposée pour la segmentation de lésions de SEP. Les images sont débruitées, les inhomogénéités sont corrigées et toutes les images sont recalées dans le même espace. Notre algorithme de segmentation est composé de trois étapes: estimation du modèle NABT, détection de données aberrantes et utilisation des règles pour sélectionner les lésions parmi toutes les données aberrantes.

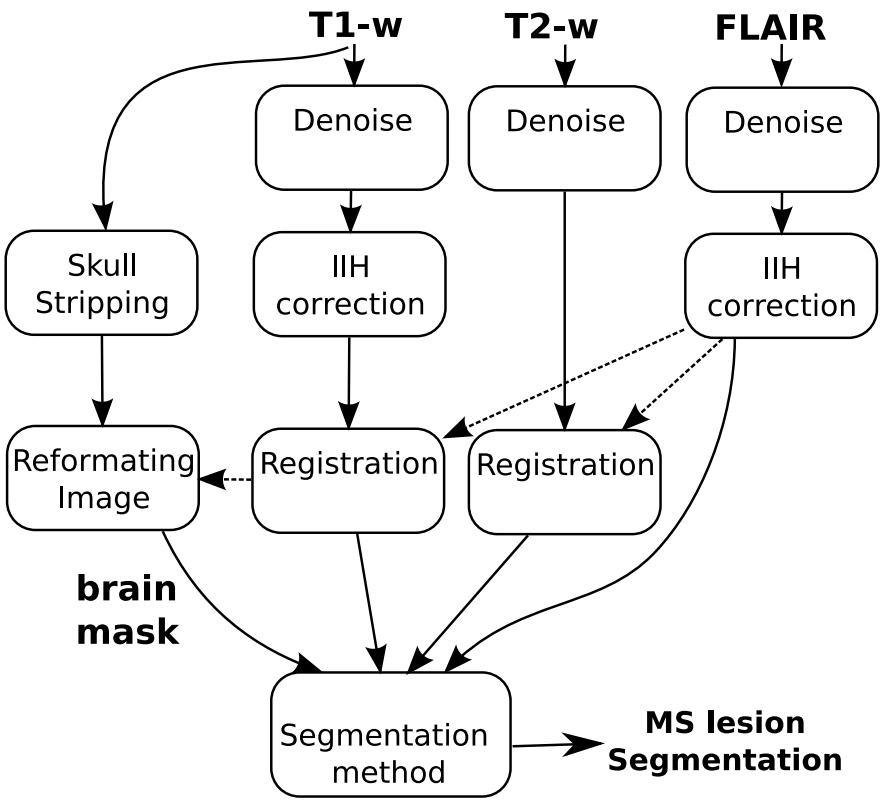


Figure 3: Chaîne de traitement proposée pour la segmentation de lésions de SEP.

Estimation du modèle NABT

Le bruit dans les IRM suit normalement une loi ricienne [Dietrich et al. 2008] qui est souvent approximé pour une loi gaussienne quand le rapport signal-bruit est élevé. La distribution des intensités de chaque tissu du cerveau peut être approximé par une loi gaussienne. Pour le cerveau, nous utilisons un modèle avec trois Gaussiens (GMM[1]) où chaque Gaussien correspond à un des tissus d'apparence normale (NABT) du cerveau : WM, GM et CSF.

Le vecteur des intensités $\mathbf{y_i} = [y_{i_1}...y_{i_m}]$ du voxel i peut être modelé comme

$$f(\mathbf{y_i}|\theta) = \sum_{j=1}^{3} \alpha_j \cdot N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{1}$$

où la moyenne $\boldsymbol{\mu}_j$ et la matrice de covariance $\boldsymbol{\Sigma}_j$ font partie de la Gaussien $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Tous les paramètres font partie de vecteur de paramètres θ.

Ce modèle est souvent calculé grâce à l'estimateur de maximum de vraisemblance (MLE):

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} f(\mathbf{y_i}|\theta) \tag{2}$$

[1]En anglais Gaussian Mixture Model

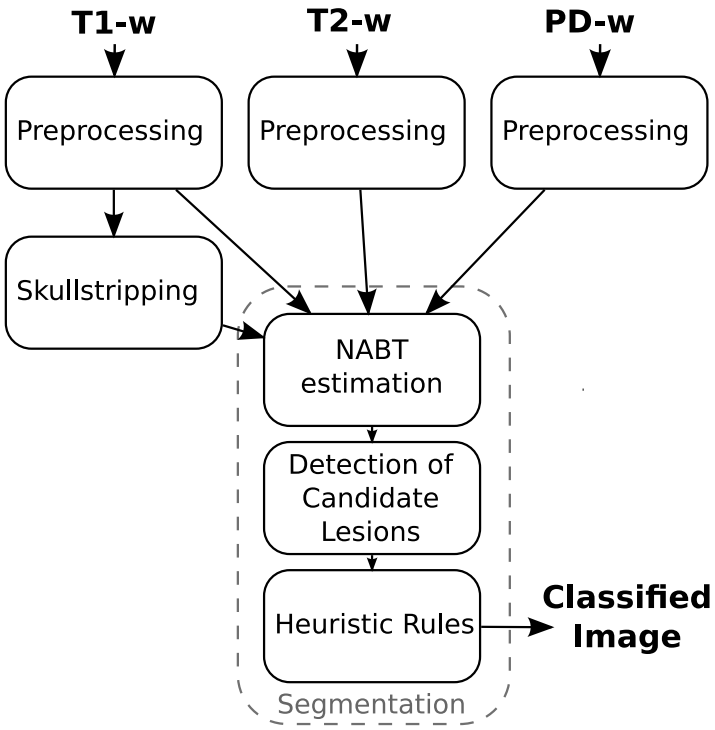


Figure 4: Schèma de la méthode STREM.

si on considère $\mathbf{y_i}$ comme étant des variables identiques et indépendantes.

Pour obtenir le MLE, l'algorithme d'expectation-maximisation (EM) est souvent utilisé parce qu'il est très simple à implémenter et qu'il converge toujours vers un maximum local de vraisemblance. Mais il a deux limites. La première limite vient du fait que l'EM garantit seulement l'obtention d'un maximum local; différentes initialisations de l'algorithme peuvent donner différentes solutions, c'est pouquoi une initialisation correcte de l'EM est importante. La deuxième limite est la sensibilité du MLE aux données aberrantes. Nous proposons deux solutions pour minimiser ces limites : utiliser une approche hiérarchique pour l'initialisation de l'EM et remplacer le MLE par l'estimateur de vraisemblance tamisée (TLE).

Initialisation hiérarchique

Notre méthode est basée sur l'utilisation des initialisations aléatoires pour minimiser la possibilité de converger vers un maximum local. En plus, nous profitons du bon contraste existant dans la séquence T1-w pour réduire le temps de calcul. Il y a deux avantages par rapport à l'initialisation par atlas [Aït-Ali 2006]: le recalage de l'atlas est long et ce type d'initialisation peut ne pas être optimale chez des patients qui ont une grand atrophie ou une large charge lésionelle.

Estimateur de vraisemblance tamisé

Neykov et al. [Neykov et al. 2007] a proposé l'estimateur de vraisemblance tamisée qui est une modification du MLE pour le rendre plus robuste aux données aberrantes. L'idée est de maximiser la vraisemblance tamisée

$$TL(\theta) = \prod_{i=1}^{k} f(\mathbf{y}_{\nu(\mathbf{i})}|\theta) \tag{3}$$

où le paramètre k détermine combien de voxels sont rejetés de l'estimation et $\nu(i)$ ordonne les voxels en fontion de sa probabilité

$$f(\mathbf{y}_{\nu(\mathbf{1})}|\theta) \geq f(\mathbf{y}_{\nu(\mathbf{2})}|\theta) \geq ... \geq f(\mathbf{y}_{\nu(\mathbf{n})}|\theta). \tag{4}$$

Il est prouvé que cet estimateur converge vers un maximum local de TL et qu'il peut être robuste même en présence de $n - k$ données aberrantes.

Détection des possibles lésions

Une valeur très élevée du paramètre k doit être choisie pour être sur que l'estimation du modèle NABT ne soit pas affectée par les lésions et autres artefacts. Pour affiner la sélection des données aberrantes nous utilisons la distance de Mahalanobis entre le voxel i et chacune des Gaussiens. Si la distance de Mahalanobis est supérieure à un seuil donné pour la loi χ^2_m, le voxel est considéré comme étant une lésion candidate.

Règles

Les voxels détectés avec la distance de Mahalanobis incluent les lésions, les vaisseaux, et d'autres artefacts de l'image. Nous utilisons trois règles pour rejeter les voxels qui ne sont pas des lésions

Intensité : Les lésions de SEP sont hyper-intenses comparées avec la WM en T2-w, PD et FLAIR. Un voxel est considéré comme lésion seulement si l'intensité dans les trois séquences est hyper-intense.

Taille : Pour éviter de fausses détections, une lésion doit être d'une taille minimale de 3 voxels.

Voisinage : Il y a de nombreux artefacts dans le CSF externe. Pour éviter de fausses détections, nous éliminons toutes les détections qui ne sont pas contigues à la WM.

5. MS4MS: Mean shift Appliqué à la SEP

Dans la dernière section, nous avons présenté STREM, une méthode pour la segmentation de lésions qui est basée exclusivement sur l'intensité des voxels. Dans ce chapitre, nous proposons une nouvelle méthode pour la segmentation où nous incluons l'information spatiale en plus de l'information d'intensité.

Notre hypothèse est que la segmentation d'un voxel isolé est plus compliquée que la segmentation d'une région de l'image. Une région est une groupe de voxels dans une même partie de l'image avec des caractéristiques d'intensité similaires. Premièrement, nous divisons le cerveau en régions grâce à l'algorithme mean shift et deuxièmement, nous classifions les lésions en utilisant la méthode STREM proposée dans la dernière section.

Mean Shift

L'algorithme mean shift est une technique non paramétrique qui sert à estimer le gradient de la densité de probabilité proposé [Fukunaga and Hostetler 1975]. Comaniciu et Meer [Comaniciu and Meer 2002] ont appliqué le mean shift pour le traitement d'images comme la segmentation ou le suivi vidéo. Dans la segmentation, cet algorithme crée un nombre indéterminé de régions en utilisant seulement l'information locale.

Récemment, le mean shift a été employé pour la segmentation d'IRM de sujets sains [Jimenez-Alaniz et al. 2006, Mayer and Greenspan 2009]. Les premiers ont utilisé le mean shift pour créer des régions qu'ils ont ensuite classifiées à l'aide d'un atlas en WM, GM et CSF. Les deuxièmes ont utilisé une version adaptative du mean shift [Georgescu et al. 2003] pour créer des régions et ils les ont classifiées en utilisant l'algorithme de k-moyennes.

Méthode

La Figure 5 montre un schéma de notre méthode. Nous utilisons le modèle NABT qui sert à normaliser toutes les séquences et à détecter les lésions candidates. Ensuite, nous utilisons le mean shift pour diviser le cerveau en régions. Finalement, chaque région est identifiée et les lésions sont déterminées en utilisant les règles heuristiques.

Estimation du modèle NABT

Nous réalisons l'estimation du modèle NABT de la même façon que pour STREM. Nous considérons un modèle avec trois Gaussiens qui représentent les trois tissus du cerveau GM, WM et CSF. Nous utilisons le TLE pour estimer les paramètres du modèle afin d'éviter que les lésions affectent l'estimation.

Normalisation des séquences

Toutes les images sont normalisées avant d'être segmentées avec le mean shift. Nous normalisons les images pour que la variance de la matière blanche dans toutes les séquences soit égale à 10.000. Cela permet de fixer les paramètres du mean shift pour toutes les images.

Segmentation par mean shift

Comme expliqué précédemment, le mean shift crée un nombre indéterminé de régions. Notre hypothèse est que chaque région appartient seulement à un tissu et

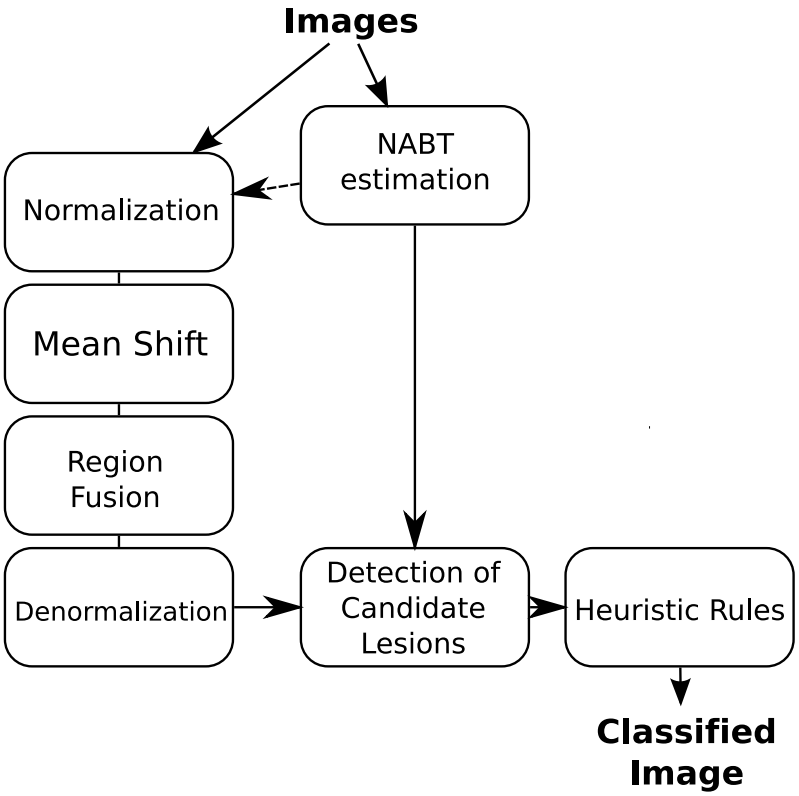


Figure 5: Schèma de la méthode MS4MS.

que l'identification d'une région est plus simple que l'identification de chaque voxel indépendamment.

Fusion de régions

Pour réduire le nombre de régions données par le mean shift, nous utilisons une méthode pour fusionner les régions. Dans les parties de l'image qui sont homogènes, le mean shift peut créer trop de régions. Pour réduire le nombre de régions, nous fusionnons deux régions si la distance entre les deux est inférieure à un seuil donné [Soille 2008].

Dénormalisation

Après la segmentation donnée par le mean shift, nous dénormalisons les images pour récupérer les intensités originales pour chaque région.

Détection de lésions candidates

De la même façon que pour STREM, nous utilisons la distance de Mahalanobis pour détecter les régions qui sont des lésions candidates.

Règles

Pour limiter le nombre de fausses détections, nous appliquons les mêmes règles que pour STREM.

6. GCEM: Segmentation de lésions de SEP par coupe de graphes

Dans la plupart des situations, les médecins voudraient corriger les segmentations fournies par les méthodes automatiques de segmentation. Les méthodes de segmentation par coupes de graphes[2] (GC) [Boykov and Funka-Lea 2006] sont des méthodes semi-automatiques qui ont été employées dans des domaines médicaux différents avec de bonns résultats.

Dans ce chapitre, nous étudions la possibilité d'automatiser les méthodes par coupe de graphes. L'objectif est de proposer une méthode automatique que soit rapide et qui donne la possibilité aux experts de corriger facilement une segmentation. Ce travail a été réalisé en collaboration avec J. Lecoeur.

La coupe de graphes

Les méthodes par coupe de graphes minimisent une énergie qui se compose de deux termes: un terme régional (regional term) et un terme de contour (boundary term). Le terme de contour mesure la similarité des voxels voisins et le terme régional mesure la similarité des voxels avec les caractéristiques d'intensité de l'objet à segmenter.

Méthode

Notre approche est décrite dans la Figure 6. Nous utilisons le gradient spectrale comme terme de contour et nous utilisons une méthode paramétrique pour déterminer les terme régional.

Terme de contour

La coupe de graphes souvent a été utilisée dans des méthodes de segmentation mono-séquence mais nous l'appliquons la segmentation multi-séquence de lésions de SEP. Nous utilisons le gradient spectral proposé pour la segmentation semi-automatique des cerveaux et des lésions de SEP [Lecoeur 2010].

Le gradient spectral considère que les les trois séquences IRM forment une pseudo-image couleur RVB, ce qui permet d'utiliser un détecteur de contour colorimétrique [Lecoeur 2010].

Terme régional

Nous utilisons une approche similaire à celle de STREM. Dans un premier temps, nous estimons les modèles d'intensité du cerveau en utilisant un modèle avec trois

[2]En anglais graph cut

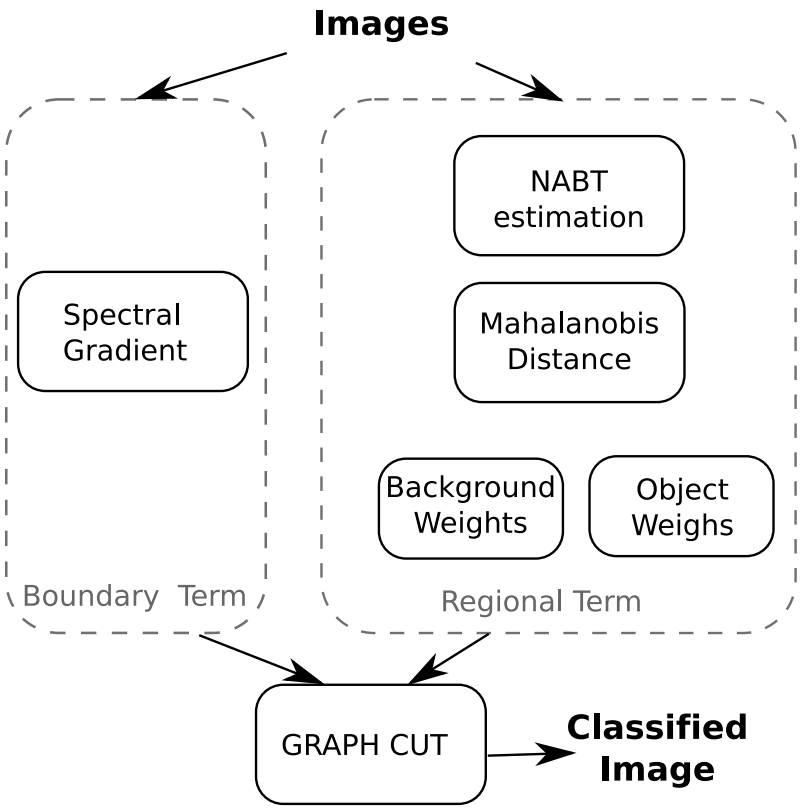


Figure 6: Schèma de la méthode GCEM.

classes Gaussiens. Ensuite, nous utilisons la distance de Mahalanobis pour définir les zones de l'images qui ont des intensités eloignées de notre modèle de tissus d'apparence normale. La grande différence par rapport à STREM, c'est la substitution des seuils pour des règles floues pour incrémenter l'information fournie à la coupe de graphes.

7. Validation

Dans cette section nous évaluons la précision de notre segmentation et nous comparons les méthodes proposées avec les segmentations manuelles réalisées par des experts. Ici nous montrons une partie de la validation réalisé dans le manuscript.

Données de MNI

Les images des dix patients ont été prises à l'Institut Néurologique de Montréal (MNI). Le protocole d'acquisition consiste en des images T1-w, T2-w et PD avec des coupes de 3 mm et une résolution de 1 mm.

Les images ont été segmentées manuellement deux fois par 5 experts, au total il y a dix segmentations par image. Pour obtenir une vérité terrain afin de comparer nos algorithmes, nous avons utilisé une méthode de consensus. Dans notre vérité terrain, un voxel est considéré comme lésion seulement si au moins 6 des 10 segmentations

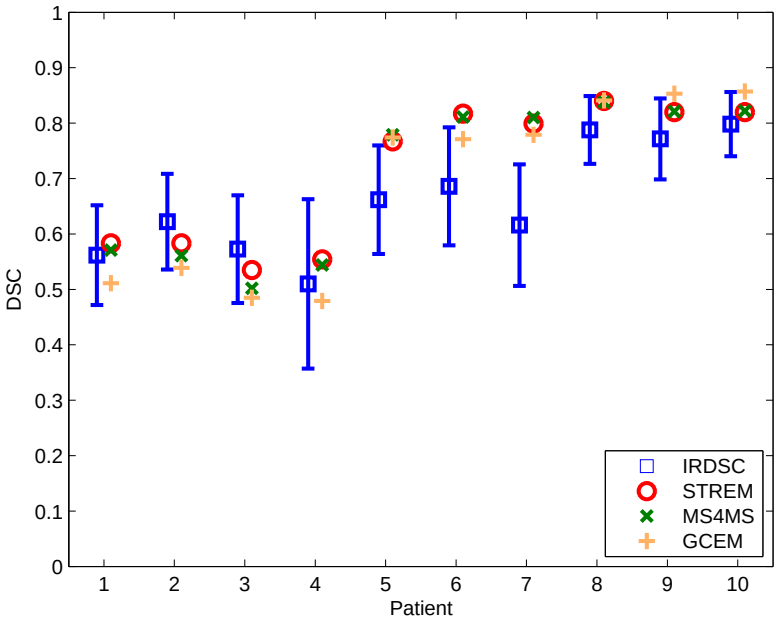


Figure 7: Résultat de la segmentation avec les données du MNI, en utilisant notre approche en deux étapes pour adapter notre méthode à la charge lésionelle. Nous utilisons le coefficient de Dice (DSC) dans l'évaluation.

considèrent que le voxel est une lésion. Cette méthode nous permet de réduire les problèmes de variabilité existant entre les experts.

Résultats

La Figure 7 montre les résultats de nos algorithmes en comparaison avec la variabilité des experts. Nous observons que nos méthodes obtiennent des résultats similaires à celles des experts (IRDSC). Dans la Figure 8, nous observons une coupe de la segmentation et nous pouvons voir la similarité entre nos méthodes et la vérité terrain.

Conclusion

Cette thèse porte sur les méthodes automatiques de segmentation des lésions de SEP dans des images par résonance magnétique. Cette segmentation joue un rôle très important dans les études cliniques où la mesure précise du volume des lésion est utilisée comme biomarqueur. Dans ce document, nous avons présenté une brève introduction de la SEP et l'utilisation de l'IRM dans l'étude de la maladie. Nous avons remarqué l'importance de l'analyse automatique des images pour réduire la variabilité et nous avons proposé un classification des méthodes de segmentation des lésions de SEP.

Les travaux de thèse de Laure Aït-Ali [Aït-Ali 2006], qui a introduit l'estimateur de vraisemblance tamisé (TLE) pour la segmentation du cerveau et des lésions de

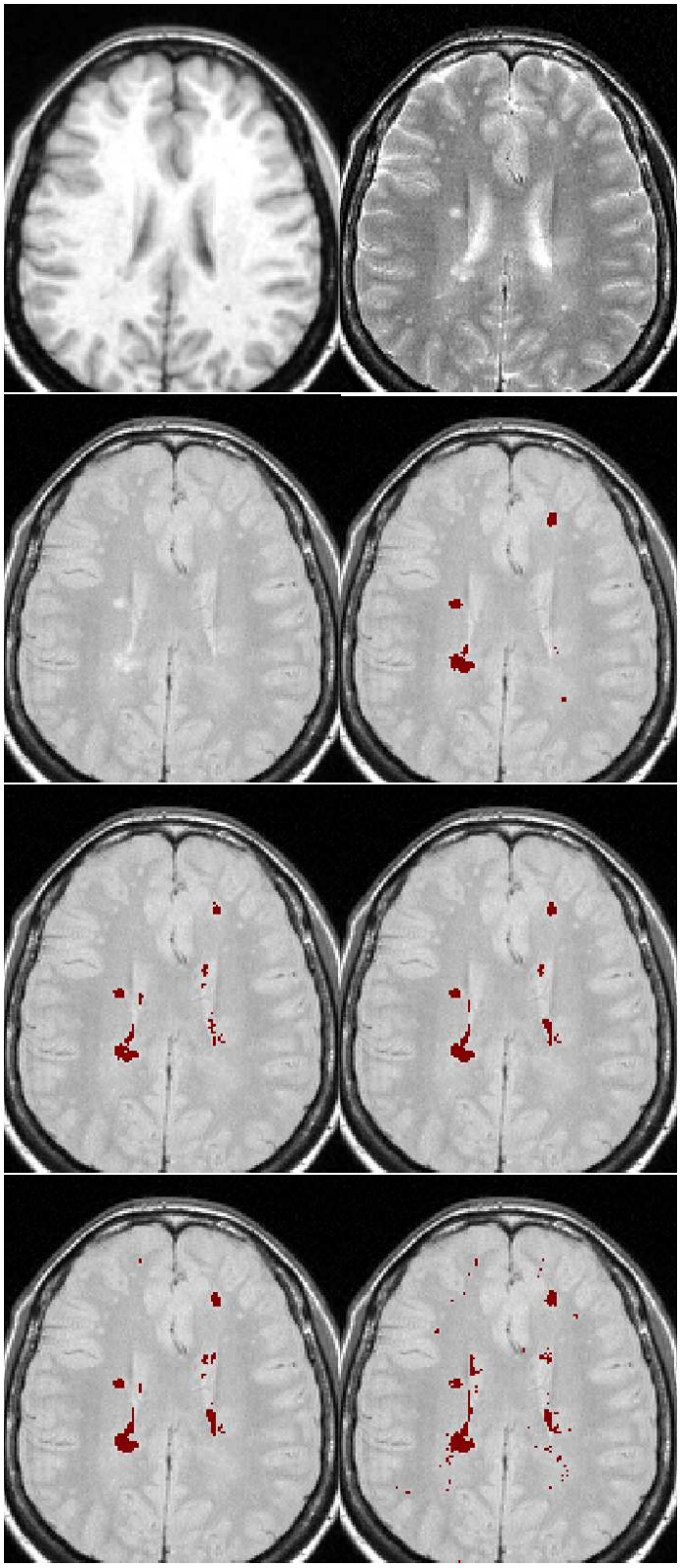


Figure 8: Résultats de la segmentation pour le patient 4. De haut en bas et de gauche a droite: T1-w et T2-w; PD et la vérité terrain; STREM et MS4MS; GCEM et EMS Van Leemput et al. [2001].

SEP, ont été notre point de départ. Dans sa méthode STREM, chaque voxel est classifié en fonction de son intensité sans utiliser d'information spatiale.

Nous avons considéré que la classification d'une région de l'image est plus facile que la classification d'un seul voxel. Nous avons donc proposé MS4MS, une méthode qui, dans un premièr temps, crée plusieurs régions dans l'image à en utilisant l'algorithme mean shift, et les segmente ensuite grâce au TLE.

Les segmentations automatiques sont très souvent vérifiées par un expert qui valide la segmentation ou la modifie en cas d'erreur. Les méthodes de coupe de graphes permettent la segmentation semi-automatique d'une façon interactive. En collaboration avec Jeremy Lecoeur, qui a travaillé sur la segmentation semi-automatique avec des méthodes de coupe de graphes [Lecoeur 2010], nous avons proposé une méthode automatique pour la segmentation de lésions de SEP. L'avantage de cette nouvelle méthode c'est que le résultat obtenu par la méthode automatique peut être modifié par un expert.

Les trois méthodes ont été validées en utilisant des images synthétiques et cliniques. Les images synthétiques ont permis d'étudier le comportement des méthodes avec des niveaux de bruit et inhomogénéité différents. En plus, les images cliniques nous ont permis d'évaluer nos méthodes dans des conditions réelles.

Contributions

Nous pouvons différentier deux contributions principales dans ce travail. Nous avons proposé deux nouvelles méthodes pour la segmentation automatique des lésions de SEP. Premièrement, nous avons étudié le comportement de STREM et proposé plusieurs améliorations. Deuxièmement, nous avons proposé deux approches différentes pour inclure de l'information spatiale qui se sont montrées efficaces contre le bruit et l'inhomogénéité. En plus, nous avons proposé une cadr de validation pour les méthodes de segmentation de lésions de SEP composé de trois étapes: étude des paramètres de la méthode, évaluation avec des données synthétiques et validation avec des images cliniques. Cette approche a été employée avec les trois méthodes ce qui a permis de comparer le méthodes à plusieurs niveaux.

Limites

Définition des lésions

La validation avec les images clinique a montré des résultats contradictoires. * Nos méthodes automatiques ont obtenu des résultats similaires à la segmentation manuelle dans une base de données cliniques mais les résultats dans l'autre base de données cliniques étaient très différents. Les deux bases de données proviennent de deux centres différents, elles ont été acquises avec des scanners différents et, différence plus importante, la définition des lésions pour la segmentation manuelle n'était pas la même. La variabilité obtenue pendant le workshop de segmentation à MICCAI'08 est un exemple de la grande variabilité existant dans la définition des lésions entre les différentes centres spécialisés en SEP. Un groupe international de travail devrait être créé pour établir une définition commune des lésions et réduire cette variabilité.

Informations globales et locales

Les deux méthodes proposées combinent des informations globales et locales. Nous avons démontré comment l'inclusion de cette information locale améliore les résultats des méthodes en présence de bruit et inhomogénéité d'intensité. Les lésions sont hétérogènes et peuvent paraître plus ou moins hyper-intense sur les images. L'utilisation d'une définition globale de hyper-intensité peut empêcher la détection de certaines lésions. L'utilisation des définitions locales d'hyper-intensité devrait réduire ce phènomène.

Chaîne de traitement

Pour utiliser une méthode de segmentation, elle doit être incluse dans une chaîne de traitement. Nous avons proposé une chaîne de traitement mais le choix de cette chaîne de traitement été réalisé à partir d'expériments réalisées seulement avectrois patients. La validation de la chaîne doit évaluer chacun des traitements indépendamment ainsi que sa position dans la chaîne dans une processus de validation plus complexe.

Perspectives

Perspectives méthodologiques

Nos méthodes détectent fauusses lésions sur les images FLAIR due aux nombreux artefacts présent dans ce type d'images. Il est nécessaire de créer d'autres règles pour éviter ces effets. Notre modèle contient seulement trois gaussiennes pour modeler les intensités du cerveau. L'utilisation de modèles paramétriques plus complexes avec l'inclusion de volumes partiels ou une estimation locale des paramètres pourrait améliorer l'estimation du modèle NABT.

Perspectives liées à la validation

L'utilisation de plusieurs segmentations manuelles proportionne plus d'information lors du processus de validation. Une priorité est d'augmenter le nombre de segmentations manuelles sur les images de Rennes pour réaliser une validation plus complète et comprendre les résultats obtenus. Nous considérons que la validation doit aussi inclure la comparaison avec d'autres méthodes de segmentation. Nous voulons segmenter les images du workshop MICCAI avec les deux nouvelles méthodes pour faciliter la comparaison de nos méthodes avec d'autres centres de recherche, malgré les limites de ce workshop.

La dernière étape de validation consisterait à utiliser nos méthodes avec des images venant des grandes études cliniques pour tester la performance dans des cas réels et étudier la corrélation des lésions avec l'évolution de la maladie. Un autre aspect important à étudier est la reproductibilité. Ce type d'analyse a besoin de protocoles d'acquisition particuliers où nous acquérons plusieurs images du même patient sur une courte période de temps pour que la maladie n'ait pas le temps d'évoluer.

Perspectives liées à l'application

L'hôpital universitaire de Pontchaillou (Rennes, France) fait partie d'un projet multicentrique qui vise à étudier l'utilisation d'un nouveau produit de contraste, l'USPIO. Notre objectif est de mettre au point un logiciel automatique et facile à utiliser pour la segmentation de lésions de SEP. De nouveaux développements seront nécessaires pour intégrer de nouvelles images obtenues avec l'USPIO.

Un autre aspect important est la distribution des méthodes de segmentation. Notre méthode de segmentation est en train d'être incluse dans NeuroLOG. NeuroLOG est un projet ANR qui vise a développer un réseau distribué avec pour but de partager les données cliniques et les chaînes de traitement pour ces données entre plusieurs sites français, faisant transparent pour les utilisateurs l'hétérogénéité existante entre les différents sites. En plus, nous sommes en train d'intégrer notre chaîne de traitement sur notre propre site web. De la même façon que d'autres méthodes sont dispobiles en ligne, nous allons proposer un service web qui permettra d'utiliser nos méthodes de segmentation en ligne[3] .

[3] http://www.irisa.fr/visages/benchmarks/

Introduction

Since ancient times, humans have tried to heal each other using different treatments varying from medical herbs to magic rituals which were often linked with traditions, superstition and religion.

Modern medicine applies the scientific method in the discovery of treatments. Evaluations of the safety and efficacy are necessary before accepting a new treatment. These evaluations are performed in form of a clinical trial.

Clinical trials study the effects of a treatment on a cohort of patients and healthy subjects so as to evaluate its efficacy, limitations and secondary effects. The outcome of these trials often includes clinical examinations, laboratory analyses and medical images. The statistical power of clinical trials depends, amongst other things, on the number of patients studied and the variability of the outcome measures.

Multiple sclerosis (MS) is the most common disabling neurological disease among young adults. Medical imaging, especially magnetic resonance imaging (MRI), is employed in clinical trials as a secondary outcome, but the interpretation of these images varies depending on the expert who performs the interpretation. This variability in the interpretation of images requires a large cohort of patients in order to obtain statistically significant results.

In this document, we focus on the focal lesions of the brain that are present in the majority of the images of MS patients. The lesions are associated with the demyelinating process that occurs in MS and their number and volume are used as measure in clinical trials. The manual delineation (or segmentation) of MS lesions is a time-consuming task which shows high inter- and intra-expert variability.

The objective of this work is to obtain an automatic, accurate and reliable segmentation of focal MS lesions in MRI of patients with MS. The objective is to simplify the processing of large cohorts of patients and eliminate the variability in experts' interpretations, which should increase the statistical power of clinical trials.

This document is organized as follows.

Chapter 1: Multiple Sclerosis and Magnetic Resonance Imaging

In this chapter, we propose an overview of multiple sclerosis and the applications of MRI used in the study of the disease.

Chapter 2: Segmentation of MS Lesions on Conventional MRI

In the second chapter, we propose a classification of MS lesion segmentation methods and describe the validation of these methods.

Chapter 3: MS Lesion Segmentation Workflow

In the third chapter, we introduce the concepts of segmentation method and segmentation workflow and propose a workflow for the segmentation of MS lesions. Part of this work has been published in a international abstract conference [García-Lorenzo et al. 2008b] and an international workshop [García-Lorenzo et al. 2008d].

Chapter 4: STREM: Spatio-Temporal Robust Expectation Maximization

In this chapter, we introduce STREM, the segmentation method developed by Laure Aït-Ali [Aït-Ali 2006], propose several improvements, and study the influence of the parameters on the behavior of this method. The improved version of STREM was entered into the segmentation challenge held during MICCAI 2008 [García-Lorenzo et al. 2008c].

Chapter 5: MS4MS: Mean Shift Clustering for MS Lesion Segmentation

In the fifth chapter, we describe a new segmentation method where global and local information is used to obtain a more robust segmentation. This method was presented at an international workshop [García-Lorenzo et al. 2008a].

Chapter 6: GCEM: A Graph Cut Approach With an Expectation-Maximization Initialization

In this chapter, another method for the segmentation of MS lesions is presented. The advantage of this method is that it offers the possibility of refining the result semi-automatically using the graph cut approach. This method was presented in an international conference [García-Lorenzo et al. 2009].

Chapter 7: Validation

In the last chapter, we validate the three methods using one synthetic database and two databases with clinical images. We also compare the three methods using a freely available state-of-the-art algorithm.

Chapter 1

Multiple Sclerosis and Magnetic Resonance Imaging

Multiple Sclerosis (MS) is one of the main causes of handicap in the young adult and affects around 80,000 people in France [ARSEP 2010]. Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that offers the possibility of visualizing the brain. It is a sensitive marker for MS and is extensively used in diagnosis, follow-up and prognosis.

In this chapter we introduce the role of MRI in MS. The first section describes MS and its main characteristics. The second section focuses on the application of MRI in MS, its advantages and its limitations.

1.1 Multiple Sclerosis

Multiple sclerosis is a chronic autoimmune disease which affects the central nervous system (CNS). Although some treatments slow the evolution, to date there is no known cure for MS.

It was first described by Charcot [Charcot 1868] and was named after the scars (in French *sclérose*) that are caused by the disease. The brain is composed of neurons that are divided into cell body and axon (Figure 1.1). The axon of the neuron transmits information and is covered by sheaths of myelin. The myelin sheaths are composed of lipids and proteins and accelerate the transmission of information along the axon. In MS, the myelin is destroyed, causing problems in the transmission of information and, in a later stage of the disease, the axon can be destroyed, causing neurological handicap.

The destruction of myelin can occur at any location in the CNS (brain and spinal cord). Depending on the location of the destruction, the symptoms vary. In this section, we will briefly describe some aspects of MS that might be necessary for the understanding of this manuscript. For more details, the reader can refer to these reference books [Compston et al. 2006, Paty and Ebers 1997].

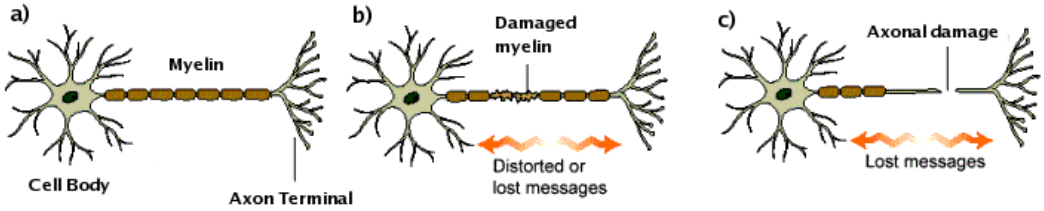


Figure 1.1: Diagram of a neuron and the destruction process: a) Structure of a healthy neuron. b) Demyelination process. c) Destruction of the axon. Modified from msdecision.org.uk .

1.1.1 Pathology

The first description of multiple sclerosis was given by the pathological analysis of the brain [Charcot 1868]. He observed destruction of myelin sheaths and injury of the axons. For more details, the reader can refer to this review paper [Lassmann 2005].

Luccinetti et al. [Lucchinetti et al. 2000] discovered the heterogeneity of MS lesions. They analyzed 51 biopsies and 32 autopsies with histologically proven MS, studied lesions accounting for structural and immunological features and differentiated four different patterns of demyelination.

This heterogeneity in MS could explain the partial results obtained by some treatments. For example, Gold and Hartung [Gold and Hartung 2005] observed that patients with pattern II responded to plasma exchange while patients with patterns I or III showed no response.

1.1.2 Causes

The cause of MS is unknown, although there is a range of factors that lead to higher chances of developing MS.

Several epidemiological studies have been performed to study the distribution of the disease in order to discover environmental factors that could be responsible for the development of the disease. The incidence of the disease varies according to geographical location (Figure 1.2). MS is more common the further we go from the tropics although no determining factors have been found [Paty and Ebers 1997].

Many studies evaluated the genetic susceptibility to MS. For example, they found that the probability of two monozygotic twins to have MS was around 25% while the probability of MS in dizygotic twins was around 4% [Ebers et al. 1986]. Researchers are trying to find the gene or set of genes that causes this susceptibility [Dyment et al. 2004].

The hypothesis of MS being caused by a virus is yet to be discarded. In 90% of patients with MS, high concentrations of IgG antibodies were found. High concentrations of IgG are also found in other inflammatory and infectious diseases where the IgG is directed against the virus. Although no virus has been isolated or directly linked with MS, Gilden [Gilden 2005] suggested that “MS is an immune-mediated

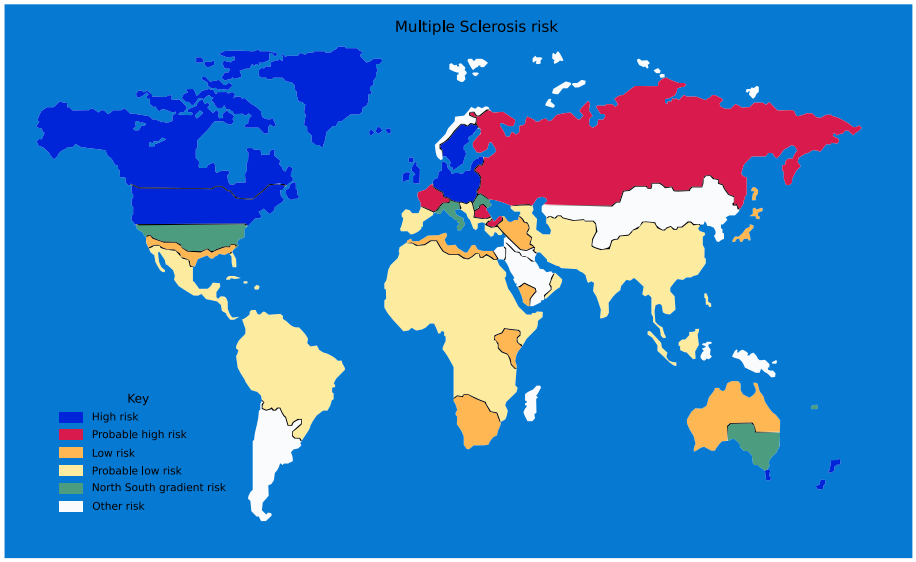


Figure 1.2: Regional prevalence of multiple sclerosis around the world (Source: http://commons.wikimedia.org).

disease, probably triggered by an infectious agent".

1.1.3 Evolution of the Disease

From a clinical point of view, MS was first described as a chronic disease with neurological attacks that could be followed by remission of the symptoms. An *attack*, also called relapse, is defined as a symptom or symptoms of neurological dysfunction which last more than 24 hours [Poser et al. 1983]. *Remission* refers to the improvement of the symptoms of an attack. To be considered remission, the improvement should last at least for one month.

An international survey among neurologists pointed out different types of evolution in patients with MS [Lublin et al. 1996]. Four different clinical types of MS were defined according to this survey (Figure 1.3). No relation has yet been found between the four patterns of demyelination and the four clinical types of MS.

The most common type is the relapsing-remitting MS (RRMS), where sudden relapses of handicap remit after a short period of time, from days to months, leaving little or no handicap. This type accounts for 55% of MS patients. In a later stage of the disease, the majority of RRMS patients converts to secondary progressive MS (SPMS). This type of MS is characterized by a progressively increasing handicap after RRMS period and affects 30% of MS patients.

Progressive-remitting MS (PRMS) is characterized by the progressive evolution of the disease in addition to relapses, and affects only 5% of patients with MS. Primary progressive MS (PPMS) shows a progressive evolution with no relapses and affects 10% of MS patients.

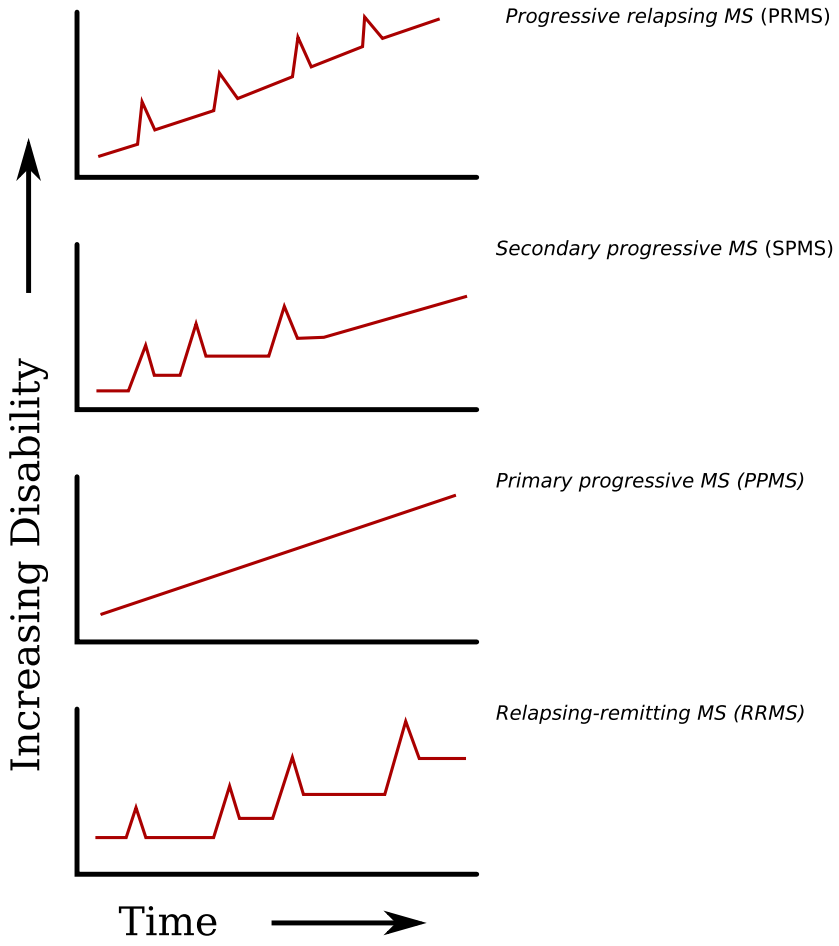


Figure 1.3: Evolution of the four different clinical types of multiple sclerosis. (Source: http://commons.wikimedia.org/)

1.1.4 Diagnosis

Treatments for MS try to preserve the brain from axonal loss. To reduce axonal loss, early diagnosis of MS is important. Multiple Sclerosis is a demyelinating disease and only the histopathological analysis of the brain after death can confirm with absolute certainty the MS diagnosis. Clinical criteria were developed for the diagnosis of MS during patient's life. The objective of these criteria is to verify the dissemination of the disease in time and space.

The first clinical criteria were defined using only neurological examinations [Schumacher et al. 1965]. Then the Poser criteria [Poser et al. 1983] included other tests: CSF analysis, evoked potentials and tissue imaging. These tests were used to confirm the diagnosis but neurological examinations constituted the main source of information. Two relapses with different symptoms were required to verify the dissemination of the disease in time and space.

MRI shows abnormalities in 95% of patients with MS. With the generalization of MRI scanners, several MRI criteria were proposed for the diagnosis of MS. In these criteria, only the number and location of the lesions visible on MRI were used to

make the diagnosis. The Barkhof criteria were the most widely used [Barkhof et al. 1997].

In 2001, the International Panel on the Diagnosis of Multiple Sclerosis presented its new diagnostic criteria [McDonald et al. 2001], known as the McDonald criteria. Basically, these new criteria incorporated the MRI information proposed by the Barkhof criteria into the Poser criteria. In these criteria, MRI plays an important role in demonstrating the dissemination of lesions in time and space. The McDonald criteria were revised in 2005 in order to account for new findings in spinal cord imaging and CSF analysis [Polman et al. 2005].

1.1.5 Disability

The variability in the symptoms and evolution of the disease complicates longitudinal and group studies. Kurtzke proposed a disability scale to evaluate the impairment of MS patients known as the Expanded Disability Status Scale (EDSS) [Kurtzke 1983]. This scale ranges from 0.0 (no disability) to 10.0 (death). and evaluates eight different functional systems: pyramidal, cerebellar, brainstem, sensory, bowel and bladder, visual, cerebral, and other. For each functional system, there is a Functional System Scale (FSS) which measures the handicap on this particular functional system. EDSS is used in clinical practice in the follow-up of patients and in clinical trials to measure the efficacy of new treatments.

EDSS is heavily weighted towards movement handicap. Other measures, such as MSFC [Fischer et al. 1999], were developed in order to take into account the cognitive handicap.

1.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI), previously called nuclear magnetic resonance imaging (NMR), is a medical imaging technique based on the magnetization properties of some nuclei of some elements such as hydrogen, which are contained in the body tissues. MRI is a non-ionizing technique well adapted for repeated examinations and follow-up studies.

In MRI, tissues are, at a given strength of the magnetic field, characterized by their relaxation time constants T1 and T2. Several MR Sequences can be used to acquire images with different intensity contrasts. Depending on the sequence parameters, the images will be weighted on T1, T2 or proton density (PD) . For a better understanding of MRI the reader can refer to these reference books [McRobbie et al. 2003, Haacke et al. 1999].

Intensity abnormalities can be found in the conventional MR images of 90% of patients with MS. Among the conventional MR sequences employed in clinical practice [Traboulsee et al. 2005], there are :

- Dual Echo: two different images are obtained: T2-weighted (T2-w) and proton density weighted (PD).
- T1-weighted without a contrast agent (T1-w).

- T1-weighted after the injection of a contrast agent (Gadolinium) (Gd T1-w): Gadolinium is a paramagnetic ion that modifies the relaxation times and increases the intensity contrast in the inflamed regions of the brain.
- Fluid Attenuating Inversion Recovery (FLAIR): This sequence suppresses the intensity of the fluids in the brain increasing the contrast between the CSF and the periventricular lesions.

Non-conventional MRI techniques [Filippi and Rocca 2005] yielding interesting findings have been proposed but their use is up to now restricted to research. These techniques include: magnetization transfer (MT), diffusion weighted imaging (DWI) and proton MR spectroscopy (^{1}H-MRS). The set of MR sequences that is acquired from a patient in one session is referred to as an MR protocol.

In the next sections we describe the three main aspects studied using MRI: focal lesions, atrophy and normal appearing brain tissues.

1.2.1 Focal Lesions

Conventional MRI images show signal abnormalities in the white matter in the majority of MS patients [Paty and Ebers 1997]. These signal abnormalities have been associated with lesions in combined MRI and histopathological analysis [Groot et al. 2001].

Although lesions also exist in the gray matter [Kidd et al. 1999], they are hard to distinguish on conventional MRI due to the low contrast between the intensity of the gray matter lesions and the intensity of the surrounding gray matter. Geurts et al. [Geurts et al. 2005] compared the detection of cortical lesions in postmortem MRI and histopathology. They only detected 5% of the gray matter lesions observed in the histopathology on FLAIR images.

Several MR sequences are necessary in order to classify the MS lesions. Signal abnormalities are classified into three types of MS lesions according to their intensities in the MR images (Figure 1.4):

T2-w lesions: These lesions have a hyper-intense signal compared to the white matter (WM) in T2-w, PD and FLAIR. These lesions are not specific to any particular process and can be due to inflammation, edema, demyelination or axonal loss.

Gd enhancing T1-w lesions: These lesions an increased of the contrast can be observed after Gd injection on T1-w images and are hyper-intense compared to WM on T2-w, PD and FLAIR. They can also be hypo-intense compared to WM on T1-w before the Gd injection. They are usually less numerous than T2-w lesions and are associated with inflammatory activity.

Black holes: These lesions are hypo-intense compared to WM on T1-w images and usually hyper-intense in T2-w, PD and FLAIR. Some hypo-intensities disappear after a month or two, thus the hypo-intensity must be present at least for three months and must not enhance with gadolinium in order to be considered a black hole. Black holes appear after T2-w lesions and are usually associated with axonal loss.

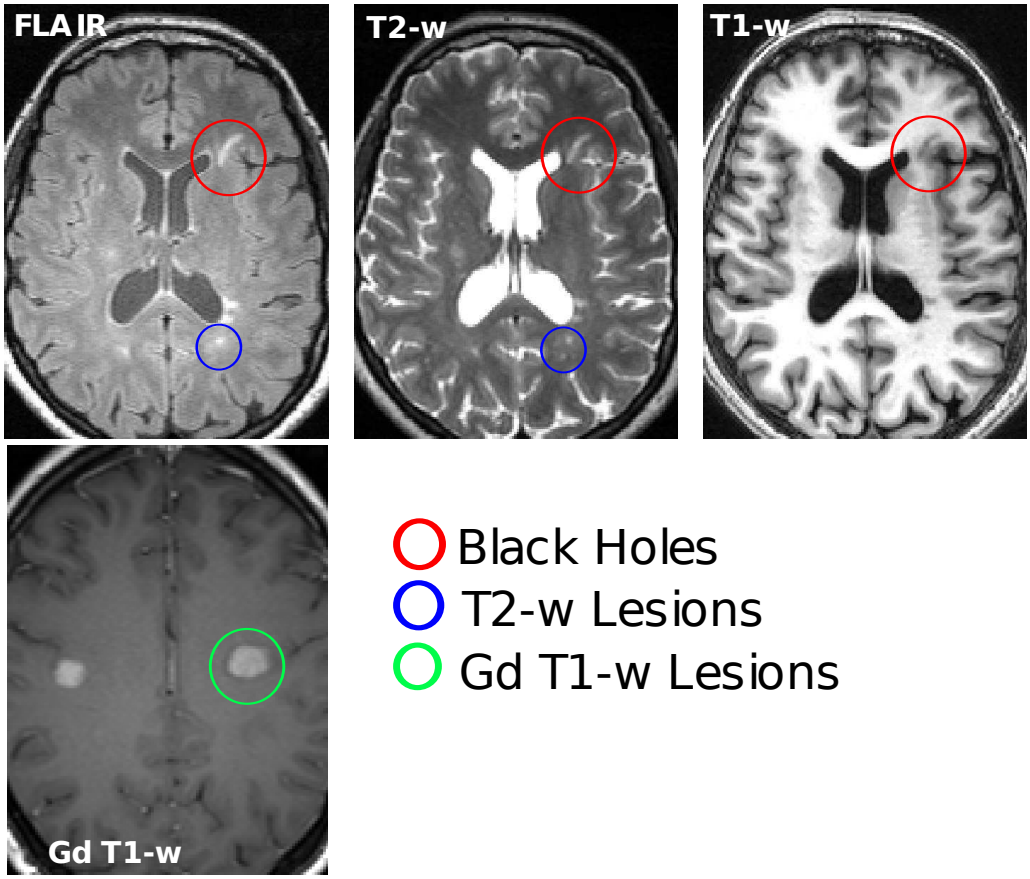


Figure 1.4: Examples of MS lesions on MRI: T1-w, T2-w, FLAIR and gadolinium enhanced T1-w (Gd T1-w).

The number and the volume of lesions that can be observed in MR images depend on the strength of the magnetic field, the slice thickness and the MR sequences employed. More lesions are detected on images acquired on a 4.7T scanner than on images acquired on a 1.5T scanner [Keiper et al. 1998]. Similarly, thin slices increase the lesion volume obtained and its reproducibility [Molyneux et al. 1998]. Finally, the lesion volume measured using FLAIR images is higher than the volume measured using dual-echo (T2-w and PD) images [Filippi et al. 1999].

Many studies correlated MRI lesions and disability, usually EDSS, with low to moderate results (see review [Zivadinov and Leist 2005]). We can point out five possible reasons for these low and moderate correlations:

- *Silent* lesions: they are MS lesions with no associated clinical disability.
- Cortical lesions: they are difficult to detect and are not taken into account in the studies, but are also associated with disability.
- Spinal cord lesions: lesions can appear in the spinal cord but MR images of the spinal cord are difficult to acquire.
- Diffuse lesions: the existence of a more diffuse pathology that cannot be observed in conventional images.

- EDSS: this measure is heavily weighted towards movement handicap.

Due to this lack of correlation, some groups tried to study the influence of the location of the lesions on the disability. In a study with 452 patients, Charil et al. [Charil et al. 2003] studied the correlation of each functional system [Kurtzke 1983] to the anatomical location of the lesions. For each functional system, they found the brain region that correlated the best with the disability. Vellinga et al [Vellinga et al. 2009] performed a similar analysis on EDSS findings, but only periventricular lesions correlated with disability and disease duration.

The limited correlation found between lesions and disability restricts the use of the volume and number of lesions as a secondary outcome in clinical trials [Miller 2004] and maintains the clinical evaluation as the primary outcome in the evaluation of new treatments.

1.2.2 Normal Appearing Brain Tissues

Conventional MRI has enabled the study of focal lesions but is unable to give information about the normal appearing brain tissues (NABT). Non-conventional techniques give more specific information about the brain and enable the study of the NABT. Here, we summarize some findings regarding MS using these techniques; a more detailed review can be found in the following book chapter [Filippi and Rocca 2005].

Diffusion-weighted imaging measures the microscopic movements, or *diffusivity*, of water molecules. Wiest-Daesslé found significant differences in the diffusivity between the NABT of MS patients and controls [Wiest-Daesslé 2009].

The magnetization transfer ratio (MTR) was strongly correlated with demyelination and axonal loss in post-mortem studies [Van Waesberghe et al. 1999]. In two studies, the MTR values in the NABT of MS patients were lower than in the NABT of controls, especially for SPMS and PPMS [Tortorella et al. 2000].

The ^{1}H-MRS, referred to as MR spectroscopy, studies the concentration of metabolites. One of these metabolites, the N-acetylaspartate (NAA), is correlated with axonal loss. A significant decrease in NAA was found in MS lesions but also in the NAWM in MS patients [Stefano et al. 2005].

1.2.3 Atrophy

Early pathological studies described the existence of brain atrophy in MS patients. With the development of imaging techniques, atrophy of the brain has been demonstrated even in the early development of the disease [Chard et al. 2002].

The causes of atrophy remain unclear but two processes seem to occur simultaneously [Bermel and Bakshi 2006]: focal atrophy and diffuse atrophy. The lesions cause the destruction of the myelin and axons, leading to focal atrophy of the brain. Studies show that the focal lesions fail to completely explain the atrophy of the brain [Meier et al. 2004]. The rest of the atrophy would appear to be due to the diffuse demyelination process that affects the NABT.

Many studies tried to correlate the disability with atrophy of the brain. While some authors were interested in atrophy of the whole brain [Rudick et al. 1999],

other authors focus on atrophy of a particular structure of the brain: gray matter, putamen or spinal cord. Spinal cord atrophy was found to have a good correlation with physical disability [Lin et al. 2003]. Recently, new techniques for the detection of local atrophy were developed using non-linear registration [Kezele et al. 2007]. For more details on atrophy in MS, the reader can refer to the following review paper [Bermel and Bakshi 2006].

1.3 Conclusion

Multiple sclerosis is a heterogeneous disease. The existence of different demyelination patterns and several clinical types makes the study of the disease very complicated. The cause of MS is unknown, which makes the search for a cure even more complicated. Nowadays, treatments try to slow down the evolution of the disability and axonal loss; early diagnosis is crucial.

MRI has been proven to be a powerful biomarker in the study of MS. Non-conventional techniques have shown the importance of studying the NABT. These techniques still require further development for their standardization and use in clinical trials. Nowadays, clinical trials use the number and volume of lesions seen in conventional images. In the next chapter, we will focus on the delineation of these lesions on conventional MRI.

Chapter 2

Segmentation of MS Lesions on Conventional MRI

The number and the volume of MS lesions are used as a secondary outcome in clinical trials but the detection and posterior delineation of MS lesions is not trivial. Lesions are usually small and numerous and have fuzzy borders which make them difficult to delineate [Miller et al. 1998].

In image processing, the delineation of lesions is called *segmentation*. Segmentation is the process that divides the images into the target (in this case the MS lesions) and the background. We differentiate segmentation from classification whereby each voxel of the image is assigned to a different class.

In this chapter we provide an overview of segmentation methods of MS lesions and how they are validated. We propose a classification of the segmentation methods in Section 2.1. Section 2.2 provides an overview of interactive segmentation of MS lesions and Section 2.3 describes the automatic methods. Finally, Section 2.4 presents the procedures for the validation of segmentation methods.

2.1 Classification of Methods for MS Lesion Segmentation

Segmentation methods are usually classified according to the mathematical method employed in the segmentation [Suri et al. 2002]. We propose to classify MS lesion segmentation methods according to three factors: human interaction, number of dimensions in the algorithm and way information in the MR sequences is merged.

Depending on the human interaction involved, segmentation methods can be classified as follows:

Interactive: These methods require a human user to segment an image. We differentiate semi-automatic methods, where a computer assists the human in the segmentation process, from the manual segmentation, which is usually used as the reference method.

Automatic: These methods carry out a segmentation without human interaction. A further classification can be made between supervised learning methods and

data-driven methods, where the first group requires images which have already been segmented in order to "learn" how to segment the lesions.

In clinical trials, several images from the same patient are acquired so as to follow the evolution of the disease. We divide the algorithms depending on the number of dimensions employed in the algorithm:

2D: Each slice of the MR image is processed independently [Grimaud et al. 1996].

3D: The MR image is processed as a whole in order to use the anatomical and spatial information [Van Leemput et al. 2001].

3D+t: Some methods include temporal information so as to create a time model of the lesions [Shahar and Greenspan 2004] or to use redundant information across the longitudinal images [Aït-Ali et al. 2005].

Finally, the methods employ different strategies to combine the information contained in the MR sequences:

Mono sequence: MS segmentation is usually done with several MR sequences so as to be as specific as possible, but some methods rely on the information found in one sequence only, for example FLAIR [Khayati et al. 2008, Anbeek et al. 2008].

Multisequence-Parallel: In this case, sequences are segmented individually in parallel, then some rules are created to merge all results in order to complete segmentation of the MS lesions [Johnston et al. 1996, Ardizzone et al. 2002].

Multisequence-Joint: These methods use all the sequences at the same time in order to improve segmentation [Van Leemput et al. 2001, Zijdenbos et al. 1994].

Multisequence-Hierarchical: These methods use each sequence for a specific step, with each step using the result of the previous step [Sajja et al. 2006, Dugas-Phocion et al. 2004a]. They are different from parallel methods in so far as in hierarchical methods, a step is dependent to the previous one, while in parallel segmentation, each sequence is processed independently.

The difference with the parallel methods is that, in hierarchical methods the steps are dependent from each order while in parallel segmentation the processing of each sequence is completely independent.

2.2 Interactive Segmentation

Manual segmentation was the first method to be used to delineate MS lesions. It implies that the expert chooses the voxels he considers to be lesions without any computer assistance. Users usually visualize several sequences at the same time, and have to "paint" the lesion or contour the edges of the lesion depending on the software employed. Volumes computed by manual segmentation are already employed in clinical trials but they require long processing time and have large intra-

and inter-expert variabilities, with up to 44% inter-expert variability in different centers [Zijdenbos et al. 2002].

In order to reduce the variability of manual segmentation, several semi-automatic methods have been developed. As MS lesions are hyper-intense compared to white matter, Wicks et al. [Wicks et al. 1992] proposed a global threshold to identify the hyper-intense regions of the brain where the threshold value was chosen by the expert. The variability of this method was found to be lower than in manual segmentation but small variations in the threshold value imply large variations in volume. Filippi et al. [Filippi et al. 1998] proposed a local thresholding technique whereby the expert chooses local thresholds in order to correctly segment each lesion, the advantage of this method being less sensitive to intensity inhomogeneity.

These two methods show how the interaction required for each method can vary. We can be classify semi-automatic methods into two categories: *lesion-by-lesion* methods and *minimal interaction* methods. On the one hand, lesion-by-lesion methods assist the expert in the delineation of the lesion but the expert needs to detect all the lesions one by one. On the other hand, minimal interaction methods seek to reduce the interaction of the expert helping the expert in both detecting and delineating the MS lesions. The interaction can vary from selecting some voxels from different tissues [Johnston et al. 1996] to segmenting one lesion only [Ashton et al. 2003]. The algorithm then detects and delineates the rest of the lesions.

In the category of lesion-by-lesion methods, region growing methods were proposed whereby the expert select a voxel at the center of the lesion and the region expands automatically to reach the contour of the lesion [Parodi et al. 2002, Ashton et al. 2003]. Grimaud et al. [Grimaud et al. 1996] proposed a gradient-based method whereby the user selects a voxel in the contour of the lesion and the computer detects the contour of the lesion using the gradient information. They showed that their method has less variability than manual segmentation and the global threshold method [Wicks et al. 1992], although the processing time increased.

Lesion-by-lesion methods reduce variability in the delineation of the lesion, but the expert must detect all the lesions. Molyneux et al. [Molyneux et al. 1999] showed that there is a low level of agreement among experienced experts regarding the number of lesions in an image. Hence minimal interaction methods were developed to assist the user in detecting and delineating MS lesions. Ghazel et al. [Ghazel et al. 2006] proposed a texture-based segmentation whereby the user only selects a ROI in the white matter where all the lesions are included. Ashton et al. [Ashton et al. 2003] presented a method based on Bayesian theory whereby only the manual delineation of an MS lesion was necessary. They compared it with a lesion-by-lesion method and manual segmentation. Their method was faster than the lesion-by-lesion method but showed a slightly larger inter-expert variability as the users might have not chosen the same lesion in each patient. According to the authors, the use of both methods can be complementary - the lesion-by-lesion method can be used both to initialize the minimal interaction method and subsequently correct the missed lesions. Some methods require the user to select some points or regions of interest (ROI) in the different tissues. Johnston et al. [Johnston et al. 1996] initialized an iterated conditional modes (ICM) algorithm separately on T2-w and PD using ROI of each tissue. The results of each sequence were then merged in a final segmentation.

Udupa et al. [Udupa et al. 1997] defined fuzzy relations among neighboring voxels using the fuzzy-connectedness theory [Udupa and Samarasekera 1996] and, from some seed points on each tissue, they segmented the MS lesions. Recently, Lecoeur et al. [Lecoeur et al. 2008] proposed a semi-automatic graph cut approach based on the spectral gradient that was applied to MS lesion segmentation.

2.3 Automatic Segmentation

When it comes to processing large databases with hundreds of images, interactive methods require too much human interaction and automatic methods are more suitable.

2.3.1 Supervised Learning Methods

Supervised Learning methods require a initial learning stage where a training database of previously segmented images is used to tune the algorithm. These methods are very powerful when the database covers completely the whole variety of possibilities. The main drawback is the initial training stage.

Creating a training database is a problem in itself. Two main limitations are the segmentation of the training database and the MR protocol of the training set. Supervised methods require a training database where the MS lesions are segmented. These segmentations are usually performed using one of the above-mentioned interactive techniques, thus the training database will be biased towards the expert that performed the segmentation and the technique employed. These methods are mainly used in mono-center studies where the MR protocol is fixed, which limits the variability of the images. In order to use the same algorithm with another MR protocol or MR scanner, another training set must be created and another learning procedure should be done which reduces its usability in multi-centric clinical trials.

Kamber et al. [Kamber et al. 1995] compared four different supervised classifiers and demonstrated the utility of an atlas in order to restrict the search for MS lesions to the white matter to reduce the number of false detections. Sajja et al. [Sajja et al. 2006] proposed a hierarchical method whereby lesions and CSF were first segmented by a Parzen-window density estimator in T2 and FLAIR, and then white matter and gray matter were segmented on T2-w and PD sequences using a Hidden Markov Random Field coupled with an EM algorithm [Zhang et al. 2001]. In addition, some heuristic rules were employed to reduce the number of false positives.

Several authors have proposed the use of the k-nearest neighbor (k-NN) algorithm [Duda et al. 2000]. This algorithm classifies each voxel using information given by the k nearest voxels from the training database. Cárdenes et al. [Cárdenes et al. 2003] combined k-NN algorithm using intensity information in conjunction with distance relabeling and connected-components operations to include spatial information. Anbeek et al. [Anbeek et al. 2004; 2008] used k-NN algorithm including both intensity and spatial information to improve the segmentation. Warfield et al. [Warfield et al. 2000] described a mono-sequence method that combines k-NN algorithm with the nonlinear registration of an atlas in order to merge intensity and

spatial information. Wu et al. [Wu et al. 2006] later extended this method to a multidimensional version.

The artificial neural network (ANN) [Duda et al. 2000] is a mathematical model inspired by biological neural networks that has been widely used in supervised classification. Zijdenbos et al. [Zijdenbos et al. 1994] proposed a trained ANN capable of classifying each voxel as one of the brain tissues or lesions. Later, they improved their method including spatial information in form of an atlas [Zijdenbos et al. 2002], while Blonda et al. [Blonda et al. 1998] proposed the use of a fuzzy neural network. Other methods combined a two-step hierarchical approach whereby candidate lesions were found using the fuzzy c-means (FCM) [Alonge et al. 2001] or fuzzy connectivity [Admasu et al. 2003] algorithms and the ANN was used to determine whether or not the lesion actually exists.

Other supervised learning techniques have also been employed in MS lesion segmentation. Lao et al. [Lao et al. 2006; 2008] classified MS lesions using a support vector machine (SVM) and Kroon et al. [Kroon et al. 2008] proposed the use of a principal component analysis (PCA) adding spatial information included in the form of an atlas. Welti et al. [Welti et al. 2001] included longitudinal acquisitions of FLAIR in PCA to characterize the evolution of the lesion. In order to increase accuracy in the detection of lesions, recent methods have included at the same time spatial, texture and intensity information in the classification. Two examples are Morra et al. [Morra et al. 2008] using the AdaBoost classifier and Akselrod-Ballin et al. [Akselrod-Ballin et al. 2006] using segmentation by weighted aggregation.

2.3.2 Data-driven Methods

Data-driven methods replace the training database with *a priori* information, which make them more robust to changes in the MR protocol. This a priori information is formalized in several ways. We can distinguish between two types of methods: parametric and non-parametric. Parametric methods incorporate a priori information in the form of a mathematical model. This model tries to explain the characteristics of the brain, usually the intensities, and the parameters of the model must be extracted from the images. Non-parametric methods, on the contrary, avoid the use of models and use other image processing techniques to segment the brain.

2.3.2.1 Non-parametric Methods

Atkins et al. [Atkins et al. 2000] proposed a transformation of T2-w and PD images into $\log(\text{T2-w}) - \log(\text{T2-w} + \text{PD})$ where, with three linear regression, they were able to find an optimal threshold for lesion segmentation. Pachai et al. [Pachai et al. 1998] implemented a multi-resolution scheme where automatic local thresholds were computed in each level to segment the MS lesions.

Other methods analyzed the intensities of the brain using clustering techniques to classify the brain voxels. Clustering methods try to separate the data points into different homogeneous clusters. One of these methods is the FCM that has been widely used for the segmentation of healthy brains [Pham and Prince 1999, Hou et al. 2007]. In FCM, each cluster is represented by its center, and each point

has associated a membership value to each of the clusters between 0.0 and 1.0. Boudraa et al. [Boudraa et al. 2000] proposed a hierarchical segmentation method whereby the FCM is used twice in PD-w images, first, to detect the lesion and CSF cluster, and secondly to refine the segmentation of lesions. Similarly, Ardizzone et al. [Ardizzone et al. 2003] employed the FCM twice in PD-w images, it is used first, to detect the parameters of an anisotropic diffusion filter to denoise images, and second, to actually segment the MS lesions. Yang et al. [Yang et al. 2003] proposed a segmentation of T1-w and T2-w images using the FCM in each sequence separately to detect MS lesions and then delineate the lesion correctly applying deformable models segmentation. Recently, Admiraal-Behloul et al. [Admiraal-Behloul et al. 2005] carried out parallel FCM segmentation of T1-w, T2-w and FLAIR images and, then, they proposed to apply some fuzzy rules and to integrate the information of an atlas in order to merge the information found on each sequence. The simultaneous use of several sequences in a joint segmentation can improve the results of clustering methods. For example, Ardizzone et al. [Ardizzone et al. 2002] also proposed a two-channel FCM in T2-w and PD at the same time. In order to include spatial information in the FCM algorithm, Shiee et al. [Shiee et al. 2008] modified the FCM algorithm to include topological constraints of the brain and atlas information.

2.3.2.2 Parametric Methods

All parametric methods follow a similar framework. First, a *model* with a reduced number of parameters is chosen to describe the brain, mainly the intensities of the brain. Second, a *cost function* is associated with the model to measure the degree of similarity between the model and the image. Third, an *estimator* is employed to determine the parameters of the model using the cost function. A widely used parametric method for healthy subjects consists in the use of a Gaussian Mixture Model (GMM), the likelihood as a cost function and the Expectation-Maximization (EM) algorithm [Dempster et al. 1977] to estimate the model parameters. The EM algorithm [Dempster et al. 1977] is an iterative method for estimating the model parameters and has been widely used for estimating the GMM in healthy patients [Wells III et al. 1996, Cuadra et al. 2005] because it is easy to implement and converges at least to a local maximum of the solution. The GMM usually consists of three Gaussians distributions which correspond to the three following classes: grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF).

For MS, this framework has been adapted in different ways in order to include MS lesions. Some authors modified the GMM and added an extra class for MS lesions: a Gaussian class [Kikinis et al. 1999, Warfield et al. 1995] or a uniform class [Rouaïnia et al. 2006]. The different types of existing lesions (T2 lesions, black holes, etc.) make it difficult to model the intensities of the MS lesions, thus other authors considered the MS lesions as *outliers* to the healthy brain model [Van Leemput et al. 2001, Aït-Ali et al. 2005]. An *outlier* is any point in the data that cannot be explained by the model.

Van Leemput et al. [Van Leemput et al. 2001] proposed a weighted EM algorithm whereby the voxels situated far from the model are less taken into account in the estimation and are considered potential lesions. Aït-ali et al. [Aït-Ali et al. 2005]

modified the cost function and employed the trimmed likelihood estimator to avoid the outliers in the estimation as proposed by Neykov et al. [Neykov et al. 2007].

GMM is a basic model for explaining the brain intensities, hence some authors have proposed more complex models. Shahar et al. [Shahar and Greenspan 2004] proposed a 3D+t model for the evolution of MS lesions in order to segment MS lesions in longitudinal images from the same patient. Harmouche et al. [Harmouche et al. 2006] used a model where the brain is divided into different regions because the intensity of each tissue varies slightly in each region. Freifeld et al. [Freifeld et al. 2007] employed a constrained Gaussian mixture model (CGMM), successfully applied to healthy brains [Greenspan et al. 2006], whereby, in addition to the image intensities, the spatial coordinates are also included in the model. CGMM allows each Gaussian to be localized in the space, making the segmentation local instead of global, as it is in the other models. In this method, the final delineation of the lesion was performed by active contours in order to correctly determine the borders.

But EM is not the only algorithm that is used to estimate the GMM. Prastawa et al. [Prastawa and Gerig 2008] used a robust estimator, called the Minimum Covariance Determinant, using an atlas to determine the regions where WM, GM and CSF are located. Khayati et al. [Khayati et al. 2008] proposed an adaptive mixture model, that automatically chooses the number of classes and then the output is reduced to four classes using certain heuristic rules before segmentation.

In order to avoid intensity inhomogeneity effects, several methods included intensity inhomogeneity correction in the estimation of the model for healthy subjects [Wells III et al. 1996, Van Leemput et al. 1999, Prima et al. 2001] and have been employed in MS lesion segmentation [Kikinis et al. 1999, Van Leemput et al. 2001, Rouaïnia et al. 2006]. Spatial information has been included in the estimation by the means of a Markov random field (MRF) whereby the neighbor voxels are taken into account for estimating the class of each voxel [Van Leemput et al. 2001, Rouaïnia et al. 2006, Harmouche et al. 2006] or by the use of a double Markov chain, whereby the information about of the neighbor voxels and of an atlas are coupled to improve the segmentation [Bricq et al. 2008]. Recently, Rousseau et al. [Rousseau et al. 2008] combined a Mean Shift segmentation with a GMM using the *a contrario* approach [Desolneux et al. 2003]. First, the mean shift segmentation created small local regions and then the *a contrario* approach determines which regions did not fit the model and were considered lesions.

Due to the limited resolution of the MRI, voxels on the border between two tissues can contain a mix of the two tissues with an intermediate intensity; this problem is called partial volumes (PV). Different solutions have been proposed to solve this problem in healthy brains with parametric methods [Santago and Gage 1993, Cuadra et al. 2005, Ruan. et al. 2000]. The idea is to model these partial volume voxels with another class. Santago et al. [Santago and Gage 1993] proposed a mathematical approach whereby the partial volume class is calculated from an integral based on the two tissues. Cuadra et al. [Cuadra et al. 2005] considered the PV classes to be a Gaussian in order to simplify the computation of PV. Ruan et al. [Ruan. et al. 2000] defined the PV as several Gaussians whose mean and variance are determined by the two normal tissues. This last model was employed in the MS lesion segmentation by Dugas-Phocion et al. [Dugas-Phocion et al. 2004b].

2.4 Validation of Segmentation

Any segmentation method must be tested in order to evaluate its accuracy, precision and robustness before being applied in clinical studies [Udupa et al. 2006]. This evaluation is usually called validation. The validation of medical image processing is, in general, very difficult due to the lack of ground truth with which to compare it [Warfield et al. 2004]. It is very complicated to evaluate and compare MS lesions for two reasons:

- Methods are not usually freely available, which makes the comparison more complex.

- MR protocols are usually specific to each center. Each method is optimized to a specific MR protocol, making it difficult for one method to be applied to another MR protocol.

Jannin et al. [Jannin et al. 2006] proposed different levels for the validation of algorithms. In segmentation, these levels extend from completely synthetic images used to evaluate the behavior of the algorithm to large multi-center clinical trials used to evaluate the robustness over different MR scanners and acquisitions.

Two of these levels have been employed in the validation of MS segmentation. On the one hand, a database of synthetic images has been developed representing brains from healthy subjects or MS patients [Cocosco et al. 1997, Collins et al. 1998]. This database enables the accurate determination of MR parameters and artifacts providing at the same time a ground truth for the evaluation of the segmentation. On the other hand, researchers used clinical images to evaluate their algorithms in actual conditions and compare their results with the segmentation performed by one or more experts. Recently, a comparison of different algorithms was done. During the Medical Image Computing and Computer-Assisted Intervention conference (MICCAI) in 2008, a workshop was organized to compare different automatic algorithms for the segmentation of MS lesion. [Styner et al. 2008].

2.4.1 Brainweb

Brainweb (BW) is a database of synthetic MR images available online[1] [Cocosco et al. 1997, Collins et al. 1998]. A healthy subject was scanned twenty times to obtain a high-SNR image. From this image, a healthy anatomical phantom was created where each voxel belongs to a specific tissue class. Simulated MR images are created thanks to this phantom and an MR simulator. In the MR simulator, we can choose the MR parameters (TE, TR, resolution, sequence, etc.) and some artifact parameters (noise and intensity inhomogeneity). Three conventional sequences are available: T1-w, T2-w and PD. From the healthy phantom, three different MS phantoms are created with different lesions loads: mild, moderate and severe. The great advantage of this database is the existence of a ground truth that enables a direct comparison of the segmentation results with the phantom.

[1] http://www.bic.mni.mcgill.ca/brainweb/

The usual measure employed in MS lesion segmentation is the Dice Similarity Coefficient (DSC) which is computed as follows

$$DSC = \frac{2 \times |R \cap S|}{|R| + |S|} \tag{2.1}$$

where R is the reference segmentation and S is the segmentation. DSC ranges from 0.0 to 1.0 (perfect segmentation), with a value of 0.7 generally considered to be a good segmentation [Zijdenbos et al. 1994].

These phantoms have been used to evaluate MS lesion segmentation algorithms due to the simplicity of the evaluation and it is a good comparison method although there are three limitations:

Only one phantom: As there is only one brain model, there is no anatomical variability, although efforts were made to address this issue [Aubert-Broche et al. 2006].

Too simple: Although it is based on real images, the final image is not completely realistic and it is easier to segment than are clinical images.

Limited to T1-w, T2-w and PD: Other useful MR sequences such as FLAIR are not present, limiting the validation of some methods.

Supervised methods require a training database and are difficult to evaluate as there is only one phantom. Zijdenbos et al. [Zijdenbos et al. 2002] used a database of clinical images as a training database and then applied their method to the Brainweb database. On the contrary, Akselrod-Ballin et al. [Akselrod-Ballin et al. 2006] employed half of the brain for training and validated their algorithm on the other half, which prevents the comparison of their results with other algorithms.

Even though many authors have evaluated their methods using Brainweb, it is difficult to compare results because the number of slices and images used in the validation varies from author to author. Some authors [Akselrod-Ballin et al. 2006, Freifeld et al. 2007, Zhu and Basir 2003] choose a limited number of slices for their comparison instead of the whole brain, making their results impossible to compare with those of other methods.

The MR parameters are always set to their default values, but there are variations in lesion load, noise and intensity inhomogeneity. The evaluation of some clinical images showed that *normal* parameters are 3% noise[2] and 20% inhomogeneity[3] [Zijdenbos et al. 2002] , which have been used by some authors on the phantom with moderate lesion load [Zhu and Basir 2003, Akselrod-Ballin et al. 2006]. Other authors preferred to remove the intensity inhomogeneity in their evaluations [Aït-Ali et al. 2005, Rousseau et al. 2008]. For instance, only two authors vary the parameters of the phantom to provide an indication of the response of their method. Freifeld

[2]For the Rician noise, the notation 3% means that the Gaussian noise used in complex domain of the image is equivalent to $N(0, v \cdot (3/100))$, where v is the value of the brightest tissue in the image (150 for T1-w and 250 for T2-w) [Coupé et al. 2008].

[3]The notation 20% refers to a maximum deviation in intensity over the ideal uniform case, ±10% over the intracraneal volume [Zijdenbos et al. 2002].

Name	Lesion Load	Noise	Inhomogeneity	Sequences
bw_mild	mild	3%	20%	T1-w, T2-w & PD
bw_moderate	moderate	3%	20%	T1-w, T2-w & PD
bw_severe	severe	3%	20%	T1-w, T2-w & PD

Table 2.1: Typical synthetic images employed in the evaluation of the segmentation methods of MS lesions

et al. [Freifeld et al. 2007] demonstrated the robustness to noise and Zijdenbos et al. [Zijdenbos et al. 2002] varied intensity inhomogeneity and noise in order to give an overall idea of the performance of their algorithm in presence of image artifacts. The performance of the segmentation can vary depending on the lesion load but few authors employed the three phantoms in their evaluation [Akselrod-Ballin et al. 2006, Rousseau et al. 2008].

In conclusion, and in spite of the utility of this synthetic database, it is difficult to compare the results obtained by the different methods due to this variability. We believe that the use of Brainweb should be standardized to allow authors to compare their methods with previous ones. This includes the use of the same metrics, acquisition parameters and validation conditions.

We employ some images of this database in order to study the parameters of our methods in next chapters. These images are summarized in Table 2.1.

2.4.2 Clinical Images

Segmenting MR images from MS patients is a necessary step in the evaluation of a segmentation method. The main issue is the lack of ground truth with which evaluate the segmentation results. As manual segmentation is considered to be the gold standard for MS lesion segmentation, semi-automatic and automatic methods have been compared to manual segmentation

Semi-automatic methods, especially lesion-by-lesion methods, allow the expert to modify the segmentation until it is found to be satisfactory. In this context, methods are evaluated not according to their precision but according to their ability to reduce inter- and intra-expert variability and processing time. Images are segmented several times by several experts in order to measure TLL variations in both manual and semi-automatic methods. Semi-automatic methods [Filippi et al. 1998, Grimaud et al. 1996, Parodi et al. 2002, Ashton et al. 2003] showed a reduction inter- and intra-expert variability compared to manual segmentation, although in some cases the processing time increased [Grimaud et al. 1996]. Other minimal interaction methods [Johnston et al. 1996, Ghazel et al. 2006, Lecoeur et al. 2008] are compared with a manual segmentation only.

Due to the absence of a ground truth, automatic methods are compared with manual segmentation that is considered to be the gold standard in MS lesion segmentation. Most methods are compared with only one expert's segmentation [Akselrod-Ballin et al. 2006, Sajja et al. 2006, Khayati et al. 2008, Lao et al. 2008, Aït-Ali 2006, Anbeek et al. 2004]. Although this evaluation provided an indication of the similar-

ity of the segmentations, the significance of the results obtained with this evaluation is limited due to inter- and intra-expert variability found in manual segmentation; it is not possible to differentiate the dissimilarities between manual and automatic segmentation due to errors in the algorithm or due to variability in the manual segmentation. For this reason, some authors compared their methods with more than one expert in order to account for manual segmentation variability [Admiraal-Behloul et al. 2005, Zijdenbos et al. 1994; 2002, Warfield et al. 1995, Wei et al. 2002, Harmouche et al. 2006]. For example, Harmouche et al. [Harmouche et al. 2006] created a consensus ground truth, whereby a voxel is considered a lesion if at least three out of five experts considered it as a lesion, thus reducing the effects of variability. Zijdenbos et al. [Zijdenbos et al. 2002] having carried out a full validation using the Brainweb phantom, compared their automatic method with manual segmentation performed in seven different MS centers. MS centers showed a large variability in the estimation of the lesion load ($44\% \pm 20\%$).

Accuracy is not the only important aspect of the validation; the reproducibility of the measures is also crucial for longitudinal trials. In order to test reproducibility, MS patients undergo MR imaging several times within a short period ranging from several hours to several days. As the images are obtained within a short space of time, it is assumed that the disease has not evolved during this period. Guttmann et al. [Guttmann et al. 1999] used this method to evaluate how each of the steps in their segmentation workflow varied and conclude that the MR acquisition generated more variability in the result than their segmentation method did. Similarly, Asthon et al. [Ashton et al. 2003] found greater variability in the semi-automatic segmentation of different MR scans of the same patient than in different segmentations of the same image, leading to the same conclusion.

2.4.3 3D Segmentation in the Clinic II: MS Lesion Segmentation

At the MICCAI 2008 conference, a workshop was organized for the comparison of segmentation algorithms [van Ginneken et al. 2007]. The objective of this workshop was to compare different segmentation methods in the form of a competition. On this occasion, one of the segmentation challenges was the segmentation of MS lesions [Styner et al. 2008][4].

The organizers gathered a database of images (T1-w, T2-w, FLAIR, DWI) from two different sites with more than 50 images. Images were manually segmented by two experts from different sites so as to measure the level of variability. The images were registered and up-sampled to isotropic 0.5 mm. The database was separated in three parts:

Training set: 20 images and their manual segmentations were given to the participants in order to adapt the algorithms to the MR protocol and the experts' definition of lesion. Supervised methods used this database in order to train their methods.

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

"Off site" test set: 25 images, without the segmentations, were given to the participants to be processed in their own laboratories. The automatic segmentation results had to be sent to the organizers for evaluation.

"On site" test set: 7 images were given to the participants during the workshop with a limited amount of time to be processed.

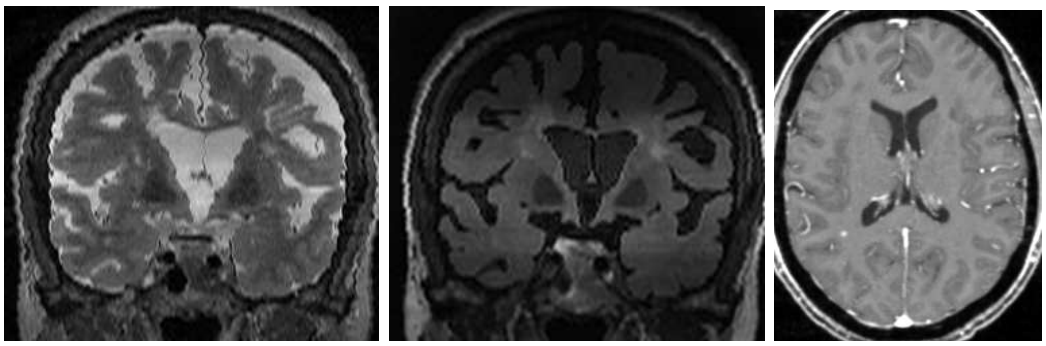


Figure 2.1: Some artifacts of the workshop database (from left to right): CHB test1 06 T2-w and FLAIR, UNC test1 Case07 T1-w (probably Gadolinium-enhancing T1-w).

There are three limitations that need to be discussed: the quality of the images, the preprocessing and the manual segmentation performed by the experts. First, multiple artifacts were found in several images due to the movement of patients during acquisition (Figure 2.1). The quality of the segmentation can be greatly affected by these artifacts, making it a challenge to evaluate the robustness of the algorithm rather than its accuracy. Second, the segmentation is usually performed after preprocessing steps, such as inhomogeneity correction and denoising. The challenge proposed already registered and up-sampled images, reducing the effectiveness of the denoising methods as the assumption of the independence of the noise was no longer valid. This could reduce the performance of the algorithms where denoising is necessary for the segmentation. Third, the manual segmentation was performed by two experts from different centers and showed great variability. For example, the two experts only agreed on 68% of the lesions, which shows the great variability between them.

Despite these limitations, this workshop is an interesting initiative to compare the algorithms. Souplet et al. [Souplet et al. 2008] were the winners of the challenge while the method we present in Chapter 4 finished fourth.

2.5 Conclusion

In this chapter, we presented an overview of methods for the segmentation of MS lesions and how they are validated. We discussed the interests of both interactive and automatic approaches, and how the latter are necessary for large clinical trials. We proposed a classification the methods found in the literature and made a distinction between supervised and data-driven methods. We described how supervised methods

have the drawback of necessitating a training period, while data-driven methods may be more robust as they are based on the information in the image. Among the data-driven methods available, we described in detail the parametric methods where the distribution of the intensities of the brain are considered to follow a mathematical model. These methods have been widely used with good results.

Finally, we stressed the importance and difficulty of validating the methods described. Synthetic images give a first evaluation of segmentation methods under controlled conditions but clinical images are necessary in order to fully assess the performance of a method. Recently, the MICCAI segmentation challenge proposed a good starting point for the comparison of methods although several important aspects required to be improved.

Chapter 3

MS Lesion Segmentation Workflow

The term *segmentation method* is somewhat ambiguous. Some authors include in this term the whole series of image processing algorithms, from the raw data to the classified image. Other authors consider the segmentation method to be just the final step, where, once the images are registered and corrected for image artifacts (e.g. noise, intensity inhomogeneity, etc.), the images are classified. In this document, we use the latter definition of *segmentation method* and speak about *segmentation workflow* when we want to consider all the steps together.

There are several processing steps that must be performed in order to prepare the data for segmentation. These processing steps usually include denoising, intensity inhomogeneity correction, registration and skull-stripping. Several authors [Zijdenbos et al. 2002, Sajja et al. 2006, Kikinis et al. 1999] proposed workflows where different algorithms are used in a different order. Each segmentation method needs a workflow which is adapted to its requirements in order to obtain the best possible segmentation.

In this chapter, we describe the workflow proposed for MS lesion segmentation. The processing steps are described in Section 3.1, and the way these steps are connected in our segmentation workflow is presented in Section 3.2.

3.1 Processing Steps

In this section, we describe each of the processing steps required for our MS lesion segmentation workflow.

3.1.1 Denoising

The MR images are corrupted by noise. This noise modifies the intensity of the voxels and is mainly caused by the MR scanner and Brownian noise. Although traditionally MR image noise is considered to have a Rician distribution [Gudbjartsson and Patz 1995], images may have other noise distributions depending on the MR sequence and the reconstruction method employed [Dietrich et al. 2008].

High levels of noise can affect the usability of the images, especially for automated registration and segmentation methods. For this reason, many denoising methods have been proposed [Buades et al. 2005]. Recently, Coupé et al. [Coupé et al. 2008]

proposed an optimized non-local means (NLM) algorithm that outperforms classical denoising methods such as anisotropic diffusion and total variation minimization.

3.1.2 Intensity Inhomogeneity Correction

Besides noise, another undesired effect on MR images is known as intensity inhomogeneity (IIH). This effect causes a smooth variation of intensity in homogeneous tissues across the images and is produced by the sum of two different sources [Vovk et al. 2007]. On the one hand, there are MR scanner related sources [Sled et al. 1998] which include the inhomogeneity of the B_0 field, the limited bandwidth of the scanner and the radio frequency. On the other hand, the dielectric and permeability properties of the imaged object can also cause intensity inhomogeneity. Sled and Spike [Sled et al. 1998] measured IIH variations of around 20% in the same tissue.

This effect hardly affects human readers' interpretation of images but it can greatly affect image processing algorithms such as registration and segmentation methods, as one of the usual assumptions is the spatial homogeneity of tissue intensity. Many methods have been proposed for IIH correction in recent years (see reviews [Vovk et al. 2007, Hou 2006]) which can be divided into two types of methods: prospective and retrospective. Prospective methods are performed before the acquisition using phantoms or special acquisitions [Vovk et al. 2007] but only correct MR scanner inhomogeneities. Retrospective methods are based on information extracted from the image and on some a priori information. The advantages of retrospective methods are that they can be automatic, correct all sources of IIH and require no further acquisitions.

One group of retrospective IIH correction methods is based on segmentation. These methods iteratively perform tissue classification and correct IIH [Wells III et al. 1996, Prima et al. 2001]. These methods work correctly on T1-w where good intensity contrast among tissues is observed, but, in sequences with less contrast among tissues between GM and WM, such as T2-w or PD, they can encounter problems to separate the two tissues. Some of these methods have been adapted, as we mentioned in the last chapter, for MS lesion segmentation [Van Leemput et al. 2001, Kikinis et al. 1999, Rouaïnia et al. 2006].

An interesting group of retrospective IIH correction methods is based on histogram analysis [Vovk et al. 2007]. These methods make little or no assumptions about the MR sequence employed and correct the IIH by information minimization [Likar et al. 2000, Mangin 2000] or by high-frequency maximization [Sled et al. 1998]. These methods can be employed directly in MS patients [Sled et al. 1998], although some undesired effects may occur if MS lesions are considered to be IIH.

3.1.3 Registration

In order to employ the information yielded from the different MR sequences during the segmentation process, images must be spatially aligned. Different image resolutions or patient's movements during the acquisition may lead to spatial misalignments.

Registration is the process whereby two images are spatially aligned. We refer to mono-modal registration for images of the same modality (or sequence) and to multimodal registration for images of different modalities. Depending on the transformation of the image, one of two types of registration may be used (linear or non-linear). Linear transformations are simple space transformations that can be represented by a 4x4 matrix, and include scale changes, rotations, translations and stretches. Non-linear transformations are mainly used to map images into a template or to detect local movements in order to measure the deformation between the two images.

In MS lesion segmentation, all MR sequences from the same patient are registered together so as to avoid slight misalignments caused by the patient's movement. A multimodal registration with a rigid transformation is used. Rigid transformation is a particular linear transformation whereby the image can only be translated and rotated. This transformation is employed for the registration of images of the same patient because only head movements are searched.

Maes et al. [Maes et al. 1997] proposed a multimodal registration technique based on mutual information that has been widely employed on medical image processing. Later, Tan et al. [Tan et al. 2002] evaluated its accuracy and its influence in MS lesion segmentation and found that errors in the repositioning of the patient in the scanner were more important than the error of misalignment after registration of images. Recently, Wiest-Daesslé et al. [Wiest-Daesslé et al. 2007] proposed a new optimization algorithm to improve the efficiency of the algorithm.

3.1.4 Skull Stripping

In order to simplify MS lesion segmentation, many methods require a preliminary step where all the non-brain voxels are removed, which is commonly named *skull stripping*. It is a difficult task because the brain is very heterogeneous and its surrounding structures can be easily confused with the brain. There are many published automatic skull stripping methods that can be roughly divided into three categories [Ségonne et al. 2004]: region-based, boundary-based and hybrid. Region-based methods are mainly based on intensity information and use morphological filtering to identify the brain. Boundary-based methods model the brain as a surface and fit this surface to the image. Hybrid methods combine the latter two methods.

Boundary-based methods focus on the border of the brain and can be more robust to intensity abnormalities such as MS lesions in the center of the brain. One of these methods Brain Extraction Tool (BET) [Smith 2002] used a deformable model that is adapted with local forces to the border of the brain.

A complete comparison of four skull striping methods was conducted recently [Fennema-Notestine et al. 2006]. Souplet et al. [Souplet et al. 2007, Souplet 2009] later compared five methods on 30 MS patients using as ground truth the combination of the five methods with STAPLE [Warfield et al. 2004]. This kind of evaluation without experts' segmentation was studied by Bouix et al. [Bouix et al. 2007] and proved to have interesting results although it "may not be sufficient to a precise performance evaluation".

3.1.5 MS Lesion Segmentation

The result of this step is a segmentation image where the voxels are classified into four classes: gray matter (GM), white matter (WM), cerebrospinal fluid (CSF) and MS lesions. An overview of the state of the art in MS lesion segmentation was given in Chapter 2. For our workflow, we propose three different methods, explained in Chapters 4, 5 and 6. The segmentation method assumes images are registered and corrected for acquisition imperfections such as IIH and noise. In addition, segmentation methods only work on brain voxels thus they required to perform skull stripping before segmentation.

3.2 Proposed Workflow

Our objective is to create a workflow that uses as input the raw images from the MR scanner, and generate a segmentation image where each voxel is classified as WM, GM, CSF, MS lesions or background.

In order to construct our workflow, we select one algorithm for each processing step. In theory, we should compare all the possible combinations of algorithms so as to obtain the best workflow, but the number of algorithms available makes this approach impossible.

The NLM denoising method [Coupé et al. 2008] showed better performance than the anisotropic diffusion method employed by other workflows [Zijdenbos et al. 2002, Sajja et al. 2006], so we selected it for our workflow. The BET [Smith 2002] skull stripping method was chosen due to the reduced number of parameters needed and because it has already been successfully used in MS lesion segmentation [Akselrod-Ballin et al. 2006, Lao et al. 2008].

Similarly, the registration implementation proposed by Wiest-Daesslé et al. [Wiest-Daesslé et al. 2007] was chosen as it improves performance over classical approaches [Maes et al. 1997]. We chose the FLAIR image as the reference image and the other sequences are registered to it. In the absence of FLAIR, the target image is the T2-w image because the PD image is already registered as both images are acquired simultaneously.

Arnold et al. [Arnold et al. 2001] compared different IIH methods but this evaluation is insufficient for us to choose an IIH correction method because T1-w is evaluated only and because only healthy subjects are used. In MS, IIH correction must be applied to all sequences and special attention has to be paid to the MS lesions, as some methods may consider MS lesions to be IIH. In order to choose a IIH method, we made a comparison of five freely available IIH correction methods which is described in the next section.

A second evaluation was performed so as to decide the order and impact of the different algorithms. Several workflows are compared in order to measure the effects of the processing algorithms and the best manner in which to order them in the workflow.

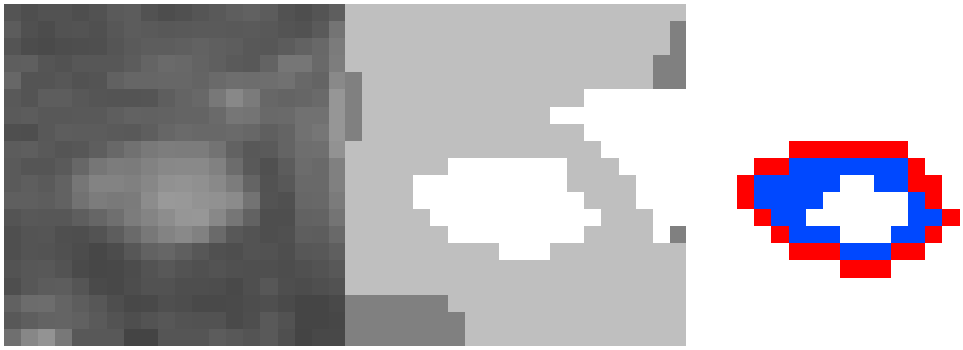


Figure 3.1: Example of lesion processing using the Coefficient of Lesion Variation (CLV). From left to right: T2-w image, manual segmentation, and resulting inner and outer rings.

3.2.1 Comparison of IIH Correction Methods

We propose a specific comparison for the correction of IIH on images from MS patients, taking into account two main aspects: the reduction of IIH in the NABT and the preservation of the contrast between MS lesions and the white matter.

Two measures have been widely employed in the evaluation of IIH correction methods: the coefficient of variation (CV_i) for one tissue and the coefficient of joint variation ($CJV_{i,j}$) for two tissues [Likar et al. 2001].

The coefficient of variation measures the variation of the tissue i is defined as

$$CV_i = \frac{\sigma_i}{\mu_i} \tag{3.1}$$

where σ_i is the standard deviation of the tissue, and μ_i its mean. An IIH correction method must reduce the CV_i for all the brain tissues.

In addition, the $CJV_{i,j}$ was proposed to ensure that the contrast between two tissues i and j was not reduced when the variation within each tissue was corrected. It is defined as

$$CJV_{i,j} = \left| \frac{\sigma_i + \sigma_j}{\mu_i - \mu_j} \right|. \tag{3.2}$$

These measures can be used to evaluate the impact of IIH correction in the NABT, but as the intensity of MS lesions may not be homogeneous, its application to MS is limited. We propose to create a new metric in order to compare the effects of IIH correction in the surroundings of MS lesions. The goal of this specific measure is to evaluate whether the IIH correction method keeps the contrast between the lesion and the surrounding NABT.

From the manual segmentation of each MS lesion, we use mathematical morphology to create two ring-like regions of 1 voxel-width, with one just inside and the other just outside the contour of the lesion (Figure 3.1). The CJV of each lesion is then calculated within those two regions and the average is computed for each image. We call this measure the Coefficient of Lesion Variation (CLV).

In this study, we apply IIH correction to T1-w, T2-w and FLAIR images. Two MS patients are included in this study. Images from the first patient were acquired on a 3T Philips (1mm isotropic 3D T1-w, 3-mm slice thickness T2-w and 3-mm slice

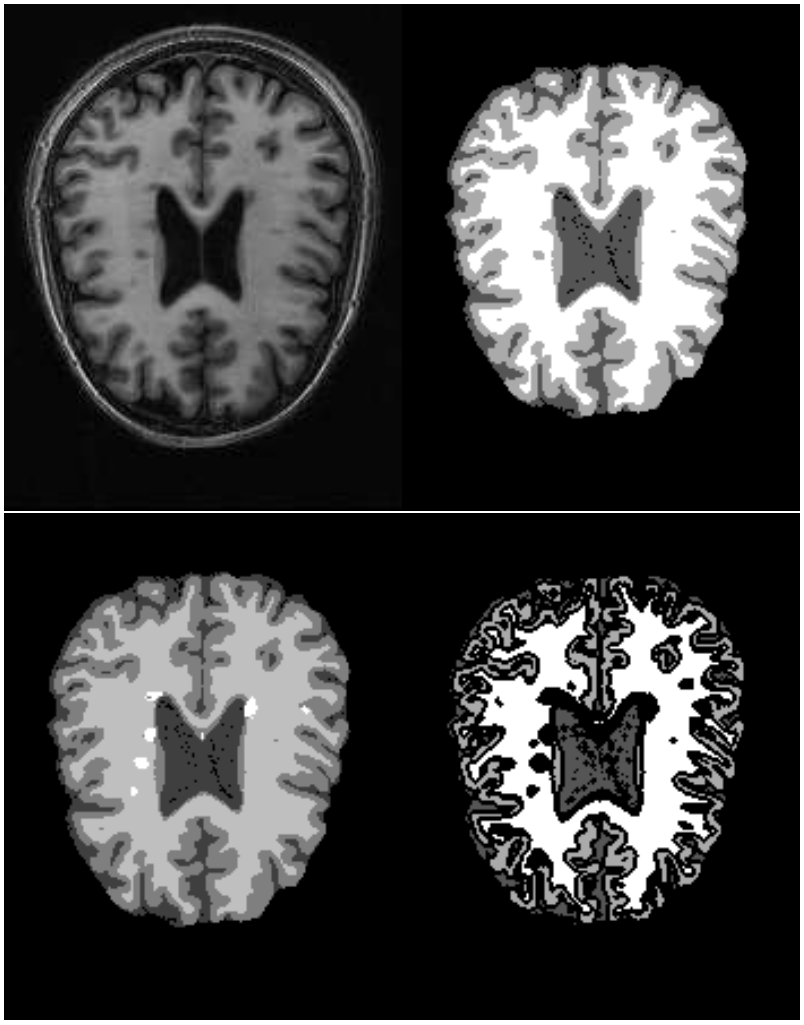


Figure 3.2: Example of regions employed for the evaluation. Top line, from left to right: T1-w image and automatic classification. Bottom line, from left to right: manual post processing and final regions.

thickness FLAIR). Images from the second patient were acquired on a 3T Siemens TRIO with two different protocols, the "2D" (T1-w, T2-w and FLAIR 3-mm slice thickness) and "fully 3D" (1mm isotropic 3D T1-w, 1mm isotropic 3D T2-w and 1mm isotropic 3D FLAIR).

The MS lesions for each patient were segmented manually. Grey matter and white matter were segmented automatically [Zhang et al. 2001] and the results were edited manually where necessary. In order to avoid partial volumes, the mask of each tissue was eroded using mathematical morphology (Fig. 3.2).

Five different and freely available methods were compared in order to select the best IIH correction method for NABT and MS lesion segmentation. Among all IIH correction methods, only those that are not based on segmentation have been used because they can be applied to any sequence without any *a priori* information. All methods are used with the default parameters. The methods compared are:

- BrainVISA [Mangin 2000] (`http://brainvisa.info/index_f.html`)

Sequence	Measure	Brainvisa	Manjon	N3	SPM2	Likar
T1-w	CV	46.4%	41.4%	26.5%	21.9%	17.3%
T1-w	CJV	-0.4%	1.3%	16.5%	5.6%	-8.0%
T1-w	CLV	13.8%	12.9%	7.4%	6.3%	6.7%
T2-w	CV	3.2%	10.7%	2.7%	8.9%	11.1%
T2-w	CJV	8.4%	29.4%	5.9%	25.8%	14.6%
T2-w	CLV	-1.1%	-0.2%	0.1%	-0.4%	0.0%
FLAIR	CV	31.1%	23.7%	16.5%	15.1%	10.1%
FLAIR	CJV	72.9%	75.%	65.1%	67.4%	1.6%
FLAIR	CLV	5.3%	5.0%	3.3%	2.1%	1.7%

Table 3.1: Results for tissue comparison for IIH correction methods.

- Manjon [Manjón et al. 2007] (`http://personales.upv.es/jmanjon/bias/index.htm`).
- N3 [Sled et al. 1998] (`http://www.bic.mni.mcgill.ca/software/minc/`)
- SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/software/spm2/`)
- Likar [Likar et al. 2000] (`http://lit.fe.uni-lj.si/shading/`)

For each method, four different measures were computed: CV, $CJV_{\mathrm{GM,WM}}$ and CLV. The results are summarized in Table 3.1.

All IIH methods were primarily designed for T1-w images (Figure 3.3), and therefore the best improvements are shown with this type of sequence. The BrainVISA method obtained the best results for CV and CLV although it failed to separate GM and WM on 3mm T1-w images. N3 proved to be a better option in terms of distinguishing GM and WM. The impact of IIH on other MR sequences has attracted little interest in the literature, but our results show that it should also be taken into account for FLAIR images. BrainVISA and Manjon's method obtained the best results, followed by N3 and SPM2. Finally, only minor improvements are found for T2-w images where in our testing images IIH is less pronounced.

We decided to include in our workflow the IIH algorithm used by BrainVISA software [Mangin 2000]. The results of N3 are almost as good as those of BrainVISA, but technical aspects simplify the inclusion of BrainVISA in our workflow. We avoid to correct IIH on T2-w images as this effect is less present in these images.

3.2.2 Evaluation of the MS Lesion Segmentation Workflow

The objective is to find the best combination of algorithms in order to improve the quality of the segmentation. We evaluated different workflows and their precision compared to manual segmentation performed by an expert. In this evaluation, we used a segmentation method developed in our laboratory by Aït-Ali [Aït-Ali 2006]. We focus our attention on the use of the denoising method, the non-local means [Coupé et al. 2008], and the IIH correction method [Mangin 2000]. Skull-stripping and registration are mandatory in order to perform the segmentation.

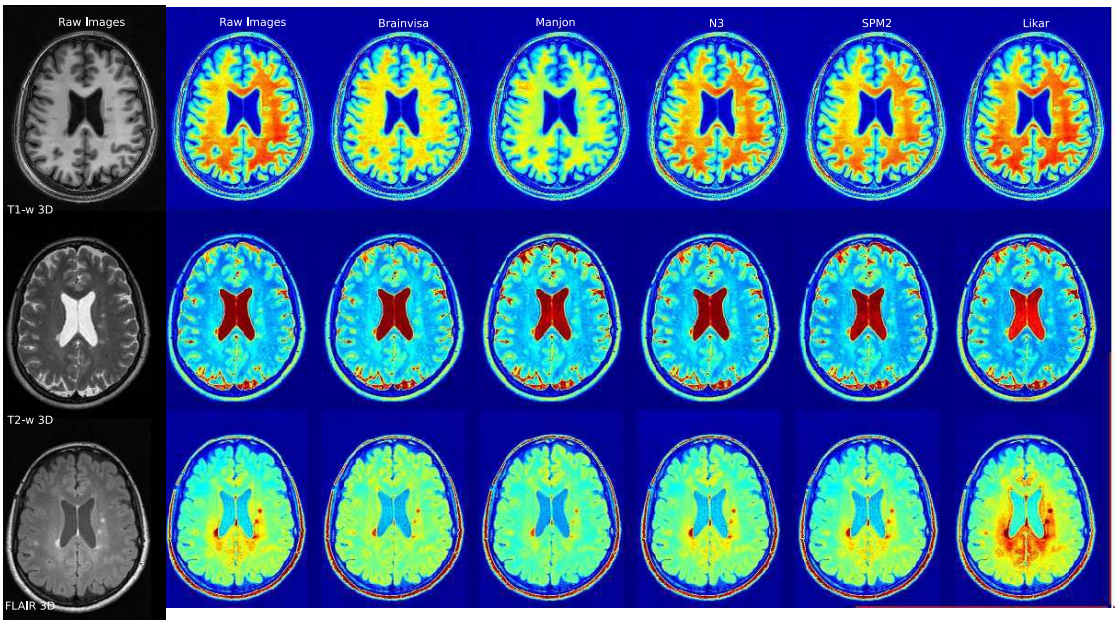


Figure 3.3: Results for IIH correction methods in the 3D protocol

Three different MR protocols are used in the experiments. For datasets including Dual Echo acquisitions, the PD image is discarded because of poor contrast. For each subject, an expert reader manually segmented the MS lesions on the FLAIR sequence. We consider the manual segmentation to be the ground truth. The protocols are:

subject1 Images acquired on a 3T Siemens TRIO: 3D 1mm isotropic T1-w, 2D 3-mm axial slice thickness Dual Echo (T2-w and PD) and 2D 3-mm axial slice thickness FLAIR.

subject2 Images acquired on a 3T Philips ACHIEVA: 3D 1mm isotropic T1-w, 2D 3-mm axial slice thickness Dual Echo (T2-w and PD) and 2D 3-mm axial slice thickness FLAIR.

subject3 Images acquired on a 3T Siemens TRIO: 3D 1mm isotropic T1-w, 3D 1mm isotropic T2-w and 3D 1mm isotropic FLAIR.

The same parameters are used in every step for all subjects. Five different workflows are tested in this experiment:

Basic: No preprocessing before registration.

NLM: Denoising before registration.

IIH: Intensity correction before registration.

IIH+NLM: Intensity correction and then denoising before registration.

NLM+IIH: Denoising and then intensity correction before registration.

The different workflows and segmentation methods are comparing by assessing the quality of the segmentation. In order to evaluate the differences between the automatic segmentation and the ground truth, we use the Dice Similarity Coefficient (DSC) that is commonly used in segmentation evaluation [Zijdenbos et al. 1994].

Table 3.2 shows the results of all different workflows. The NLM+IIH workflow displays better results for each subject. The preprocessing task which most impacts the segmentation results is IIH correction. This can be explained by the fact that the segmentation method [Aït-Ali 2006] is based exclusively on intensities and therefore spatial IIH causes poor detection of MS lesions.

The effects of denoising are more difficult to evaluate. Denoising without IIH correction may not be of any help if images have a large IIH as in Subject 3. Subject 1 shows little or no improvements after the correction; the T1-w image has a low contrast between GM and WM that leads to a poor segmentation.

In addition, the order of preprocessing steps is not obvious. Denoising methods are supposed to work better with piecewise constant regions, that is why IIH correction is usually used before denoising [Montillo et al. 2003]. The drawback is that IIH correction locally changes the nature of the noise when correcting a multiplicative inhomogeneity field affecting the denoising algorithm.

	Subj1	Subj2	Subj3
Basic	0.31	0.20	0.42
NLM	0.33	0.29	0.42
IIH	0.31	0.49	0.53
IIH+NLM	0.30	0.46	0.48
NLM+IIH	0.31	0.49	0.56

Table 3.2: DSC values for all subjects and workflows.

In this evaluation, there are not enough patients to be able to make a firm statement about the best workflow. In any case, we can deduce two main ideas from the results: IIH correction is very important and greatly improves the results, and the image denoising should be performed before IIH correction.

3.3 Conclusion

In this chapter, we explained the concept of a workflow. The segmentation of MS lesions is a complex task that can be divided into smaller tasks. We described each of these tasks and explained our choice for each task as well as the optimal combination we found among the processing methods.

We proposed a workflow for MS lesion segmentation (Figure 3.4). First, images are denoised and corrected for intensity inhomogeneity (except T2-w images). Then FLAIR is used as the reference, all images are registered rigidly and T1-w is used to obtain the brain mask. The last step is to apply the segmentation method with all the registered images and the brain mask.

In the next three chapters, we propose three different segmentation methods. Any of the three methods can be used as the final step in this workflow.

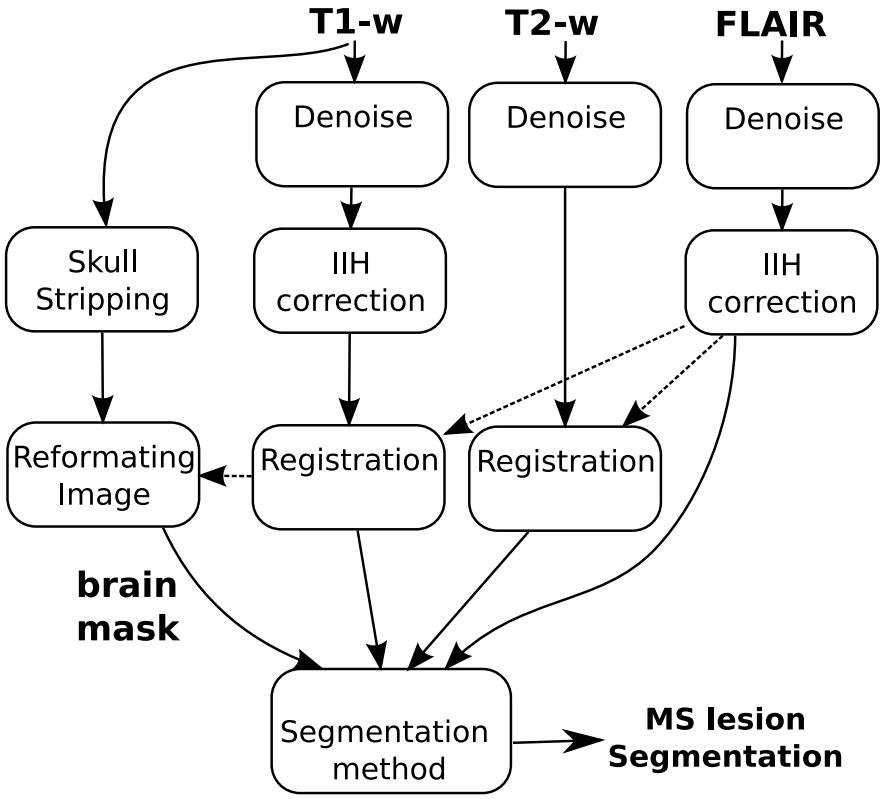


Figure 3.4: Proposed workflow for the automatic segmentation of MS lesions.

Chapter 4

STREM: Spatio-Temporal Robust Expectation Maximization

In the previous chapter, we proposed a segmentation workflow and used a parametric algorithm called STREM [Aït-Ali 2006].

In this chapter, we describe STREM which will serve as a basis for the methods described in the following chapters. STREM is a parametric segmentation method that was developed to segment longitudinal data but in this chapter we present the method only as a one time-point segmentation method because no longitudinal data was available for validation. In addition, we describe some modifications we perform in order to make this method faster, more robust and accurate.

4.1 Method

Our method classifies each voxel of the brain as one of four classes: MS lesions, white matter (WM), gray matter (GM), or cerebrospinal fluid (CSF). We consider a typical MR protocol for MS lesions (T1-w, T2-w and PD) as input of our method. However, other sequences can be added with few modifications, for example FLAIR. Figure 4.1 illustrates the workflow proposed for the segmentation of MS lesions and NABT. MR images go through a preprocessing stage composed of three steps: correction of intensity inhomogeneities, reduction of image noise, and registration of all images in the same space. The T1-w image is used for skull stripping in order to focus the segmentation on the brain voxels. Our segmentation algorithm is composed of three steps: estimation of the NABT model, detection of outliers, and use of heuristic rules to extract the MS lesions from these outliers.

4.1.1 Estimation of the NABT Model

MRI noise usually follows a Rician distribution [Dietrich et al. 2008], which can be approximated by a Gaussian distribution for high SNR [Sijbers et al. 1998]. Following this approximation, the distribution of intensities within each brain structure can be approximated by a Gaussian distribution. We model the image intensities of a healthy brain with a 3-class GMM, where each Gaussian represents one of the

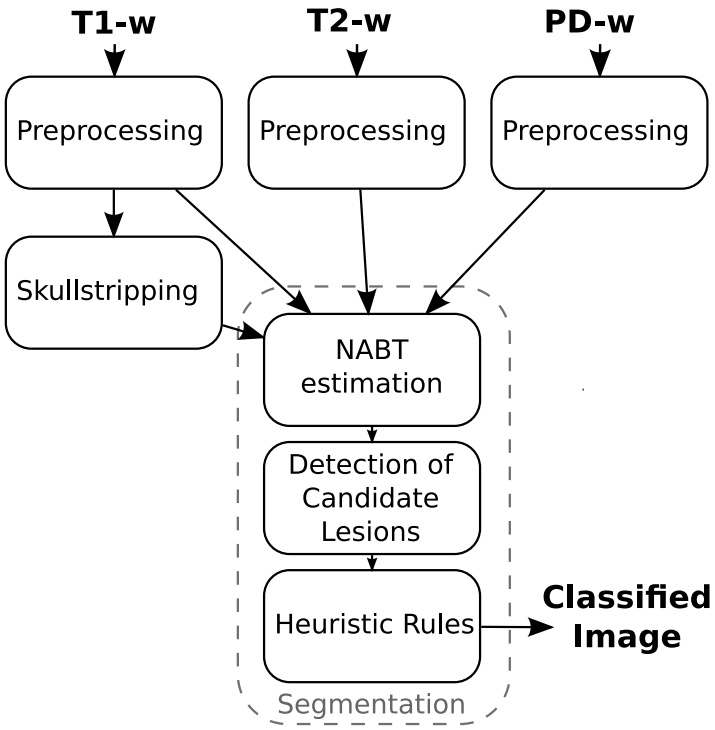


Figure 4.1: Workflow of STREM.

brain tissue WM, GM and CSF. We consider the m MR sequences as a multidimensional image with n voxels. The intensity vector $\mathbf{y_i} = [y_{i_1}...y_{i_m}]$ of the voxel i can be modeled as follows

$$f(\mathbf{y}_i|\theta) = \sum_{j=1}^{3} \alpha_j \cdot N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{4.1}$$

where the mean $\boldsymbol{\mu}_j$ and the covariance matrix $\boldsymbol{\Sigma}_j$ define the parameters of each Gaussian $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. These parameters and the mixing parameter α_j are merged in the parameter vector θ.

These parameters can be estimated using the Maximum Likelihood Estimator (MLE)

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} f(\mathbf{y_i}|\theta) \tag{4.2}$$

if we consider $\mathbf{y_i}$ as *i.i.d.*.

In order to obtain the MLE, we can employ the EM algorithm [Dempster et al. 1977], a technique which is used to iteratively estimate $\hat{\theta}$. From θ_l, the EM algorithm obtains another θ_{l+1} where $L(\theta_{l+1}) \geq L(\theta_l)$.

This method is usually chosen because is easy to compute and always converges to, at least, a local maximum of L, but it has two main drawbacks. The first drawback is that the EM algorithm enables only a local maximum to be reached; different initial parameters θ_0 may enable different solutions to be reached, which makes the choice of θ_0 an important issue. The second drawback is the sensitivity of

MLE to outliers. In statistics, this sensitivity is measured by the Breakdown Point (BP), which can be defined as the smallest number of outliers that can cause the estimator to take arbitrary large values [Müller and Neykov 2003], and, in the case of the MLE, BP is equal to zero.

We propose two solutions to minimize the effect of these two drawbacks: employing a hierarchical initialization in order to increase the chances of reaching the global maximum, and replacing the MLE with the trimmed likelihood estimator TLE [Neykov et al. 2007] which was introduced for the estimation of the NABT model by Aï-Ali [Aït-Ali 2006].

4.1.1.1 Hierarchical Initialization

In general, finding a good initialization is a trade-off between accuracy and computational cost [Meilă and Heckerman 2001]. A general approach uses the EM algorithm with different starting parameters and then selects the solution with maximum L; to gain more chances of reaching the global maximum, more starting parameters are needed which increases the computational time.

The initialization can be given by a probabilistic atlas [Aït-Ali 2006], where each voxel contains the probabilities of belonging to the WM, GM or CSF. The atlas is linearly registered to the patient images and the mean and variance can be computed using the probability of each tissue given by the atlas. The atlas-based initialization method has two drawbacks: the registration is a time-consuming task, and the atlas may provide improper initializations in MS patients displaying considerable brain atrophy or lesion load.

Biernacki et al. [Biernacki et al. 2003] proposed to reduce the computational cost of the above-mentioned random technique with a four-step method. First, they chose multiple starting parameters at random. Second, they ran the EM algorithm with a fixed number of iterations for each starting parameters. Third, they selected the solution providing the best likelihood and fourth, they ran the EM algorithm again until the convergence was reached starting with the solution selected.

We propose a new method to initialize our multi-sequence NABT estimation which includes *a priori* information in order to reduce the computation cost. First, we perform a NABT estimation on the T1-w only, applying the Biernacki et al. initialization, and classify each voxel to one of the three classes according to their highest probability. This NABT model on T1-w provides us with the mean and variance on T1-w of the three tissues. Second, we have to estimate the mean and variance for the other sequences. Using this information, we can therefore compute the histogram for each tissue t (CSF, GM and WM) and sequence s (T1-w, T2-w and PD), and find all local maxima. For each tissue and sequence, we choose one of the local maxima as the mean $\mu_{t,s}$ (see Table 4.1). The objective is to avoid missclassification in the T1-w estimations due to vessels and skull stripping errors. Then the standard deviation is computed using a robust variance estimator

$$\sigma_{s,t}^2 = (1.4918 \cdot med|y_i - \mu_{t,s}|)^2. \tag{4.3}$$

and the final covariance matrix for each tissue t is given by

Sequence	Tissue	Selected Maximum
T2-w/PD	WM	Absolute maximum
T2-w/PD	GM	Absolute maximum
T2-w/PD	CSF	Brightest maximum
FLAIR	WM	Absolute maximum
FLAIR	GM	Absolute maximum
FLAIR	CSF	Absolute maximum

Table 4.1: Table of the selected maximum of the histogram for each tissue t and sequence s. In T2-w and PD the brightest maximum is chosen to avoid vessels and skull stripping voxels missclassified as CSF in T1-w.

$$\begin{pmatrix} \sigma^2_{T1,t} & 0 & 0 \\ 0 & \sigma^2_{T2,t} & 0 \\ 0 & 0 & \sigma^2_{PD,t} \end{pmatrix}. \tag{4.4}$$

4.1.1.2 Trimmed Likelihood

Neykov et al. [Neykov et al. 2007] recently proposed a modification of the MLE in order to make it more robust to outliers. The basic idea consists in maximizing not (4.2) but the trimmed likelihood

$$TL(\theta) = \prod_{i=1}^{k} f(\mathbf{y}_{\nu(i)}|\theta) \tag{4.5}$$

where the trimming parameter k $(n/2 \leq k \leq n)$ determines how many voxels are rejected from the estimation and the permutation function $\nu(i)$ sorts the voxels

$$f(\mathbf{y}_{\nu(1)}|\theta) \geq f(\mathbf{y}_{\nu(2)}|\theta) \geq ... \geq f(\mathbf{y}_{\nu(n)}|\theta). \tag{4.6}$$

We use the parameter h instead of k, where $h = \frac{n-k}{n}$ and $0 \leq h \leq 0.5$. For $h = 0$, the TL is equivalent to MLE. We can prove that this algorithm converges to, at least, a local maximum of the TL and that $BP = h$ [Müller and Neykov 2003], which means that the TLE can obtain a good estimation of the data even if the data are contaminated with $h\%$ of outliers.

4.1.2 Detection of Candidate Lesions

A high value should be chosen for the trimming parameter h in order to ensure all MS lesions voxels and other artifacts are rejected from the estimation of the NABT model. In practice, the h rejected voxels contain inliers that actually fit the NABT model reasonably well. Thus, to refine the detection of outliers, we define distance d_i as the minimal Mahalanobis distance of the voxel i from one of the Gaussians in the NABT model.

$$d_i = \min_{\forall j} \left\{ \sqrt{\left(\mathbf{y}_i - \boldsymbol{\mu}_j\right)^T \Sigma_j^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}_j\right)} \right\}. \tag{4.7}$$

If we consider that voxel intensities of each tissue follow a Gaussian law, the Mahalanobis distance follows a χ^2_m law with m degrees of freedom [Aït-Ali et al. 2005, Dugas-Phocion et al. 2004a], where m is the number of MR sequences. The voxel i is considered a *candidate lesion* when the distance d_i is greater than a threshold that is defined by the χ^2_m law for a given p-value p_{maha}.

4.1.3 Heuristic Rules

Candidate lesions detected with the Mahalanobis distance include MS lesions, vessels, registration errors, flow artifacts, noise, etc. Aït-Ali [Aït-Ali 2006] proposed some intensity rules to discriminate true MS lesion voxels from the other voxels. We included two other rules to reduce the number of false positives: the small lesions rule and the neighbor information rule.

4.1.3.1 Intensity Rules

MS lesions are known to be hyper-intense compared to the WM intensity on T2-w and PD-w and FLAIR sequences.rm We use the information given by the NABT model to define hyper-intensity [Aït-Ali 2006].

A voxel is considered to be hyper-intense for a given sequence (e.g. T2-w) if its intensity y is greater than a threshold y_{th} that is defined by the probability of the Gaussian distribution

$$p_{\text{hyper}} = \int_{y_{\text{th}}}^{\infty} N(\mu^{\text{T2}}_{\text{WM}}, \sigma^{\text{T2}}_{\text{WM}})\, dy. \tag{4.8}$$

If the voxel is not considered hyper-intense on every sequence, it is discarded as a lesion. We focus our method on T2-w hyper-intense lesions, but rules can also be defined for other types of lesions [Aït-Ali et al. 2005].

4.1.3.2 Small Lesions Rule

In order to avoid false positives, candidate lesions smaller than 3 voxels in size are rejected. These small candidate lesions are usually produced by noise or flow artifacts. In clinical practice, lesions must have a radius of 3 mm to be considered as such [Barkhof et al. 1997].

4.1.3.3 Neighbor Information Rule

In MRI, external CSF usually contains artifacts such as fluid flow or partial volumes. These effects can cause voxels in the cortex or external CSF to have similar intensities to MS lesions. In order to reduce the number of false positives due to these effects, we remove all candidate lesions that are not contiguous to WM voxels or that are contiguous to the brain mask border.

4.2 Parameters Setup

In this section, we perform two experiments on the Brainweb images in order to fix the three parameters of our method. The first experiment studies the trimming likelihood estimator for the estimation of the NABT model. The second experiment studies Mahalanobis distance and heuristic rules for the detection of lesions.

4.2.1 Estimation of the NABT Model

We study the behavior of the TLE in the presence of outliers as well as the influence of h. Typical outliers, other than MS lesions, come from errors occuring in the brain extraction step. We create imperfect masks of the brain where skull voxels are included and perform the estimation of the NABT model. We classify each voxel using the NABT model and compute the DSC for each tissue (CSF, GM and WM). In this section, we do not perform lesion detection because we focus on the estimation of the NABT model only.

We employ the images *bw_moderate* (Table 2.1) and compute the imperfect masks dilating the perfect brain mask $mask_{gt}$ from the phantom with spherical structuring elements of different sizes: 1, $mask_{d1}$ (4% of outliers); 2, $mask_{d2}$ (8% of outliers); and 3, $mask_{d3}$ (12% of outliers). We perform the TLE with h varying from 0 (equivalent to MLE) to 0.49 (the limit at which convergence is guaranteed). We assign to each voxel the class of the model with the highest probability and we compute the DSC for each tissue (CSF, GM and WM). The DSC is only computed in the brain voxels and not in those added in the dilating operations.

For MLE ($h = 0$), we observe a decrease in the DSC for all the dilated masks compared to the perfect mask (Figure 4.2); this effect is more pronounced on the CSF. For TLE, we can divide the h range into three ranges: $h <$ outliers, outliers $< h$ and outliers $\gg h$. For $h <$ outliers, the DSC increases when we increase h. Once outliers $< h$, the DSC values are stable and similar to the DSC obtained for the perfect mask. But when outliers $\gg h$, the algorithm considers many voxels to be outliers and the DSC decreases.

The TLE estimation of the NABT model was stable even in the presence of outliers when h is higher than the number of outliers. The MLE ($h = 0$) utilizes all the voxels in the estimation of the NABT model and therefore, in the presence of outliers, the estimation is biased by these outliers making the classification of tissues less accurate. The TLE is able to deal with the outliers when h is higher than the number of outliers although for very high h ($0.35 < h$), many voxels are considered outliers, and this leads to a poor estimation of the model and low DSC scores. In our method, we set $h = 0.20$ for real images, although for Brainweb images where no outliers are present we use $h = 0.05$.

4.2.2 Detection of Lesions

In our method, lesion detection consists of two steps: the detection of candidate lesions and the use of heuristic rules to discriminate the real lesions from the other outliers. The detection of candidate lesions depends on the p_{maha} and the use of

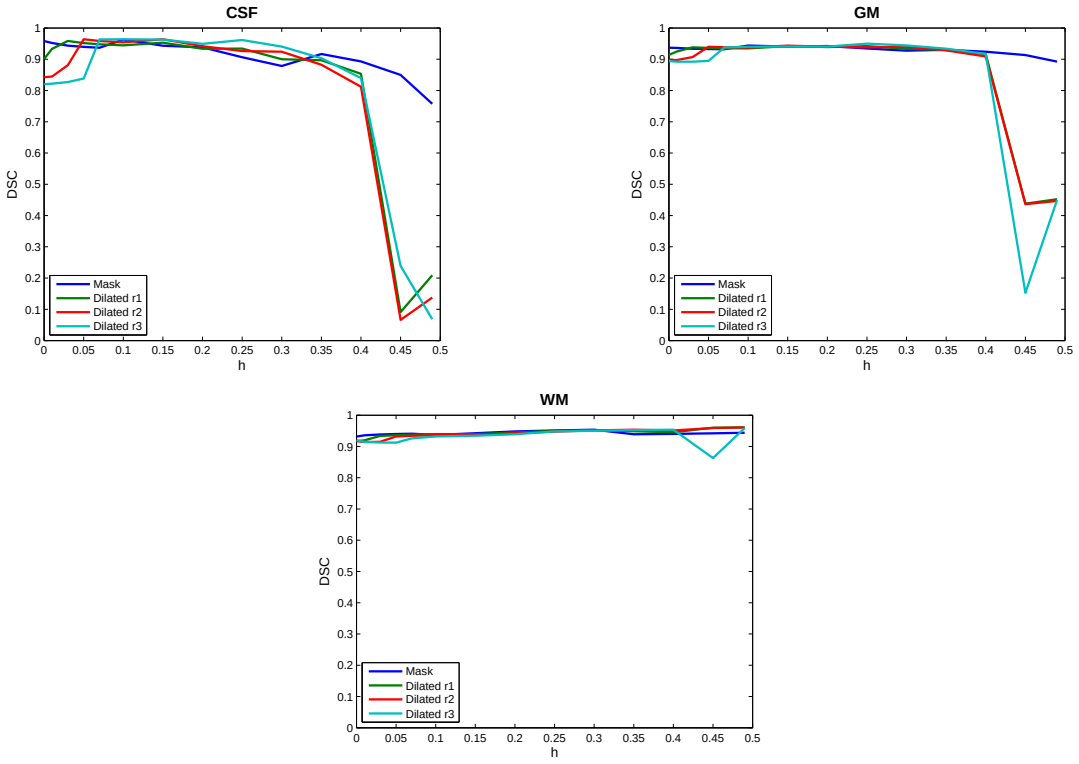


Figure 4.2: DSC for each brain tissue when we vary h on Brainweb images. TLE algorithm shows good stability when we increase the number of outliers compared to the MLE ($h = 0$).

heuristic rules depends on p_{hyper}.

In this experiment, we employ the images from Table 2.1. We perform our segmentation using the ground truth mask and by changing the values of p_{maha} and p_{hyper}. We compare our segmentation with the ground truth using the DSC.

The results are displayed in Figure 4.3. We observe that the best DSC value varies for each lesion load, increasing for the lesion load: mild (> 0.7), moderate (> 0.8) and severe (> 0.85). If we intersect the zones of the graph using best results for each lesion load, we obtain that the best parameters found are $0.3 \leq p_{\text{maha}} \leq 0.35$ and $1 \cdot 10^{-3} \leq p_{\text{maha}} \leq 5 \cdot 10^{-4}$.

The DSC is sensitive to the size of the segmentation, which may explain the lower results of the mild lesion load compared to the severe lesion load [Zijdenbos et al. 1994]. For our algorithm, we choose $p_{\text{maha}} = 0.3$ and $p_{\text{maha}} = 1 \cdot 10^{-3}$.

4.3 Conclusion

We presented a new version of STREM that was already proposed by Aït-Ali [Aït-Ali 2006] where some improvements were made in order to improve the performance of the algorithm. We analyzed the different parameters of the algorithm in order to optimize their values using the Brainweb database.

In the following two chapters, we propose two new segmentation methods that

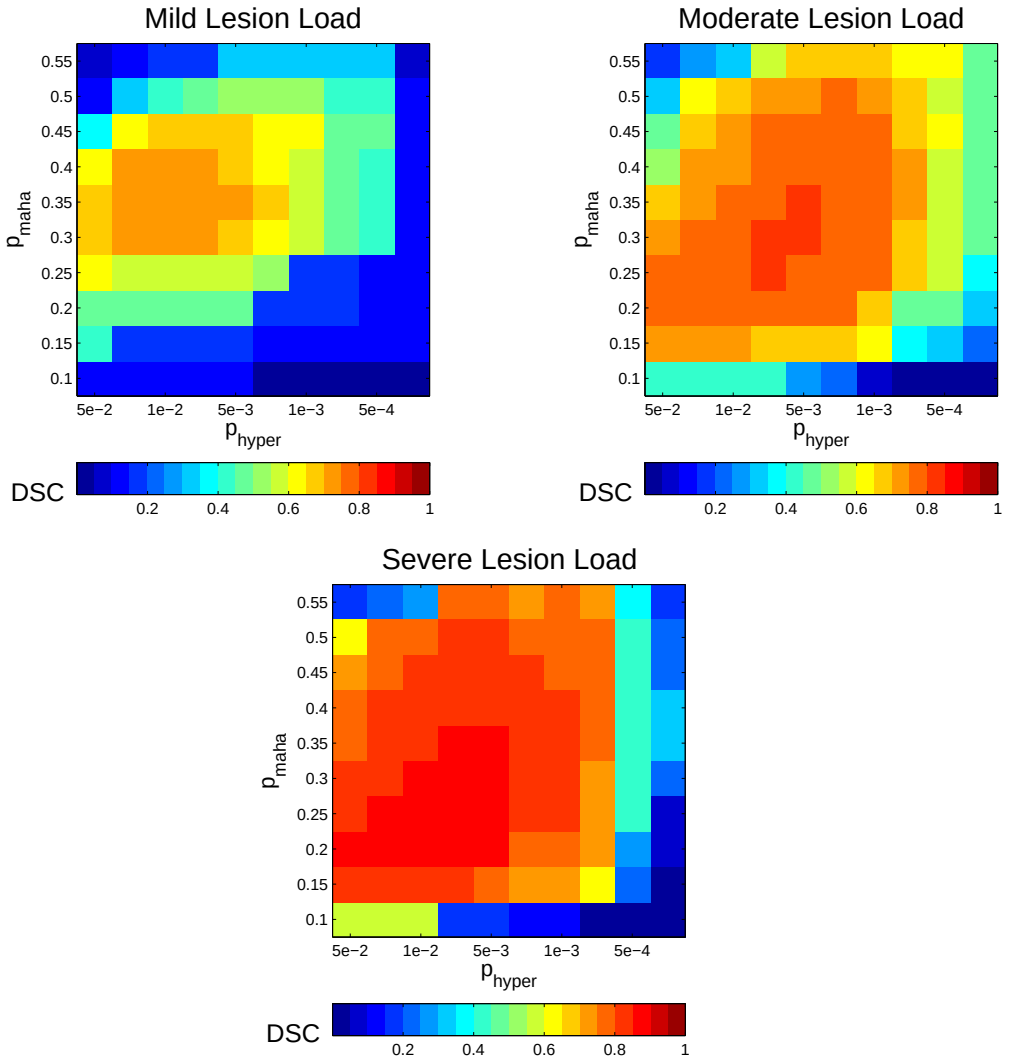


Figure 4.3: DSC values for the automatic segmentation varying the Mahalanobis threshold (p_{maha}) and the hyper-intensity definition (p_{hyper}) on the Brainweb images.

are based on this method but incorporate spatial information to improve the segmentation.

Chapter 5

MS4MS: Mean Shift Clustering for MS Lesion Segmentation

In Chapter 4 we presented STREM, a segmentation method that uses intensity information only in the segmentation of MS lesions.

The intensity of a voxel can be altered by the image noise, which can cause problems in the segmentation. Including spatial information in the segmentation can reduce the effects of noise, providing a more reliable segmentation. Some methods include spatial information in the model estimation using Markov Random Fields [Van Leemput et al. 2001] or a probabilistic atlas [Dugas-Phocion et al. 2004a].

In this chapter, our assumption is that the segmentation of a voxel is more difficult than the segmentation of a region of the image. A *region* is a group of voxels in the same spatial location with similar intensity characteristics. We first divide the brain into regions where each region contains only voxels of the same tissue, and then determine whether each region is an MS lesion or a healthy tissue.

We use the mean shift algorithm (MeS)[1] [Fukunaga and Hostetler 1975] to create these regions in the image. Mean shift is a non-parametric algorithm that uses local spatial information only and can be employed in data clustering, image smoothing and data filtering.

In this chapter, we improve the segmentation method proposed in Chapter 4. We use the mean shift algorithm to create meaningful regions in the brain and then classify these regions using a Gaussian Mixture Model. The chapter starts with a description of the mean shift algorithm. Then the method is described and finally some experiments are done to explain the parameters of our method.

5.1 The Mean Shift Algorithm

The mean shift algorithm is a non-parametric technique for the estimation of the probability density gradient. Fukunaga and Hostetler [Fukunaga and Hostetler 1975] described the mean shift approach and applied it to clustering and data filtering.

[1]Mean shift is usually abbreviated to MS, but in this document MS is reserved for multiple sclerosis.

Cheng [Cheng 1995] generalized the formulation and gave a mathematical description of the mean shift algorithm. Later, Comaniciu and Meer [Comaniciu and Meer 2002] applied mean shift in image processing tasks such as segmentation, video tracking and discontinuity preserving smoothing.

Recently, the mean shift algorithm was employed in the segmentation of brain MRI of healthy subjects [Jimenez-Alaniz et al. 2006, Mayer and Greenspan 2009]. The first method [Jimenez-Alaniz et al. 2006] used an atlas to label the regions given by mean shift in three classes: WM, GM and CSF. It uses a mean shift approach coupled with an edge detector [Meer and Georgescu 2001]. The second method [Mayer and Greenspan 2009] applied an adaptive version of the mean shift clustering [Georgescu et al. 2003], and then performed a k-means clustering on the results in order to classify the three tissues.

Given n data points x_i, $i = 1, .., n$ in the d-dimensional space R^d, the Parzen window density estimator [Duda et al. 2000], with kernel $K(x)$ and bandwidth b, is given by

$$\hat{f}_{b,K}(x) = \frac{1}{nb^n} \sum_{i=1}^{n} K_b(x - x_i). \tag{5.1}$$

We are interested in radially symmetric kernels that verify that

$$K_b(x - x_i) = c \cdot k\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right) \tag{5.2}$$

where the profile $k(x)$ is only defined for $x \geq 0$ and c is a constant [Cheng 1995]. Typical kernels are described in Figure 5.1. Thus we can rewrite equation (5.1) as

$$\hat{f}_{b,k}(x) = \frac{c}{nb^d} \sum_{i=1}^{n} k\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right). \tag{5.3}$$

The gradient of the density can then be estimated as

$$\hat{\nabla} f(x) \equiv \nabla \hat{f}_{b,k} = \frac{2c}{nb^{d+2}} \sum_{i=1}^{n} (x - x_i)\, k'\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right). \tag{5.4}$$

We define $g(x) = -k'(x)$, where $g(x)$ is the profile of the kernel $G(x)$. The kernel $K(x)$ is called the shadow of $G(x)$.

$$\hat{\nabla} f(x) \equiv \nabla \hat{f}_{b,k} = \frac{2c}{nb^{d+2}} \sum_{i=1}^{n} (x_i - x)\, g\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right) = \tag{5.5}$$

$$= \frac{2c}{nb^{d+2}} \left(\sum_{i=1}^{n} g\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right)\right) \left(\frac{\sum_{i=1}^{n} x_i g\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right)}{\sum_{i=1}^{n} g\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right)} - x\right) \tag{5.6}$$

where $\sum_{i=1}^{n} g\left(\left|\left|\frac{x - x_i}{b}\right|\right|^2\right)$ must be positive. The first term is proportional to the Parzen window of the kernel G and the second term is called the mean shift

$$m_{b,g}(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left|\left|\frac{x-x_i}{b}\right|\right|^2\right)}{\sum_{i=1}^{n} g\left(\left|\left|\frac{x-x_i}{b}\right|\right|^2\right)} - x. \tag{5.7}$$

The mean shift vector always points towards the direction of the maximum increase of the density [Comaniciu and Meer 2002].

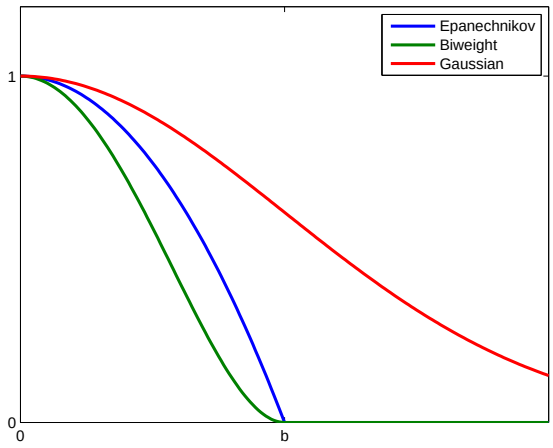


Figure 5.1: Typical profile kernels in the mean shift algorithm

The *mean shift procedure* is performed by the iteration of two steps for each point x_i (Algorithm 1). In every iteration j, the point x_i is moved towards a zone of zero gradient and of local maximum of density called mode that we note $M(x_i)$. The convergence becomes slow when the gradient is near zero; mean shift is stopped when the mean shift vector is smaller than a constant c, which we set as $c = 0.005$.

Algorithm 1 Mean shift procedure

1. While $||m_{b,g}(x_i^{j-1})|| > c$
   2. compute the mean shift vector $m_{b,g}(x_i^j)$, equation (5.7)
   3. move the point $x_i^{j+1} = m_{b,g}(x_i^j) + x_i^j$

Clustering is one of the applications of the mean shift algorithm, whereby points are classified into an undetermined number of clusters (Algorithm 2). Each cluster is characterized by a mode M_l, which is the local maximum of density of the cluster, and c_{mode} defines the minimal distance existing between two modes and is usually equal to half of the bandwidth.

The choice of bandwidth parameter influences the final clustering obtained by the mean shift algorithm. A large bandwidth finds less modes in the image but small clusters might disappear while a small bandwidth might create too many clusters.

Algorithm 2 Mean shift clustering

0. define L, a list containing the modes
1. for $x_i,\ i = 1, ..., N$
   2. compute the mean shift procedure, obtain x_i^{conv}
   3. for M_l in L
      4. compute $|M_l - x_i^{conv}|$
      5. if $|M_l - x_i^{conv}| \leq c_{mode}$
         6. $M(x_i) = M_l$
         7. go to 1.
   8. add mode $M(x_i) = x_i^{conv}$ to L

5.1.1 Mean Shift for Images

There are two different ways to use the mean shift clustering algorithm in image segmentation. The first option is to perform the clustering using only the intensity information. In this case, the point x_i will correspond to the intensities of the voxel i for the different MR sequences. Mean shift clustering is then performed in m dimensions, where m is the number of different MR sequences. In this case, no spatial information of the image is used. The second option consists in adding the three spatial dimensions to the clustering. The point x_i contains m intensities and the three spatial dimensions x, y and z, thus mean shift is performed in $m+3$ dimensions [Comaniciu and Meer 2002]. The advantage of this option is that it enables spatial and the intensity information to be incorporated into the segmentation process. In this case, the clusters formed by the algorithms are usually called *regions*.

The kernel $K(x)$ can be decomposed into two independent kernels with two different bandwidth parameters — one for spatial components b_s and another one for intensity components b_r — as in

$$K_{b_s,b_r}(x) = K\left(\frac{x^s}{b_s}\right) K\left(\frac{x^r}{b_r}\right) \tag{5.8}$$

where x^s contains the three spatial dimensions and x^r contains m intensity components.

Separated kernels were already employed in medical image segmentation [Jimenez-Alaniz et al. 2006] and simplify the concept of bandwidth, as independent bandwidths are chosen for the intensity and spatial features. In the adaptive mean shift [Mayer and Greenspan 2009], the kernel is not computed using a bandwidth parameter but with the k-nearest neighbors. In this method, the relation between spatial and intensity features remains unclear.

5.1.2 Optimization of the Mean Shift Algorithm

One of the limitations of the mean shift algorithm is its high computational cost. The computational cost of the mean shift depends on three main aspects: the kernel computation, the number of iterations of the mean shift procedure and the number of points to process.

Kernel computation: In order to compute the mean shift vector, it is necessary to find all neighboring points in order to compute the distance between the points. In high-dimensional spaces, the search for these points can be computationally expensive. For generic clustering, advanced neighborhood searching techniques have been employed [Georgescu et al. 2003]. In image segmentation, we can take advantage of the regular spatial distribution of voxels to find the neighboring points without the need for these search methods [Carreira-Perpinan 2006].

Number of iterations: When the kernel is Gaussian, mean shift can be seen as a EM algorithm and its convergence is mainly linear [Carreira-Perpinan 2007]. When the density gradient is low, e.g. near the mode, the convergence of the mean shift is very slow. Some authors have proposed accelerating the mean shift procedure to reduce the number of iterations, although these accelerations can reduce the accuracy of the regions found. Exampes of such accelerations include adaptive over-relaxed mean shift [Shen and Brooks 2005], mean shift with quasi-Newton Methods [Yang et al. 2003] and sparse mean shift [Carreira-Perpinan 2006].

Number of points: Other optimizations methods reduce the number of points on which perform the mean shift. One option is spatial discretization [Carreira-Perpinan 2006], the basic idea of which is that points in a similar spatial position converge to the same mode. This method creates a super-resolution grid from the image, and during the mean shift procedure, the grid points are assigned to the same mode as the processed point. When another point is found in those grid points, it is directly assigned to the same mode.

5.1.3 Implementation Details

Our implementation of the mean shift clustering reduces the execution time by two different means. We implement an optimization method called *the basin of attraction* to reduce the number of points to be processed. In addition, we propose a multi-threaded implementation to accelerate the execution on machines with more than one processor.

5.1.3.1 Basin of Attraction

The objective here is to reduce the number of points on which we perform the mean shift procedure. We discarded spatial discretization [Carreira-Perpinan 2006] because the creation of a super-resolution grid will cause memory allocation problems since the images are rather large already. Instead, we propose another implementation of the same idea: points with similar features converge to the same mode. It can be seen as if each point has a basin of attraction where all the points that fall into this basin of attraction converge to the same mode (Figure 5.2).

To explain the basin of attraction, we consider a unique bandwidth parameter b; the extension to two different bandwidths is straightforward. We define a parameter

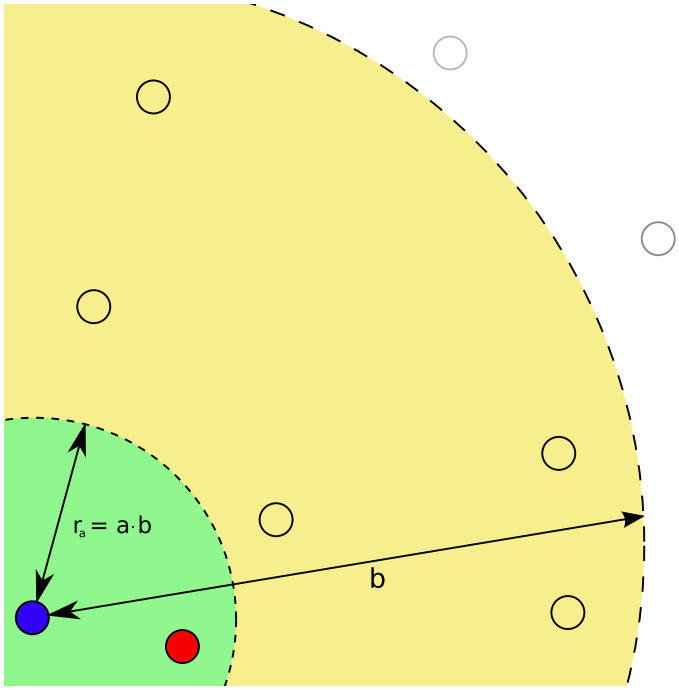


Figure 5.2: Example of the basin of attraction in 2D. We observe an iteration of the mean shift procedure on the blue point. In yellow, the mean shift kernel with bandwidth b. In green, the basin of attraction of radius r_a. In red, a point in the basin of attraction of the blue point.

a ($0 \leq a < 1$) that defines a radius of attraction $r_a = a \cdot b$, and we consider that two points within a radius of attraction converge to the same mode (Algorithm 3).

The advantage of this implementation is that there is no significant increase in the computational cost, because in the computation of the mean shift vector, we already need to compute $\left|\left|x_k - x_i^j\right|\right|$. Then we just verify if this value is smaller than r_a.

5.1.3.2 Multi-threaded Implementation

The use of a multi-threaded implementation does not modify the computational cost of the algorithm, but reduces the execution time in multi-core machines, the use of which is nowadays widespread. A thread is essentially an independent subprogram that runs independently of the main program. The objective is to create multiple threads that work in parallel, thus accelerating the execution. We expect to divide the execution time by almost the number of processors in the machine.

The mean shift procedure can be easily parallelized because each point is computed independently of the result of other points. There are only two aspects that limit the performance of the parallelization: the management of the modes and the basin of attraction.

Once a mode is found using the mean shift clustering, it is stored in a list of found modes. If two threads try to access the mode list at the same time, one of the threads will have to wait, limiting the performance.

Algorithm 3 Mean shift procedure with basin of attraction

1. While $||m_{b,g}(x_i^{j-1})|| > c$
   2. for neighbor x_k
      3. if $\left|\left|x_k - x_i^j\right|\right| < r_a$
         4. if x_k has an associated mode $M(k_x)$
            5. $M(x_i) = M(k_x)$
            6. stop
         7. if x_k has no associated mode
            8. when finished, $M(x_k) = M(k_i)$
   9. compute the mean shift vector $m_{b,g}(x_i^j)$
   10. move the point $x_i^{j+1} = m_{b,g}(x_i^j) + x_i^j$

The basin of attraction requires the result of the mean shift procedure on the neighboring points, which removes the independence of the mean shift clustering. If one thread tries to access a point that is being processed by another thread, it will have to wait for the thread to finish, thus reducing the performance.

To avoid these decreases in performance, we propose to process the image in blocks. We divide the image into rectangular blocks where the number of blocks is greater or equal to the number of threads. Each block is processed independently by just one thread and has its own independent list of modes in order to eliminate interaction between threads. The basin of attraction problem is reduced as only one thread is processing the neighboring points inside the same block thus avoiding conflicts.

Creating a local list of modes for each block may cause two neighboring blocks having a similar mode. This issue is solved in a subsequent step called region fusion (details in Section 5.2.4). The limitation due to the basin of attraction can still occur on the border between two blocks. We propose to process the blocks non-consecutively, thus reducing the chances of two neighboring blocks being processed at the same time.

5.2 Method

As explained previously, the mean shift algorithm creates an undetermined number of regions in the image. In this section, we describe how we use mean shift in order to segment MS lesions.

Figure 5.3 shows the workflow of our method. First, we estimate the NABT model as in Chapter 4. This model serves, on the one hand, to normalize the different sequences, and, on the other hand, to detect candidate lesions.. Second, we apply the mean shift segmentation followed by a region fusion procedure to obtain meaningful regions of the brain. Finally, we apply heuristic rules to discriminate the lesions from other tissues in order to obtain the final segmentation.

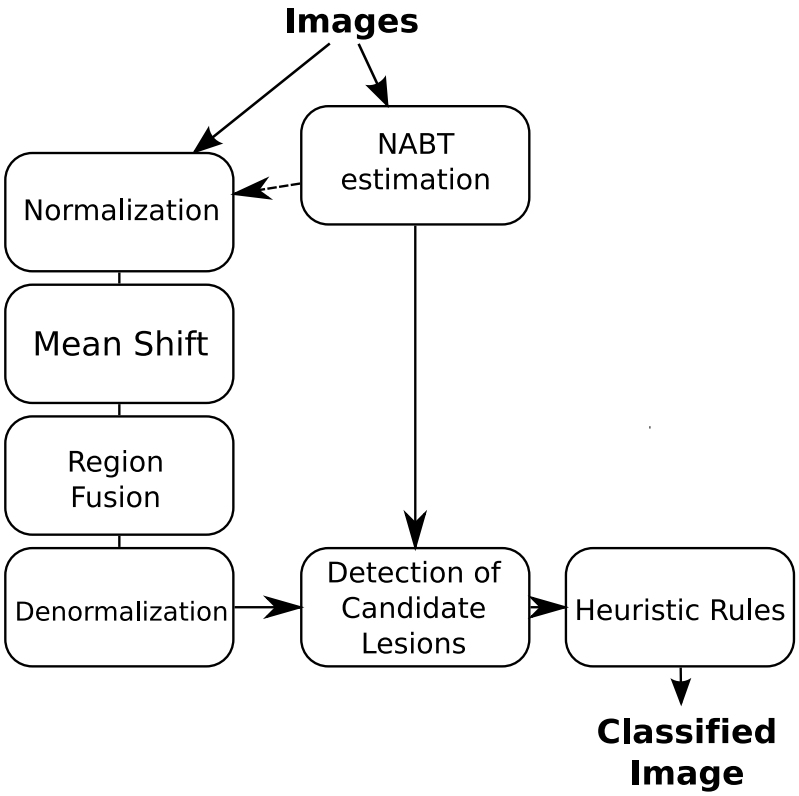


Figure 5.3: Workflow of MS4MS.

5.2.1 Estimation of the NABT Model

As explained in Section 4.1.1, we consider the intensities of the brain to follow a 3-class GMM, where each Gaussian represents a brain tissue: GM, WM or CSF. This model excludes the MS lesions that are detected afterwards as outliers to this model. We employ the TLE [Neykov et al. 2007] to compute the model parameters because the TLE is more robust to outliers than the traditional MLE.

5.2.2 Sequence Normalization

Each MR sequence has a different intensity range and therefore the bandwidth parameter b_r may be inefficient. In (5.8), the kernel can be decomposed again to have one b_r parameter per sequence, but it will increase the processing time of the algorithm.

We normalize all the images so that the variance of the standard deviation of WM is equal to 100 using the previously estimated NABT model. This normalization enables us to have a common bandwidth parameter for all the sequences, and to use the same bandwidth parameter for different MR protocols.

5.2.3 Mean Shift Segmentation

As described earlier, the mean shift algorithm creates an undetermined number of output regions using both spatial and intensity information. Our assumption is that

each region belongs to one tissue class only and that classifying one region is easier than classifying each voxel individually.

5.2.4 Region Fusion

To reduce the number of regions given by the mean shift algorithm, we employ a region fusion method. In homogeneous regions, the density gradient is near zero and the mean shift vector is very small, which may cause oversegmentation. Essentially, region fusion methods merge neighboring regions if their intensities are similar.

Two regions are usually merged if the distance from one another is lower than a fixed threshold α, which can be called α-connected [Soille 2008]. This method can have a *chaining effect* when the intensity increases slowly. The reason for this is that an α-connected relation is not transitive; when a region a is α-connected with regions b and c, b and c might not be α-connected. To resolve this issue, we employ another definition of connectivity known as (α, ω)-connected [Soille 2008], which is transitive. Basically, the distance between neighbor regions must be lower than α and the maximum distance between all connected regions regions must be lower than ω. Both parameters are set to half of the feature bandwidth.

5.2.5 Denormalization

After performing mean shift segmentation, we denormalize to recover the original intensities of the brain for each region. This step is necessary for classifying each region using the NABT model.

5.2.6 Detection of Candidate Lesions

In the same way as in Section 4.1.2, we compute the Mahalanobis distance between the mode of each region in the image and the previously estimated NABT model. Considering that voxels intensities in each NABT follows a Gaussian law, the Mahalanobis distances follow a χ^2 law with m degrees of freedom [Aït-Ali et al. 2005, Dugas-Phocion et al. 2004a], where m is the number of MR sequences. Each region in the image is considered to be a candidate lesion when the Mahalanobis distance for every class is greater than a threshold that is defined by the χ^2_m law for a given p-value p_{maha}.

5.2.7 Heuristic Rules

The outliers detected by the previous step come from different sources: MS lesions, vessels and skull-stripping. In order to reduce the number of false positives, we use the same rules as in Section 4.1.3.

5.3 Parameter Setup

As in Chapter 4, we describe the behavior of our method when we vary its parameters. We first describe a new measure created to study the mean shift segmentation.

Then we study three aspects of this method: the mean shift bandwidth, the optimization of the mean shift implementation, and the MS rules.

5.3.1 Optimistic Overlap

In order to select the parameters of the mean shift algorithm, a measure to evaluate the accuracy of the mean shift regions is required. The only requirement for a mean shift region is that all the voxels in the region belong to the same tissue class; regions with voxels belonging to more than one class are considered to be wrong.

We define the Optimistic Overlap (OO) as the best possible overlap between the regions and the ground truth. We call this overlap measure as "optimistic" because each region is assigned to the class that gives the best score without the use of any classification method. The advantage of this technique is that we do not evaluate the performance of the classification algorithm and we focus exclusively on the specificity of the mean shift regions.

We consider image I with n voxels, the mean shift regions $r_i, i = 1...n$ and the ground truth with K classes. For each r_i, we count the number of voxels that belong to each class k, $v_i^1, ..., v_i^K$. We write the optimistic overlap as

$$OO = \frac{\sum_{i=1}^{n} \max_{k=1}^{K}(v_i^k)}{n}. \tag{5.9}$$

The values range from 1 (perfect) to $1/K$. This measure can be seen as the best result we could obtain if our subsequent classification is perfect. Taking into account that the volume of MS lesions is very small compared to the volume of the brain, the value of OO should be very high in order to ensure accurate segmentation.

5.3.2 Mean Shift Bandwidth

Our objective is to obtain a reduced number of regions, but each region must gather voxels from the same class only. Jimenez-Alaniz et al. proposed a spatial bandwidth of 6 mm and a feature bandwidth of 9 for the segmentation of healthy brains on T1-w images [Jimenez-Alaniz et al. 2006]. The spatial bandwidth depends on the size of objects we wish to segment. In our case, we consider that the spatial bandwidth proposed for healthy patients is adequate for the segmentation of MS patients and therefore we set the same spatial bandwidth, $b_s = 6\ mm$.

The feature bandwidth should take into account the contrast and intensity range of the image. The feature bandwidth parameter depends on the way in which the images are normalized. Jimenez-Alaniz et al. omitted the normalization process they used. As described earlier in Section 5.2.2, we normalize the standard deviation of the white matter to 100.

To study the feature bandwidth, we perform the mean shift on the images *bw_moderate* (Table 2.1). We want to measure two different aspects: the reduction and accuracy of the mean shift regions. The reduction is evaluated using the ratio between the number of modes found and the number of voxels in the brain. To assess the accuracy, we employ the optimistic overlap (OO) described earlier.

We perform the mean shift segmentation using different feature bandwidths ranging from 50 to 300. No optimization is used in the mean shift.

Figure 5.4 shows the results of the experiment. We observe that there is a small reduction in the number of modes when the feature bandwidth is 100 or lower. The number of modes is reduced to less than 10% when $b_r \leq 125$. The OO is around 100% for $b_r < 125$ and decreases for higher bandwidth values.

We observe that b_r has to be higher than the normalization vale in order to reduce the number of modes significantly. We notice that the values of OO are higher than 98% for all parameters, which means that less than 2% of the voxels of the brain were incorrectly classified. MS lesions occupy only 0.2% of the volume of the brain; small variations of the OO can modify our segmentation significantly (Figure 5.5). In order to obtain a reduced number of regions yet minimize overlap errors, we set our bandwidth to 125.

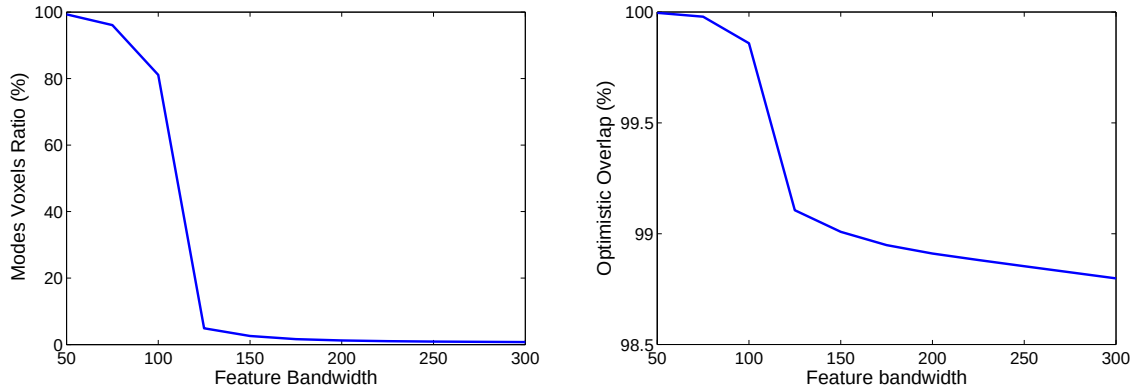


Figure 5.4: Results on *bw_moderate* with varying feature bandwidths. Left: Ratio between the number of modes and voxels. There are fewer regions when the feature bandwidth is high. Right: Optimistic Overlap. There are more overlap errors when the feature bandwidth is high.

5.3.3 Implementation Optimization

In this section, we evaluate the efficiency of the optimization proposed for the mean shift and select the best parameters for this optimization. In the first experiment, we study the efficiency of the basin of attraction and in the second one we focus on the execution time.

5.3.3.1 Basin of Attraction

As explained earlier, the basin of attraction has one parameter a that determines which voxels are close enough together to be attracted by each other. If a is large, this optimization can attract voxels that should belong to different regions, causing segmentation errors.

We perform the mean shift segmentation on *bw_moderate* (Table 2.1) with a feature bandwidth of 125 and a spatial bandwidth of 6 mm, and with the parameter

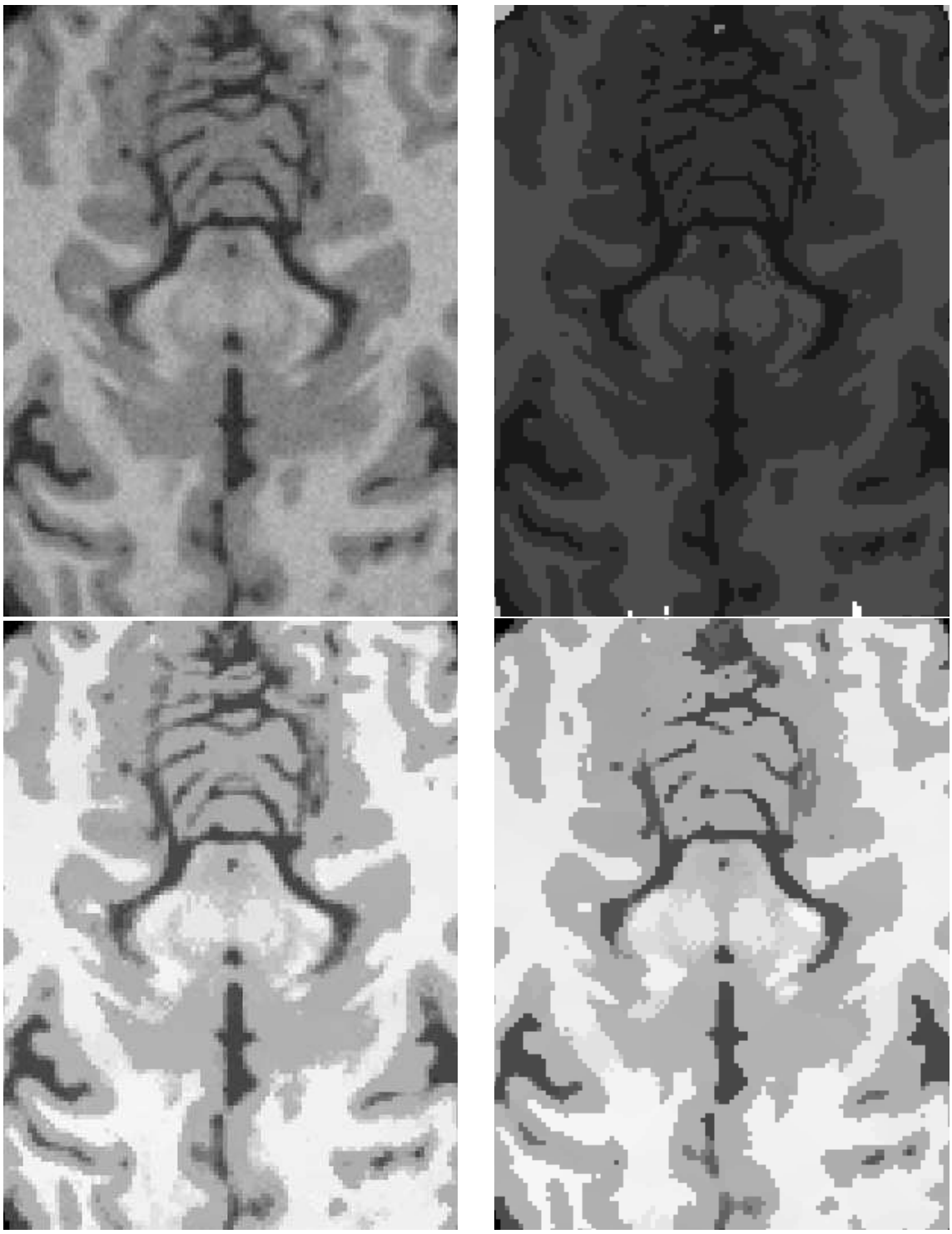


Figure 5.5: From tom to bottom, from left to right: Brainweb T1-w image, ground truth, and mean shift results with bandwidths 125 and 300. When the bandwidth is equal to 300, regions seems more homogeneous but some CSF regions were merged with GM leading to problems in the segmentation.

a varying from 0 to 0.9. We compute the optimistic overlap and count the number of voxels that are processed using our basin of attraction.

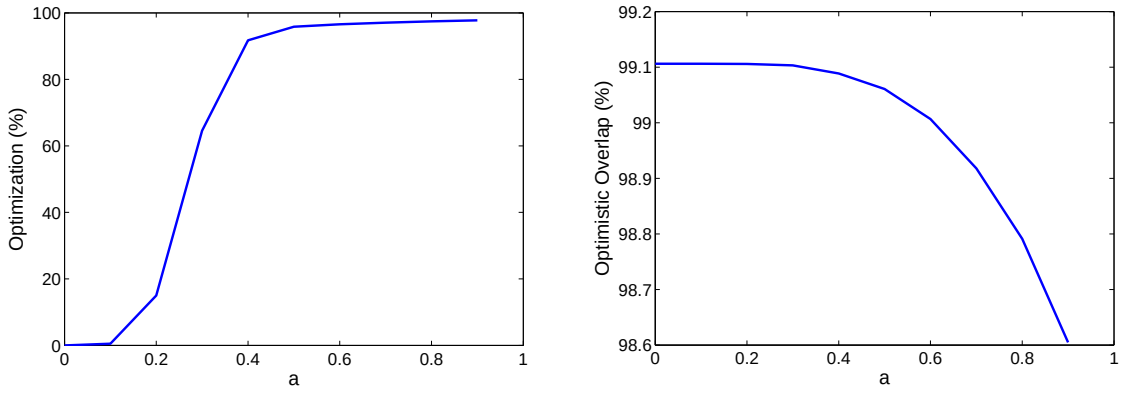


Figure 5.6: Results on *bw_moderate* with varying parameter a of the basin of attraction. Left: Percentage of voxels processed by the basin of attraction, which is a measure of the efficiency of the optimization. Right: Optimistic Overlap.

The results are shown in Figure 5.6. On the left, we observe that when a increases, the number of voxels that are attracted also increases. When $a \geq 0.3$, the number of attracted voxels is higher than 50%. On the right, the optimistic overlap is almost constant for values of $a < 0.4$. When $a > 0.4$, the optimistic overlap decreases rapidly.

The objective of the basin of attraction is to reduce the number of operations of the mean shift yet yielding a similar result to the mean shift without optimization. For our method, we choose $a = 0.3$ in order to obtain the best possible optimization without reducing the quality of the mean shift segmentation.

5.3.3.2 Execution Time

We measure the execution time of the mean shift algorithm using the multi-threaded version of the algorithm and the basin of attraction. We perform the mean shift segmentation on *bw_moderate* (Table 2.1) on a machine with four cores. We vary the number of threads from 1 to 8 with and without the basin of attraction and measure the execution time.

The results are shown in Table 5.7. We can observe that the use of the basin of attraction reduces the execution time to 14% of the original time showing a significant improvement. As for the multi-threaded implementation, we can observe that the reduction is close to optimal with and without the basin of attraction. When the number of threads is greater than or equal to the number of cores, the reduction is almost 25% compared to the non-threaded version.

The basin of attraction gives the best time reduction. In addition, multi-core machines are nowadays very common and the multi-threaded version improves the execution time depending on the number of cores in the machine.

	Basin of Attraction			
Threads	Without		With	
1	27082	(100.0%)	3865	(14.0%)
2	12701	(46.9%)	1965	(7.3%)
4	6843	(25.3%)	1104	(4.1%)
8	6875	(25.4%)	1130	(4.0%)

Figure 5.7: Execution time of the mean shift algorithm (in seconds) when we vary the number of threads with and without the basin of attraction. Experiments were run once only.

5.3.4 Rules

We perform a similar study to that conducted in Section 4.2.2. We perform the MS lesion segmentation on *bw_mild*, *bw_moderate* and *bw_severe* (Table 2.1) , varying the p_{maha} of the Mahalanobis threshold and the p_{hyper} of the hyper-intensity definition. The automatic segmentation is compared with the ground truth using DSC.

Figure 5.8 shows DSC for the three lesion loads. For *bw_mild*, the best DSC values are found around $p_{maha} = 0.4$ and $p_{hyper} = 5 \cdot 10^{-3}$. The best results for *bw_moderate* are found around $p_{maha} = 0.35$ and $p_{hyper} = 5 \cdot 10^{-4}$. For *bw_severe*, the best DSC values are obtained for $p_{maha} = 0.3$ and $p_{hyper} = 1 \cdot 10^{-4}$.

We observe that for the three lesion loads, there is a flat zone where the algorithm is stable to small variations in the parameters. For moderate and severe lesion loads, the algorithm is robust to variations in p_{maha}, and obtain high DSC scores for almost the whole range of values once p_{hyper} is correctly chosen. We set the parameters of the algorithm as $p_{maha} = 0.35$ and $p_{hyper} = 1 \cdot 10^{-3}$.

If we compare these results with the ones obtained in the last chapter (Figure 4.3), we observe that the selection of parameters is less critical because the range of parameters leading to good results is wider for this new method. Creating mean shift regions simplifies the subsequent classification and small variations in the parameters leave the results unchanged.

5.3.5 Conclusion

In this chapter, we have presented a new method for the segmentation of MS lesions which integrates the use of the mean shift algorithm into our previous method. This method is based on the assumption that a region of similar voxels is easier to classify than each voxel independently. Mean shift creates these regions using local spatial information that are subsequently classified using our model estimation. The experiments we conducted showed that this method is less sensitive to the choice of parameters than STREM.

The main drawback of mean shift is the computational cost involved. We proposed a multi-threaded implementation and the basin of attraction in order to reduce considerably the execution time without losing accuracy.

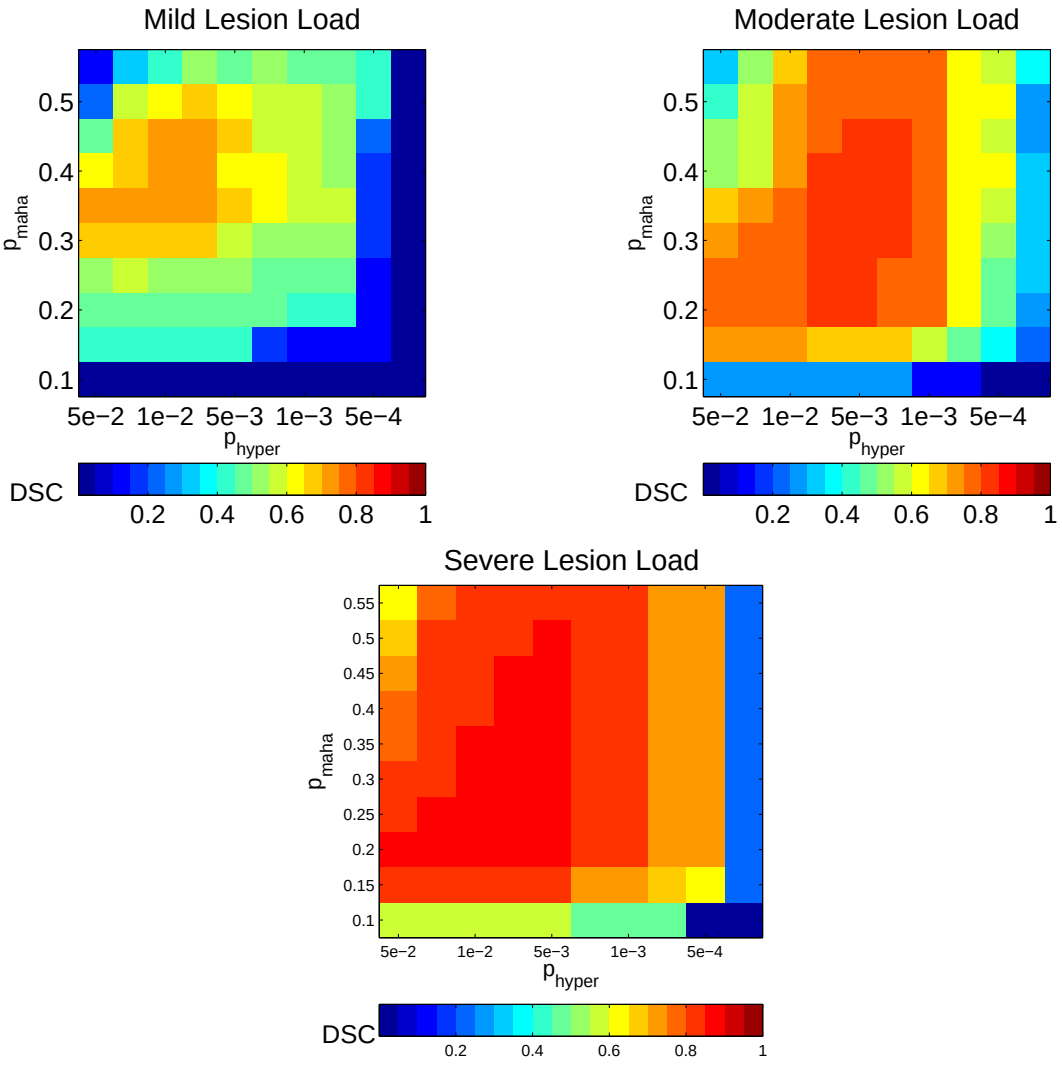


Figure 5.8: DSC values for automatic segmentation, varying the Mahalanobis threshold (p_{maha}) and the hyper-intensity definition (p_{hyper}) on the Brainweb images..

Chapter 6

GCEM: A Graph Cut Approach With an Expectation-Maximization Initialization

In Chapter 5, we described a segmentation method that combined both a global and a local algorithms so as to include spatial information in the segmentation of MS lesions using mean shift.

In many situations, clinicians would like to refine a particular segmentation given by an automatic segmentation method. Graph cut (GC) [Boykov and Funka-Lea 2006] is a recently developed technique for semi-automatic segmentation which has been successfully employed in different medical areas. It is based on both regional and contour information, is very fast to compute and the result can easily be refined interactively.

In this chapter, we explore the possibility of improving our automatic MS lesion segmentation using a graph cut framework. We propose to automate the graph cut algorithm so as to segment MS lesions using several MR sequences. The initialization for graph cut is given by a Gaussian Mixture Model (GMM) estimated by a robust version of the Expectation-Maximization (EM) algorithm [Neykov et al. 2007]. The advantage of this method over previous methods is the possibility for an expert to easily refine the segmentation semi-automatically thanks to the graph cut framework.

This work was carried out in collaboration with Jeremy Lecoeur, who worked in the semi-automatic segmentation of brain images with graph cut. The objective is to automatize the spectral graph cut [Lecoeur et al. 2008] for the segmentation of MS lesions.

6.1 The Graph Cut Framework

The rapid computation of the graph cut has opened up the possibility of using a new family of graph-based semi-automatic segmentation methods for medical images. Boykov and Jolly [Boykov and Jolly 2000] proposed the use of graph cut for the segmentation of cardiac MRI and CT. The user only needed to select some seeds from the object and the background to perform the segmentation. The advantage is that the segmentation can be refined iteratively in a fast and simple way. Esneault

et al. [Esneault et al. 2007] used graph cut to segment liver tumors on CT images. Recently, the graph cut was also applied to brain MRI. Song et al. [Song et al. 2006] employed graph cut with atlas priors for the segmentation of brain tissues in neonatal brains and Lecoeur et al. [Lecoeur et al. 2008] semi-automatically segmented tumors and MS lesions in multi-sequence brain MRI.

Figure 6.1 describes the graph cut algorithm in an intuitive way. The segmentation problem can be described by a flow graph $\mathcal{G} = \langle \mathcal{P}, \mathcal{N} \rangle$ which represents the image [Boykov and Jolly 2001]. In the graph $\mathcal{G}$, each voxel of the image corresponds to a node p in $\mathcal{P}$. The node set $\mathcal{P}$ also contains two particular nodes called terminal nodes - also known as *source* and *sink* - which represent the classes *object* and *background* respectively. Two neighbor nodes are connected with undirected links and $\mathcal{N}$ contains all the links of the image. All the nodes are connected to both object and background nodes. The segmentation is represented by the vector V, where the value of the binary variable V_p can be either object or background.

We define an energy function $E(\mathrm{V})$ that includes spatial and intensity information as

$$E(\mathrm{V}) = \alpha \cdot \sum_{p \in \mathcal{P}} R_p(V_p) + \sum_{\substack{\{p,q\} \in \mathcal{N} \\ V_p \neq V_q}} B_{\{p,q\}}. \tag{6.1}$$

The regional term $R_p(\cdot)$ expresses how the voxel p fits the given models of the object and background. In the graph, this relation is expressed by the connection of all the nodes to the object and background nodes, called t-links.

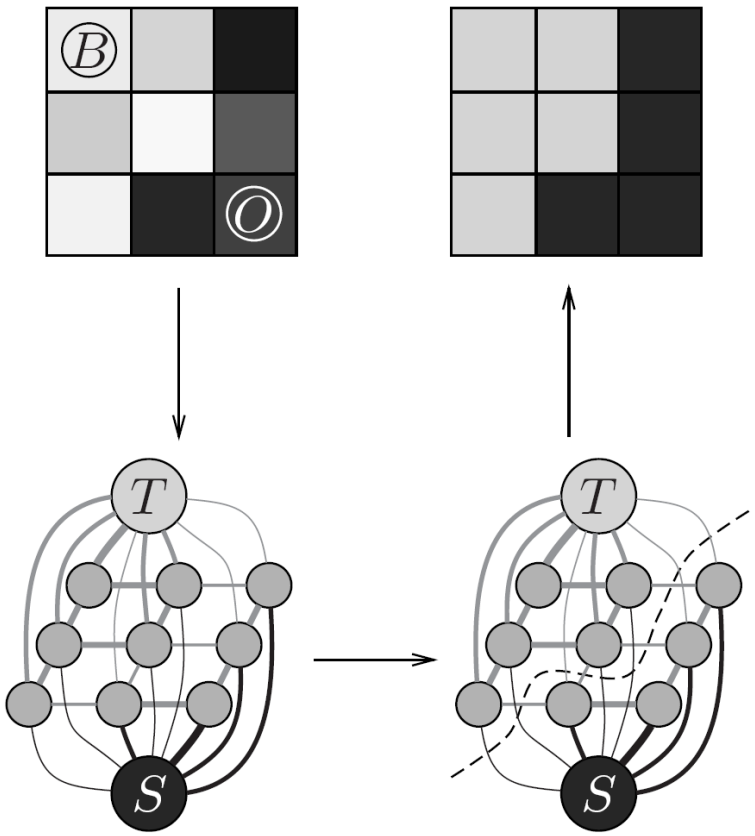
The boundary term $B_{\{p,q\}}$ reflects the similarity between the voxels p and q. Neighboring nodes are connected in the graph, n-links, with weight $B_{\{p,q\}}$. Their values are close to zero when the existence of a contour between p and q is very likely and are large otherwise. The coefficient α is used to adjust the importance of the region and boundary terms.

When the energy function (6.1) is minimized, the resulting vector V gives the result of the segmentation. Greig et al. [Greig et al. 1989] proposed an efficient method to compute the global minima using the max-flow/min-cut algorithm. The basic idea of this algorithm is that the cut with the minimal cost can be found using the maximum flow from the object to the background. Recently, Boykov and Kolmogorov [Boykov and Kolmogorov 2004] proposed an optimization of the algorithm allowing faster computation of the graph cut. For more information, the reader can refer to the work by Lecoeur [Lecoeur 2010].

6.2 Method

As explained before, graph cut requires two different sources of information: the regional term and the boundary term. In the semi-automatic approach, the regional term is given by the seeds chosen by the user. In our method, we avoid the user interaction and compute the regional term automatically.

Figure 6.2 illustrates the workflow employed in this method. The boundary term is computed using the spectral gradient proposed for the segmentation of multi-sequence MRI [Lecoeur 2010]. The regional term is computed in three steps: the



1. In the image, the user chooses seeds from the *background* (B) and the target *object* (O).

2. The image is transformed into a graph; each pixel is represented by a node in the graph. Neighbor voxels are connected by a link. In addition, two other nodes corresponding to the object (S) and the background (T) are created, which are connected to all the pixels in the images.

3. Each link between two nodes has an associated weight that represents the similarity between the two nodes (this is represented by the width of each connection). The algorithm "cuts" the weakest links in order to split the graph into two sub-graphs.

4. Once the cut is made, the graph is converted back into the result image.

Figure 6.1: Example of a graph cut segmentation. (Images extracted from [Lecoeur 2010])

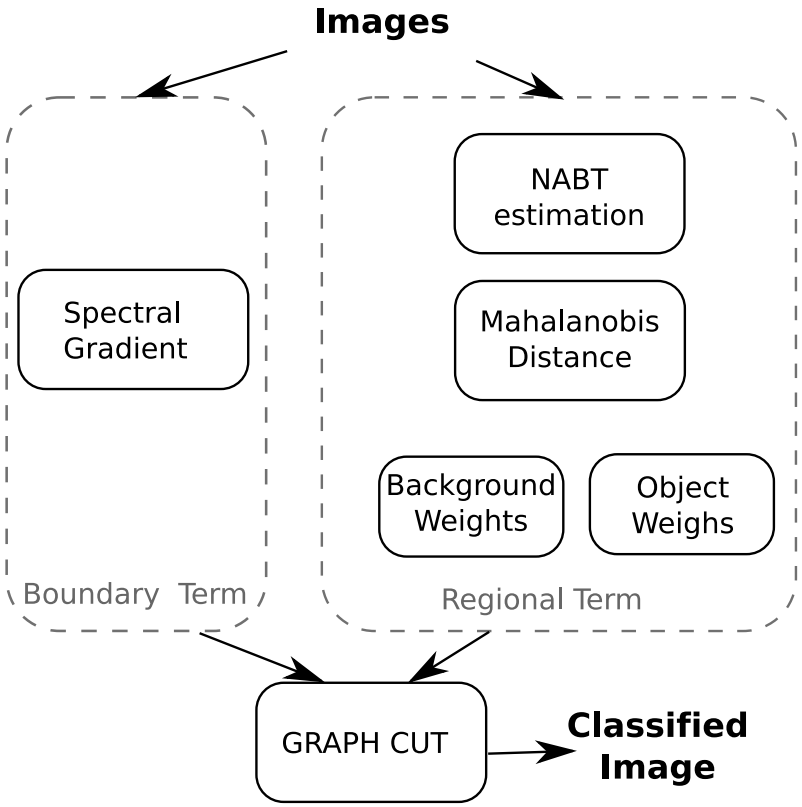


Figure 6.2: Workflow of the proposed automatic segmentation method based on graph cut.

estimation of the NABT model (as described in Chapter 4), the Mahalanobis distance and the estimation of the object and background weights for the regional term.

In the following sections, we consider the MR sequences as a unique multidimensional image of dimension m equal to the number of sequences. We assume that all the MR sequences of the same patient have been previously registered in the same space, that intensity inhomogeneity correction has been performed and that the brain has been extracted.

6.2.1 Boundary Term

The boundary weights $B_{\{p,q\}}$ are usually computed by an edge detection technique such as the local intensity gradient [Boykov and Kolmogorov 2004] or the Lorentzian error norm [Song et al. 2006]. The majority of these techniques are performed on one sequence only although recently some techniques have been proposed for the use of graph cut in the tensor space [Weldeselassie and Harmaneh 2007]. As the boundary term, we choose the spectral gradient that was already employed for the semi-automatic segmentation of brain multi-sequence MRI [Lecoeur et al. 2008].

The objective is to consider our three MR sequences as a RGB color image and use an invariant color-edge detector [Lecoeur et al. 2008, Lecoeur 2010]. This detector is based on a physical property of color, the spectral intensity e, and its

derivatives with respect to the light wavelength λ, e_λ and $e_{\lambda\lambda}$, that are simple to compute [Geusebroek et al. 2000].

For the graph, the detector is discretized, giving the following boundary term [Lecoeur et al. 2008]:

$$B_{\{p,q\}} = \exp\left(-\frac{(\varepsilon(p)-\varepsilon(q))^2 + (\varepsilon_\lambda(p)-\varepsilon_\lambda(q))^2}{2\sigma^2}\right)\cdot\frac{1}{dist(p,q)} \tag{6.2}$$

where σ is a smoothing parameter (in our experiments $\sigma = 1$), and

$$\varepsilon = \frac{1}{e}\cdot\frac{\partial e}{\partial\lambda} = \frac{e_\lambda}{e} \quad \text{and} \quad \varepsilon_\lambda = \frac{\partial\varepsilon}{\partial\lambda} = \frac{e\cdot e_{\lambda\lambda} - e_\lambda^2}{e^2}. \tag{6.3}$$

6.2.2 Regional Term

In semi-automatic frameworks, the weights $R_p(\mathcal{B})$ and $R_p(\mathcal{O})$ for the voxel p are usually defined as follows

$$R_p(\mathcal{B}) = \begin{cases} \infty & \text{if } p \in \mathcal{B} \\ 0 & \text{if } p \in \mathcal{O} \\ -\ln P(I_p|\mathcal{O}) & \text{elsewhere} \end{cases} \quad R_p(\mathcal{O}) = \begin{cases} 0 & \text{if } p \in \mathcal{B} \\ \infty & \text{if } p \in \mathcal{O} \\ -\ln P(I_p|\mathcal{B}) & \text{elsewhere} \end{cases} \tag{6.4}$$

The point sets $\mathcal{B}$ and $\mathcal{O}$ are the seeds of the background and the object respectively. The probability $P(I_p|\mathcal{B})$ reflects how the intensity vector I_p of voxel p fits into the intensity model estimated using $\mathcal{B}$. Most authors assume that seeds follow a Gaussian distribution [Song et al. 2006, Lecoeur et al. 2008, Boykov and Funka-Lea 2006].

Our objective is to eliminate dependency on the seeds in order to create a fully automated method. We replace $P(I_p|\mathcal{B})$ and $P(I_p|\mathcal{O})$ in (6.4) for some weights W_B and W_O in order to have a more general definition, $W_O, W_B \in [0,1]$. For example, W_O should be close to 0 when the voxels are very likely to be a MS lesion and close to 1 otherwise.

$$R_p(\mathcal{B}) = \begin{cases} \infty & \text{if } p \in \mathcal{B} \\ 0 & \text{if } p \in \mathcal{O} \\ -\ln W_B & \text{elsewhere} \end{cases} \quad R_p(\mathcal{O}) = \begin{cases} 0 & \text{if } p \in \mathcal{B} \\ \infty & \text{if } p \in \mathcal{O} \\ -\ln W_O & \text{elsewhere} \end{cases} \tag{6.5}$$

In order to enable the subsequent semi-automatic processing by a user, the point sets $\mathcal{B}$ and $\mathcal{O}$ are maintained in (6.5) althought, in the case of automatic segmentation, weights W_O and W_B are used only .

In order to determine these weights, we follow a similar approach to the one in Chapter 4. First, we consider the NABT intensities to follow a 3-class Gaussian model and we compute the model parameters using the trimmed likelihood estimator [Neykov et al. 2007]. Second, we compute the distance of each voxel from the

model using the Mahalanobis distance and compute the p_{maha} for each voxel (see Section 4.1.1). Voxels with low p_{maha} are more likely to be outliers. In the graph cut framework, weights can have any value from 0 to 1 thus we do not apply any thresholding regarding p_{maha}, contrary to the methods in Chapters 4 and 5.

For each type of weight, W_B and W_O, we define specific rules so as to include more information in the graph cut framework:

Background weights (W_B): Voxels that follow the NABT model should have a weight close to zero. We define these weights using the p_{maha} mentioned earlier as:

$$W_B = 1 - p_{\text{maha}}. \tag{6.6}$$

Object weights (W_O): Outliers, voxels with low p_{maha}, are usually MS lesions, but they can also occur due to veins or registration and skull-stripping errors. This is why we need more information to precisely determine the weights. To select MS lesions from other outliers, we apply *a priori* knowledge about the intensity of the lesions.

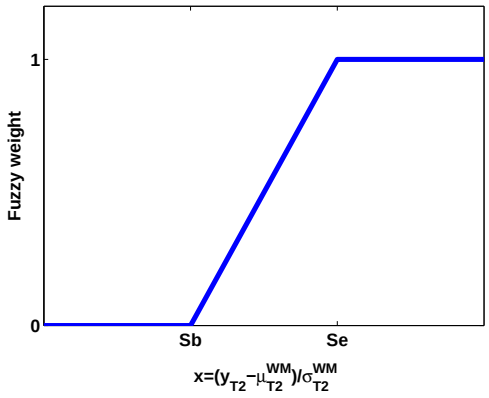


Figure 6.3: Fuzzy definition of *hyper-intensity*. When the intensity $x < S_b$, we consider the voxel as not being not hyper-intense. Between S_b and S_e, the function varies between 0 and 1. Finally when $x > S_e$, the voxel is considered as being completely hyper-intense.

As we mentioned before, MS lesions are usually described as *hyper-intense* compared to the WM in T2-w, PD-w or FLAIR images. In Chapter 4, we described the hyper-intensity using hard thresholds because we needed to obtain a binary MS lesion segmentation. In this case, the weights can have any value between 0 and 1, thus we choose a fuzzy approach to model this experts' knowledge. Instead of defining a binary threshold for hyper-intensity, we define a fuzzy weight. This fuzzy weight is computed for each sequence and needs two parameters to define the fuzzy function: slope beginning S_b and slope end S_e (Figure 6.3).

As an example, for T2-w images, considering the intensity of the T2-w image y_{T2}, we can normalize the intensity according to the WM mean $\mu_{\text{T2}}^{\text{WM}}$ and standard deviation $\sigma_{\text{T2}}^{\text{WM}}$, as follows

$$x = \frac{y_{\text{T2}} - \mu_{\text{T2}}^{\text{WM}}}{\sigma_{\text{T2}}^{\text{WM}}} \tag{6.7}$$

the equation of the fuzzy weight then becomes

$$W_{\text{T2}} = \begin{cases} 0 & \text{if } x < S_b \\ \frac{x - S_b}{S_e - S_b} & \text{if } S_b \geq x \geq S_e \\ 1 & \text{if } x > S_e \end{cases} \tag{6.8}$$

The information given by the Mahalanobis distance and the hyper-intensity fuzzy rules is merged into one weight using the fuzzy AND operator:

$$W_O = \text{AND}\{W_{\text{PD}}, W_{\text{T2}}, W_{\text{FLAIR}}, p_{\text{maha}}\}. \tag{6.9}$$

One post-processing step is performed after graph cut. As many false positives occur due to artifacts in the external CSF, all lesions detected neighboring the brain border are removed from the segmentation.

6.2.3 Semi-automatic Post-processing

Users of automatic segmentation methods need to be able to verify the automatic segmentations and refine the segmentation by adding missed lesions and removing false positives. As mentioned in Section 6.2.2, our graph cut method offers the possibility of user interaction.

Figure 6.4 shows an example of the possible interaction. Once the automatic segmentation is displayed, the user discovers a missed lesion in the images of the patient. The user adds an object seed in the lesion, see equation (6.5). The graph cut is then recalculated and the new lesion is added.

6.3 Parameter Setup

In this section, we propose two experiments to analyze the behavior of our method with respect to α of the graph cut framework and to the fuzzy rules of the regional term.

6.3.1 Fuzzy Rules

MS lesions are considered as *hyper-intense* compared to white matter, but there is no consensus on how to define hyper-intensity. In the two last chapters we defined a hard threshold with which determine the hyper-intensity, but we proposed to use a fuzzy definition of hyper-intensity, as a binary decision is already made by the graph cut.

The idea of this experiment is to vary our fuzzy definition of hyper-intensity. In practice, we have to vary two parameters: the slope beginning S_b and the slope end S_e. We note that when $S_b = S_e$, the fuzzy rule becomes a binary rule. We perform

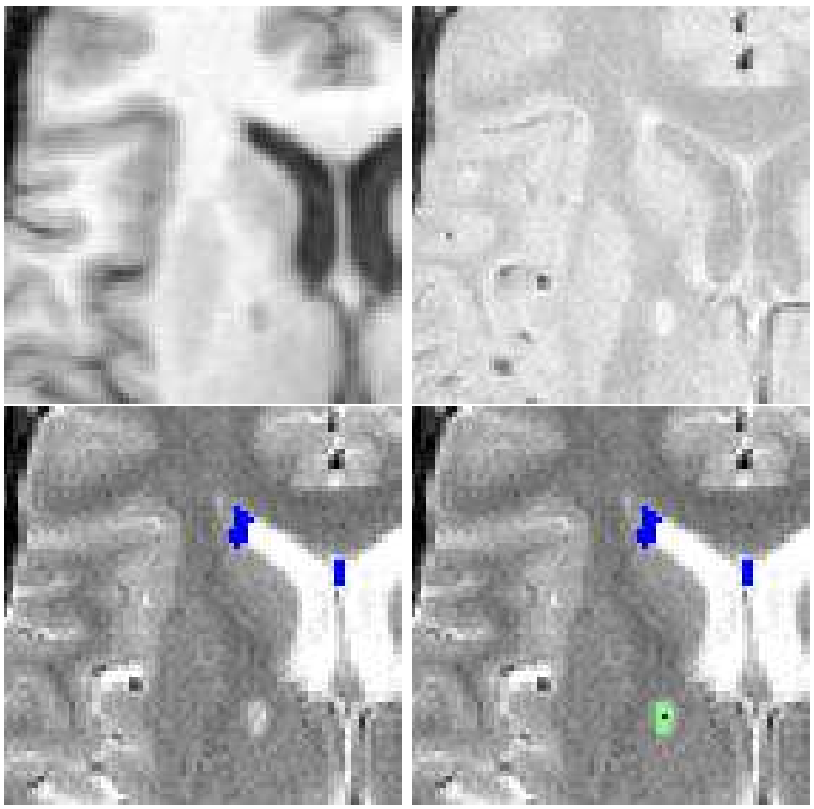


Figure 6.4: Example of semi-automatic postprocessing. Top line: T1-w and PD-w. Bottom line: automatic GC solution and semi-automatic solution (Red: object seed, Green: graph cut solution, Blue: automatic segmentation).

the segmentation on *bw_mild*, *bw_moderate* and *bw_severe* (Table 2.1) using our method modifying S_b and S_e from 1 to 4 in steps of 0.5. The segmentations are compared to the ground truth using the DSC (see Chapter 2). As proposed by Lecoeur et al. [Lecoeur et al. 2008], α was set to 10 in this experiment.

The results for every pair of $S_b; S_e$ values are shown in Figure 6.5. The best set of parameters varies for each lesion load: mild $(2.5; 3.5)$, moderate $(3; 4)$ and severe $(3.5; 4)$. For every lesion load, the fuzzy rule is better than any of the hard threshold rules $(S_b = Se)$. The best overall set of parameters is $S_b = 2.5$ and $S_e = 3.5$.

6.3.2 Alpha

Regional and boundary terms are influenciated by α. Lecoeur et al. [Lecoeur et al. 2008] set $\alpha = 10$ but we modify α in order to gain a better understand its influence on the graph cut algorithm.

We choose from the Brainweb database the images with moderate lesion load, 20% inhomogeneity and two levels of noise (3% and 5%). In this case, we set the fuzzy rules $S_b = 2.5$ and $S_e = 3.5$ and vary α from 0 to 100. The segmentation is then compared to the ground truth of the lesions using the DSC.

It is interesting to notice the two extreme situations of the analysis. When $\alpha = 0$,

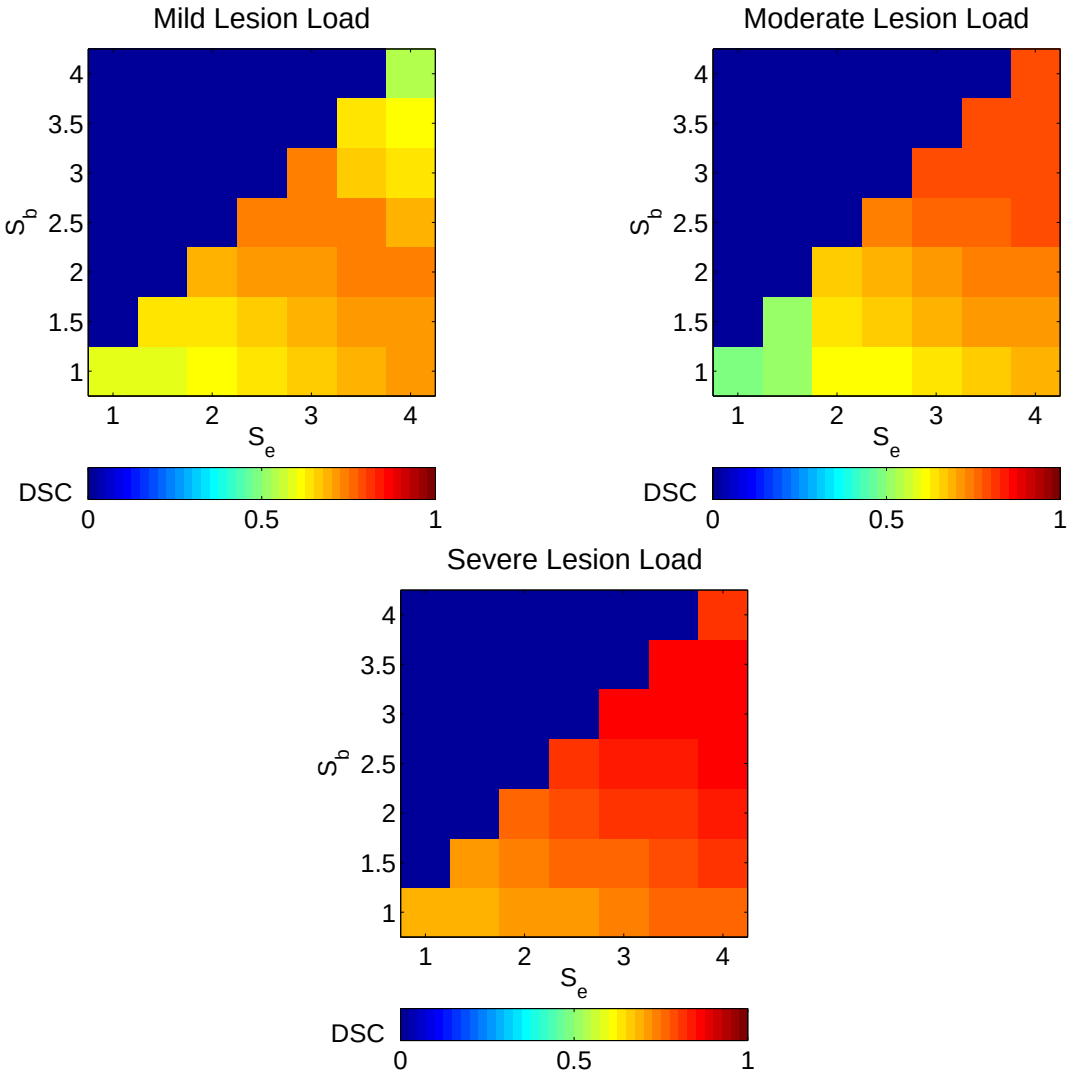


Figure 6.5: Results of DSC for different fuzzy values on Brainweb images. Note that when $S_b = S_e$ the fuzzy rule becomes a hard threshold.

the regional term is not employed and only the information of the boundary term is employed. If we replace $\alpha = 0$ in (6.5), we can observe that only the seeds can influence the segmentation, thus for the automatic segmentation this is of little use. On the contrary, when α increases, the regional term has a greater weight in the cost function and the spatial information is no longer employed in the segmentation. When $\alpha \rightarrow$ inf, the graph cut becomes a threshold where only the value of the two weights is taken into account.

Figure 6.6 describes the behavior of α. We observe that the best performance of the algorithm is situated around the 3-10. When $\alpha \leq 1$, the boundary information has more influence than the regional information and the algorithm fails to yield a good result.

When $\alpha > 60$, the boundary information has little influence and the graph cut becomes an automatic threshold technique; no spatial information is employed and the method is sensible to the noise.

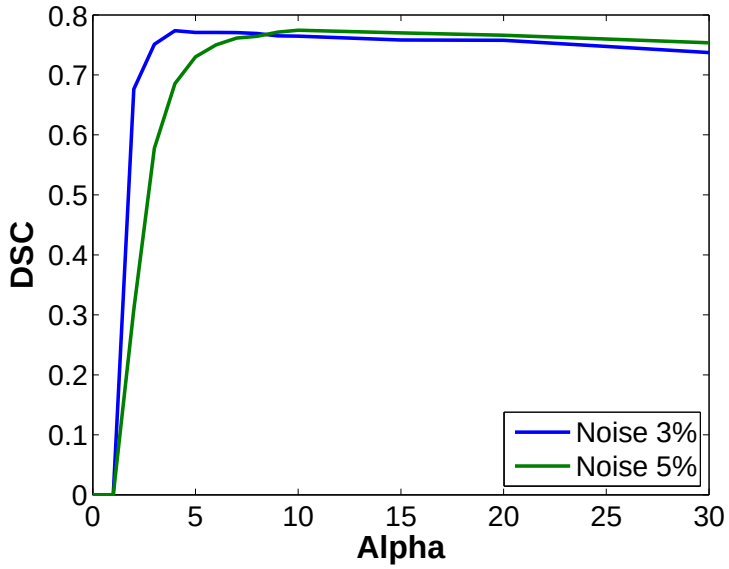


Figure 6.6: Results on Brainweb with different values of α.

Figure 6.7 shows some visual results. When $\alpha = 100$, the boundary information has little influence and the graph cut becomes an automatic threshold technique, small points are detected as lesions, thus producing multiple false positives. When $\alpha = 2$, we observe that the top right elongated lesion is truncated. This happens because the cost of the boundary term is higher than the cost of the regional term, which is commonly known as *shrinking bias* [Vicente et al. 2008], and causes that elongated targets to be incorrectly segmented. When the regional term becomes is higher ($\alpha = 10$) this effect disappears.

6.4 Comparison with the Semi-automatic Approach

We compared our results with the semi-automatic method proposed by Lecoeur et al. [Lecoeur et al. 2008]. Comparing of an automatic method with a semi-automatic method is a complicated process as the semi-automatic method is user-dependent; the results of the segmentation varies depending the expert's experience in MS lesion detection and expertise on the segmentation method employed.

In order to simplify the evaluation, we simulated the seed selection by two means: Random and Erosion. In the Random selection, a percentage of the ground truth voxels are selected as seeds at random. In the Erosion selection, a percentage of the ground truth voxels is selected, starting with the voxels in the center of the lesions. This second selection method should be more reliable as raters usually select voxels in the center of a lesion.

We compute the semi-automatic method in *bw_moderate* and compare it with our automatic method in terms of DSC (Figure 6.8). The results of our automatic method correspond to a percentage of seeds of 5% for the random seeds and a 7% for the seeds selected by erosion. This means that the rater has to select around 175 voxels at random or up to 245 voxels from the center of the lesions, which shows

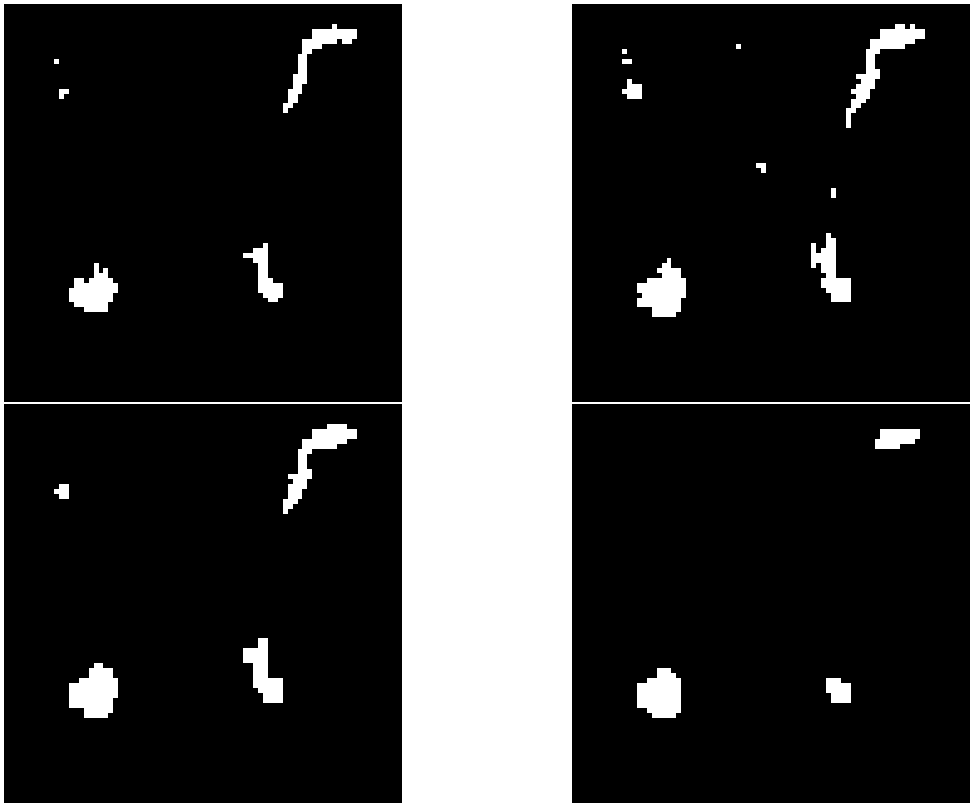


Figure 6.7: Top line, from left to right, top to bottom: Brainweb ground truth and automatic segmentation with $\alpha = 100$. Bottom line, from left to right: automatic segmentations with $\alpha = 10$ and $\alpha = 2$.

that the gain obtained by our algorithm is significant, although a further analysis would be necessary using real raters.

6.5 Conclusion

We proposed an automatization of the graph cut framework for the segmentation of MS lesions. Based on the method proposed in Chapter 4, we proposed a new segmentation method that incorporates spatial information thanks to the spectral gradient.

This new method offers two main advantages. First, it includes gradient information which should make the method more robust to noise and intensity inhomogeneity. Second, the method offers the possibility of subsequently refining the segmentation if the expert is not completely satisfied.

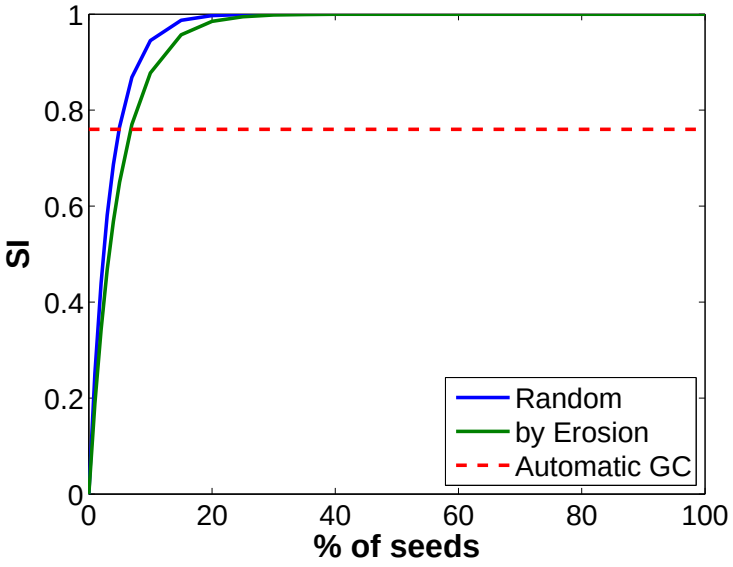


Figure 6.8: Results on Brainweb ($n = 3\%$ and $rf = 20\%$) for the semiautomatic version of graph cuts [Lecoeur et al. 2008], with random selection of seeds and with selection of seeds by erosion.

Chapter 7

Validation

In Chapter 2, we gave an overview about the validation of MS lesion segmentation methods. In this chapter, we evaluate the three method proposed in the last three chapters to measure their accuracy.

We start describing the metrics employed in the validation of our methods and compare our methods using a database of synthetic images and two databases of clinical images. We then finish with our conclusions of the comparison.

7.1 Metrics

The evaluation of a segmentation method changes whether the exact solution is known or not. On synthetic images, this solution is available and is known as the *ground truth*. When the ground truth is available, the evaluation consists in measuring the dissimilarities between the results of the segmentation method and the ground truth.

In clinical images, the ground truth is not available and the evaluation of the segmentation is more complex. In MS lesion segmentation, automatic methods are usually compared against gold standards based on manual segmentations. A basic evaluation method considers the manual segmentation to be the ground truth.

Dissimilarities between the results of an automatic segmentation method and gold standards can be due to both errors from the automatic segmentation method or variability in inter- and intra-experts' segmentations [Grimaud et al. 1996]. When several manual segmentations are available, other evaluation methods can be applied where the expert's variability is taken into account in the comparison.

7.1.1 Evaluation Against the Ground Truth

The evaluation of a segmentation method on synthetic images consists of comparing the segmentation result against the ground truth that served to create the images. This approach is also employed in clinical images when only one expert's manual segmentation is available and we consider the manual segmentation to be the ground truth.

Many metrics have been employed for the comparison of a segmentation result with a reference image. As we mentioned in Chapter 2, the most popular measure

in MS lesion segmentation is the Dice Similarity Coefficient

$$DSC = \frac{2 \times |R \cap S|}{|R| + |S|}. \tag{7.1}$$

Values ranges from 0.0 to 1.0, indicating complete failure or perfect segmentation respectively, with 0.7 typically considered to be a good agreement [Zijdenbos et al. 1994]. We can rewrite this equation using the information in Figure 7.1 as

$$DSC = \frac{2 \times TP}{2 \times TP + FN + FP}. \tag{7.2}$$

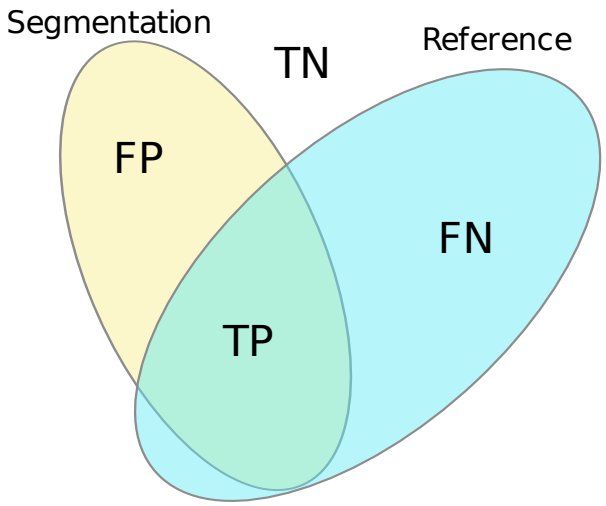


Figure 7.1: Description of the four possibilities when comparing two images: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

7.1.2 Evaluation Against Several Manual Segmentations

In this kind of evaluation, the question we want to answer is whether the automatic segmentation method is as good as an experts' manual segmentation.

First, manual segmentations are employed to create an improved segmentation, which is called a *silver standard*. The automatic segmentation methods are then evaluated using this silver standard and compared to the variability obtained among experts.

7.1.2.1 STAPLE

The STAPLE (Simultaneous Truth And Performance Level Estimation) algorithm was designed to study the performance of different experts when the ground truth

is not available [Warfield et al. 2004]. This algorithm takes into account that all segmentation methods or experts are somehow imperfect and that their sensitivity and specificity can be measured. The sensitivity and specificity of each method is estimated using an EM approach that estimates at the same time the STAPLE silver standard (SSS).

Once the SSS is created, we can compute the sensitivity Se and specificity Sp of other methods as follows

$$Se = \frac{\sum_{i=0}^{n} D_i \cdot W_i}{\sum_{i=0}^{n} W_i} \tag{7.3}$$

$$Sp = \frac{\sum_{i=0}^{n} \bar{D}_i \cdot (1 - W_i)}{\sum_{i=0}^{n} (1 - W_i)} \tag{7.4}$$

where D_i is the voxel i of the binary segmentation and W_i is the probability of the voxel i on SSS.

There are three different options to use STAPLE with both manual and automatic segmentations. The first option is to include all manual and automatic segmentations in the STAPLE computation and obtain simultaneously the SSS, and the sensitivity and specificity for every segmentation. In this option, the hypothesis is that the quality of the automatic segmentations is similar to the quality of manual segmentations and thus they can be included in STAPLE. The number of manual segmentations should be larger than the number of automatic segmentations or the SSS might be biased against the automatic methods.

The second option assumes that segmentations from experts are better than segmentations from automatic methods and therefore STAPLE is employed to create SSS with manual segmentations only. The Se and Sp of automatic methods are computed afterwards using equations (7.3) and (7.4). Since only manual segmentations are used in the creation of the SSS, it would be biased towards automatic segmentation methods.

The third option requires the experts to perform two manual segmentations for each image. The first manual segmentation of every expert is employed in the computation of the SSS. Then, we compute sensitivity and specificity for all automatic segmentations and the second manual segmentations using equations (7.3) and (7.4). The advantage of this option compared to the former two options is that the creation of the SSS and the evaluation of the sensitivity and specificity are performed with different images, which leads to a less biased comparison. The drawback of this option is the requirement of additional time-consuming manual segmentations.

7.1.2.2 Voting Silver Standard

Due to the variability between experts, one strategy is to employ the consensus among experts to create a silver standard. A voxel is considered a lesion if the majority of the manual segmentations consider it as a lesion, we call this resulting image the Voting Silver Standard (VSS).

We propose a measure based on the DSC to measure the variability among experts and to compare the variability with the automatic method. We call this measure the raters DSC (rDSC). If we consider R experts, for every pair of experts

r and s, we compute the $\mathrm{DSC}_{r,s}$, the mean and the standard deviation of the rDSC gives us an indication of the level of agreement among experts.

$$\mu_{\mathrm{rDSC}} = \frac{1}{\binom{2}{R}} \sum_{r=0}^{R} \sum_{s=0, s\neq r}^{R} \mathrm{DSC}_{r,s} \tag{7.5}$$

$$\sigma_{\mathrm{rDSC}} = \sqrt{\frac{1}{\binom{2}{R}} \sum_{r=0}^{R} \sum_{s=0, s\neq r}^{R} (\mathrm{DSC}_{r,s} - \mu_{\mathrm{rDSC}})^2}. \tag{7.6}$$

To evaluate an automatic segmentation method, we compute the DSC using the VSS as reference and compare the values obtained with the rDSC. The advantage of this strategy is that the creation of the silver standard is intuitive and the DSC is a well known measure in the domain of segmentation of MS lesions.

7.2 Brainweb Database

In Section 2.4.1, we introduced the Brainweb database, which is a realistic synthetic database of MR images. These images were already used to study the parameters of the three proposed segmentation methods.

7.2.1 Data

This database contains three MS phantoms with different lesion loads (mild, moderate and severe). From each of these phantoms, we can produce synthetic MR images with different MR parameters and image artifacts. For all levels of noise (1%, 3%, 5%, 7% and 9%) and IIH (0%, 20% and 40%), we obtain the three sequences (T1-w, T2-w and PD) from the three MS phantoms The TLL of each phantom is: mild, 0.4 cm^3; moderate, 3.5 cm^3; and severe, 10.1 cm^3.

7.2.2 Results

We process all images in the Brainweb database using default parameters for our three automatic segmentation methods. The images are already aligned and no further preprocessing was employed on images. We use the ground truth to determine the brain mask and DSC was computed for each method (Figures 7.2, 7.3 and 7.4).

For the mild lesion load (Figure 7.2), we obtained lower DSC than for the other two lesion loads, We also notice that with the mild lesion load phantom with 40% inhomogeneity, all three methods obtained poor DSC. The DSC is more sensitive when the segmentation region is small [Zijdenbos et al. 1994].

Our methods obtained a lower DSC for 1% than for 3% noise. On images with 1% noise, variance on each class is very small when there is 1% noise which results in poor detection of the candidate lesions. The variance on clinical is greater than Brainweb with 1% noise and therefore this case is of little practical interest.

MS4MS obtains higher DSC than STREM in presence of IIH and noise. GCEM shows higher DSC than the other two methods for high levels of noise. It also

exhibits a similar behavior to MS4MS in regards to the IIH. GCEM obtains the best score in overall, however MS4MS obtains the highest DSC for the normal conditions (noise=3% and IIH=20%), which are the conditions employed in the optimization of parameters.

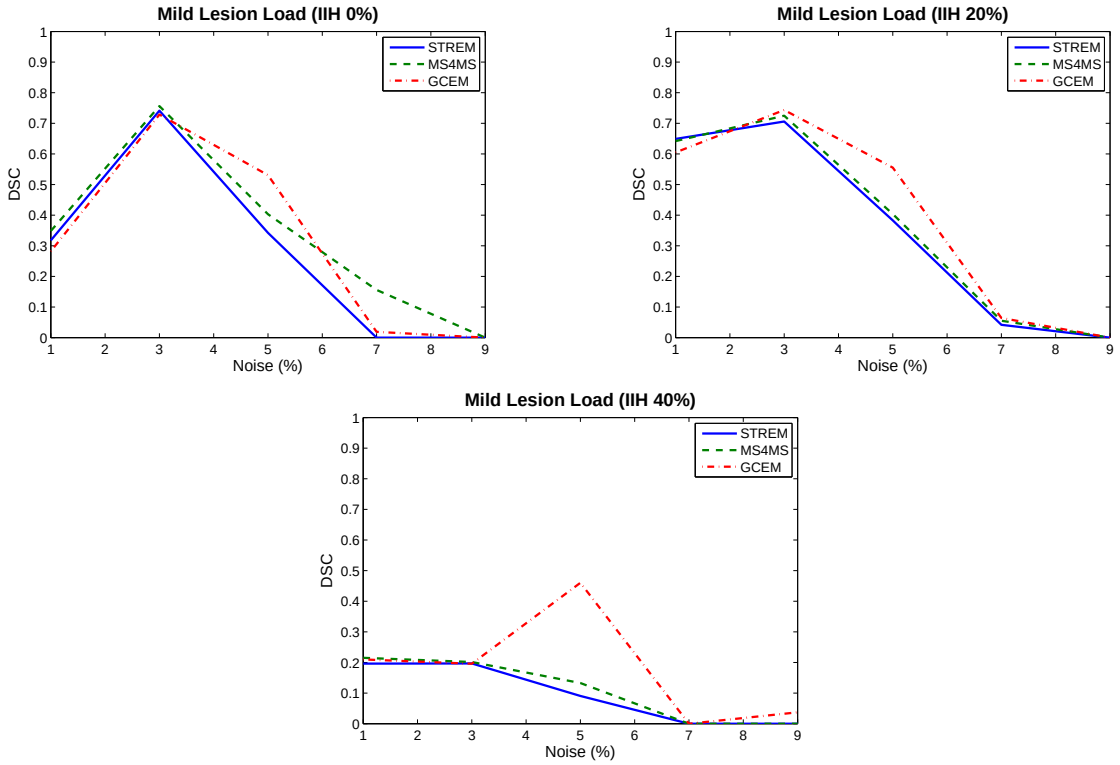


Figure 7.2: DSC results for images with mild lesion load from the Brainweb database for our three methods: STREM, MS4MS and GCEM. Each graphic represents one level of intensity inhomogeneity (IIH): 0%, 20% and 40%.

7.2.3 Conclusion

The comparison on synthetic data gives information about the behavior of the methods with different levels of noise and inhomogeneity. The parameters of all methods were optimized for 3% of noise and therefore they obtained their best results.

STREM accuracy decreases rapidly in presence of noise and inhomogeneity. The creation of regions with MS4MS improves the robustness of the method against noise and inhomogeneity yielding better DSC values than STREM because a region is easier to classify than a voxel alone.

GCEM is more robust to noise than the other two methods. The inclusion of contour information given by the intensity gradient makes the method robust to both noise and inhomogeneity.

Table 7.1 show the results of our methods and several state-of-the-art methods on the Brainweb database. Each method uses only some images from the Brainweb database making the comparison complicated. We observe that our methods obtain similar results to those in the literature.

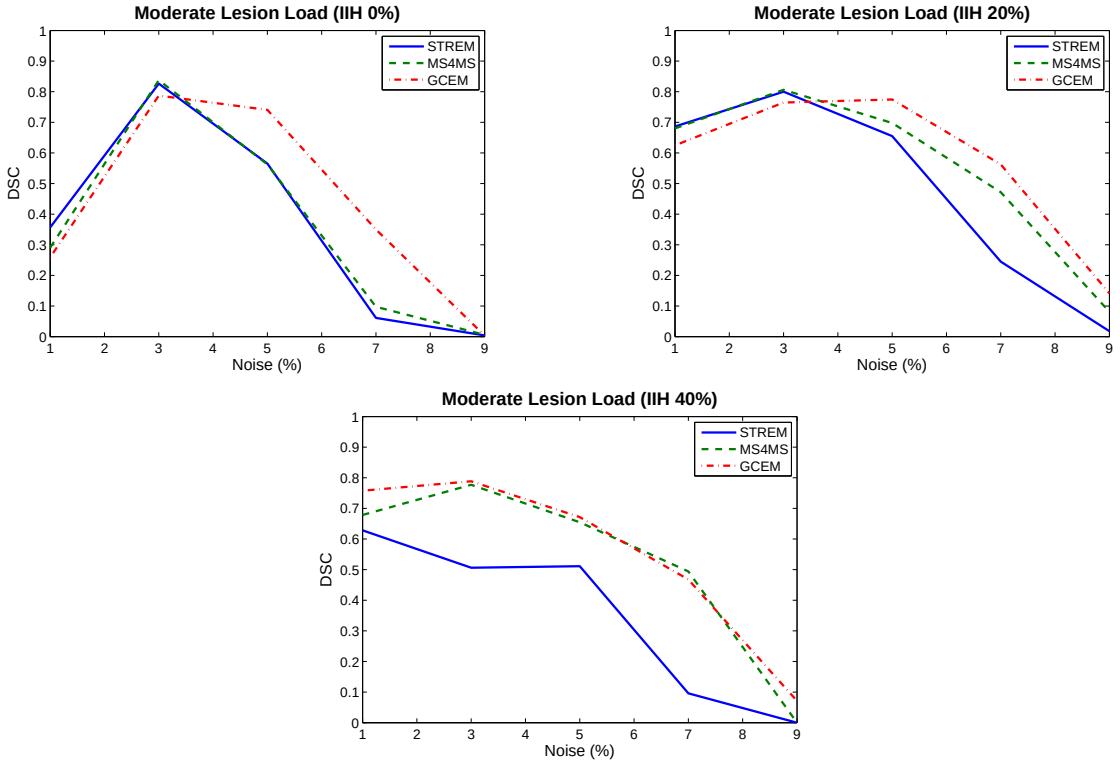


Figure 7.3: DSC results for images with moderate lesion load from the Brainweb database for our three methods: STREM, MS4MS and GCEM. Each graphic represents one level of intensity inhomogeneity (IIH): 0%, 20% and 40%.

7.3 Rennes Database

Images in this database are available with no preprocessing. We then apply the complete segmentation workflow proposed in Chapter 3 using our three segmentation methods.

7.3.1 Data

Images of thirteen MS patients were acquired on a Philips Achieva 3T System (Philips Medical Systems, Best, The Netherlands). The MR protocol consisted of:

T1-w: SENSE=4, TE=4.6 ms, voxel size $1 \times 1 \times 1$ mm.

Dual Echo: TSE, TR=2000 ms, TE1/TE2=10/90 ms, voxel size $0.93 \times 0.93 \times 3$ mm.

FLAIR: TSE, TR=11000 ms, TE=125 ms, TI=2800 ms, voxel size $0.93 \times 0.93 \times 3$ mm.

Images were acquired at the Pontchaillou University Hospital (Rennes, France) on clinical routine. There were no selection criteria for the patients, as such there was a great variability on disease progression. Because of confidentiality reasons, no clinical information of the patients was available. In all patients, MS lesions were

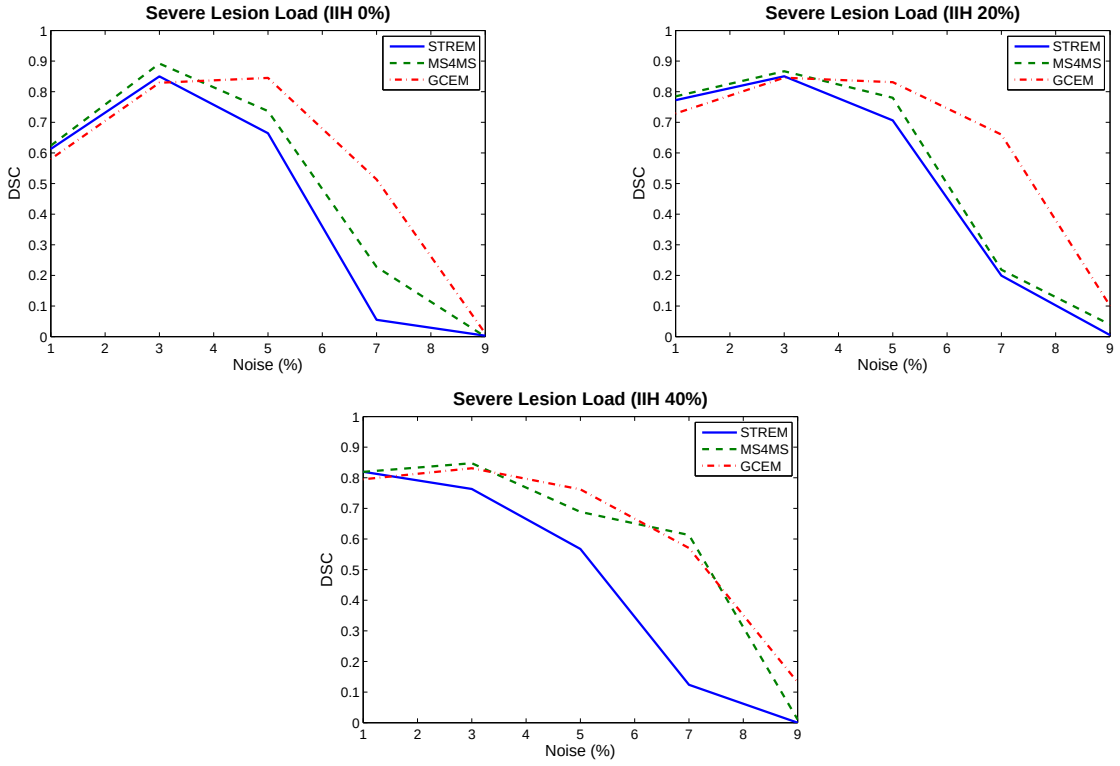


Figure 7.4: DSC results for images with severe lesion load from the Brainweb database for our three methods: STREM, MS4MS and GCEM. Each graphic represents one level of intensity inhomogeneity (IIH): 0%, 20% and 40%.

manually delineated on FLAIR images by one expert using the Anatomist software[1]. In this database, all images are in the DICOM format without any processing thus the complete segmentation workflow can be evaluate. PD images showed a intensity contrast similar to the T2-w images and was discarded for the segmentation.

Patients were ordered by TLL which was computed from the expert's manual segmentation (Figure 7.5). We observe that four patients have a similar TLL to *bw_mild*, five patients obtain a similar TLL to *bw_moderate* and four patients have a TLL higher than *bw_severe*.

7.3.2 Results

We perform the segmentation using the workflow proposed in Chapter 3 using T1-w, T2-w and FLAIR images with the three proposed segmentation methods. The segmentation results are compared with manual segmentation using DSC.

Figure 7.6 shows the DSC values for all segmentation methods. We observe that the three segmentation methods obtain similar results. GCEM obtains a higher DSC on the last three patients, and MS4MS obtains the best average for the four patients with lower lesion load. On average (Table 7.2), GCEM is slightly better than the other two methods although results show a low agreement between the

[1] `http://brainvisa.info/`

Method	*bw_mild*	*bw_moderate*	*bw_severe*
STREM	0.72	0.72	0.80
MS4MS	0.72	0.80	0.87
GCEM	0.74	0.76	0.84
[Rousseau et al. 2008]	0.52	0.63	0.82
[Freifeld et al. 2007]		0.77*	
[Van Leemput et al. 2001]		0.80*	
[Zijdenbos et al. 2002]		0.79	
[Shiee et al. 2009]		0.81	

Table 7.1: DSC values on the Brainweb images *bw_mild*, *bw_moderate* and *bw_severe* for the three methods and other results obtained from the literature. (* DSC was only computed on slices from 60 to 120)

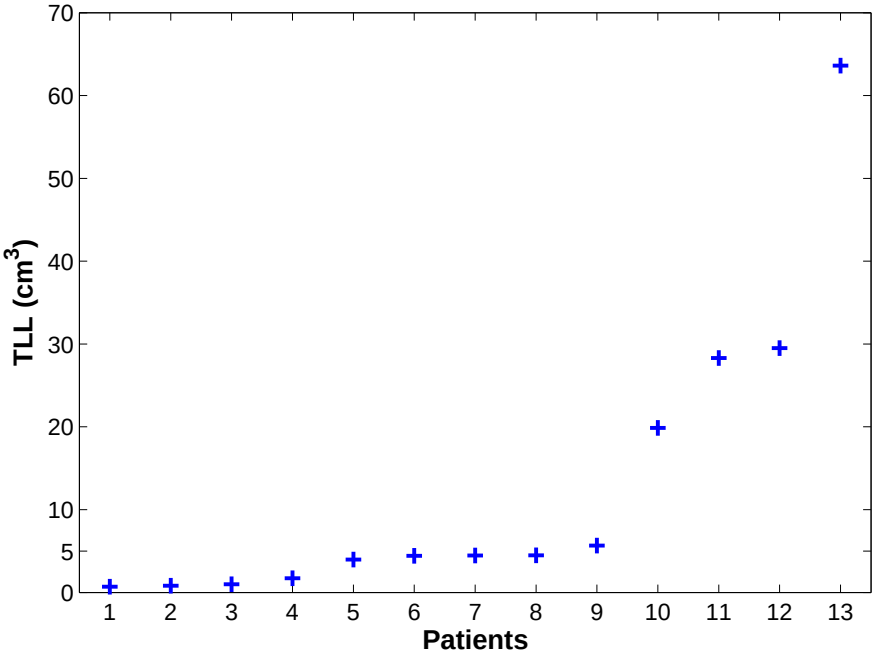


Figure 7.5: Total lesion load (TLL) computed from the manual segmentation of the expert on the Rennes database.

manual segmentation and the automatic methods.

7.3.3 Visual Comparison

The observed effects of the preprocessing workflow are described in this section (Figure 7.7). In the raw images, we observe an important intensity inhomogeneity in the T1-w image, where the white matter in center of the brain is brighter than the anterior part of the brain. We also observe intensity noise, especially in T1-w and FLAIR images. Once the correction methods are applied, we obtain images that contain less noise and are more homogeneous than the originals without any

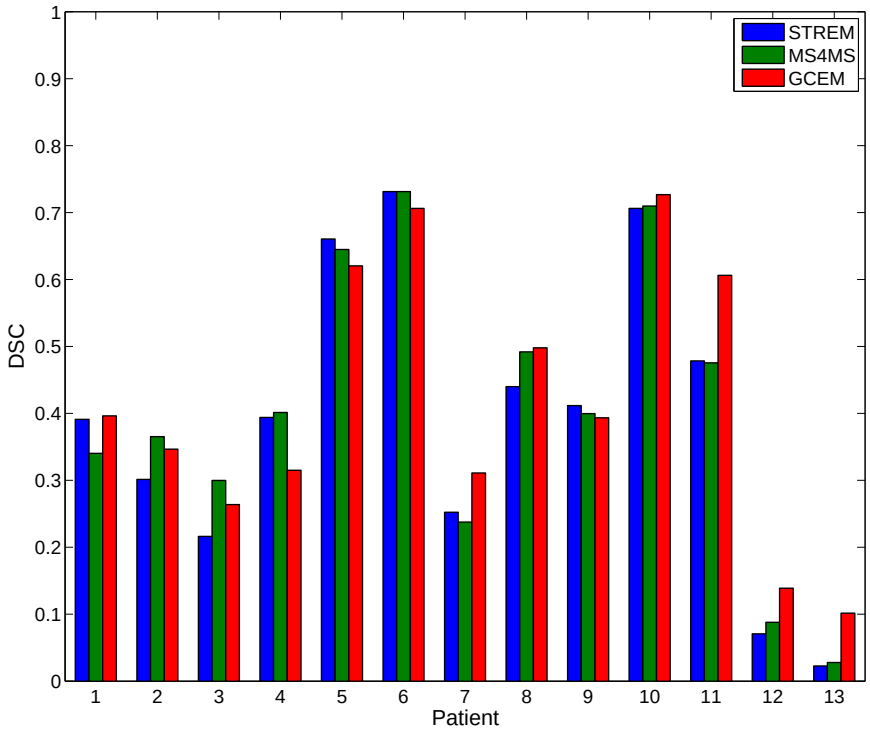


Figure 7.6: DSC results for Rennes database.

Method	Average DSC
STREM	0.39
MS4MS	0.40
GCEM	0.42

Table 7.2: Average DSC values for the three workflows on the Rennes database.

lesions to disappear.

We visually compare patients 5 and 7 that have similar lesion loads but different DSC score.

Figure 7.8 shows the results on patient 5, on which the automatic methods obtain a DSC around 0.6. We observe a good agreement in the contour of the lesions between manual segmentation and the automatic methods. In the FLAIR images, the interface between the ventricles and the white matter is bright resulting many false positives in our segmentation.

Figure 7.9 shows the results on patient 7, on which the automatic methods obtain a DSC around 0.3. We also observe false positive in the bright region between the two ventricles in the FLAIR images. One of the four lesions is missed by the three segmentation methods, the intensity contrast of these lesion is less important than the contrast of the others. We observe that the lesions in T2-w are smaller than the lesions in FLAIR, which has been pointed out in the literature [Filippi et al. 1999]. Our methods consider voxels as part of a lesion only when the intensity of

both FLAIR and T2-w are hyper-intense. This could explain the systematic under-segmentation of our methods.

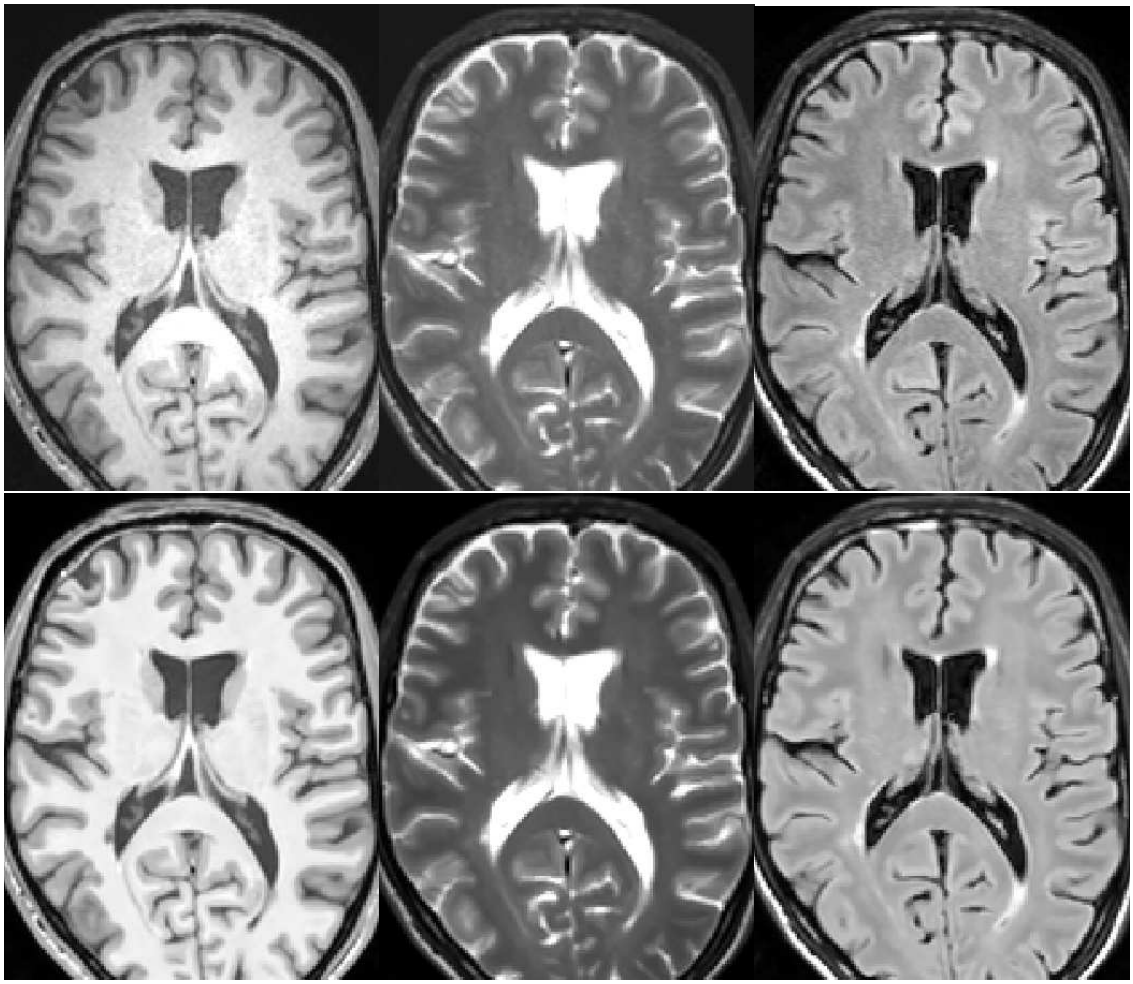


Figure 7.7: Preprocessing results of patient 5 from the Rennes database. Top line, from left to right: original T1-w, T2-w and FLAIR. Bottom line, form left to right: processed T1-w, T2-w and FLAIR.

7.3.4 Conclusion

In this database, only one expert segmented the images and therefore the variability of the manual segmentation was not evaluated. We can point out two main reasons for the low agreement found between our methods and the manual segmentation.

First, the manual rater segmented lesions on FLAIR images without comparing the contour with T2-w images. Our methods only segment lesions visible in both FLAIR and T2-w which explains our systematic under segmentation. A solution would consist in detecting the lesions using both images and delineating the exact contour only using FLAIR.

Second, FLAIR shows a high contrast between white matter and lesions but also contains many bright regions in between ventricles and white matter that causes multiple false positives in our segmentation.

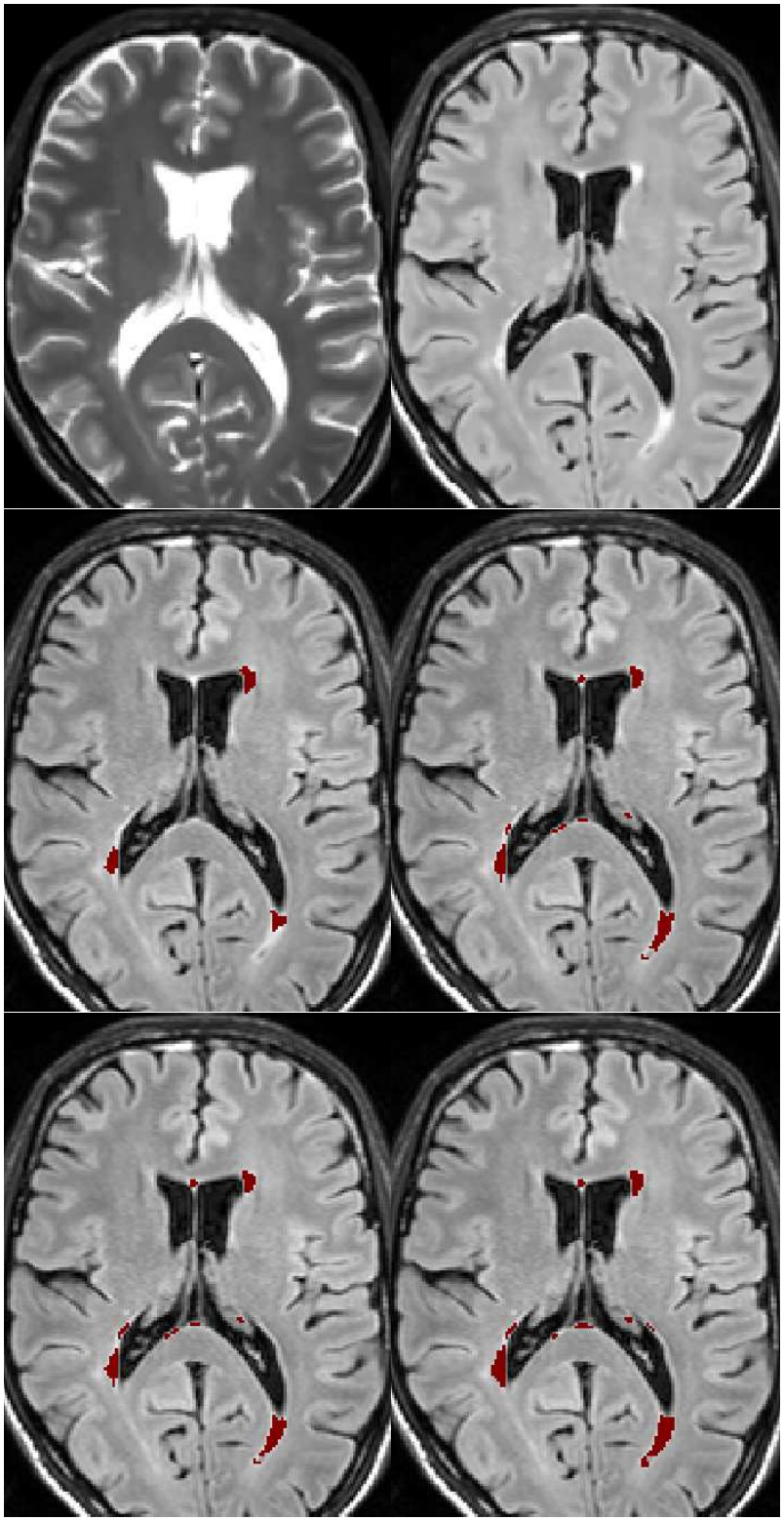


Figure 7.8: Segmentation results of patient 5 from the Rennes database. Top line, from left to right: processed T2-w and FLAIR. Middle line, form left to right: manual segmentation and STREM. Bottom line, from left to right: MS4MS and GCEM.

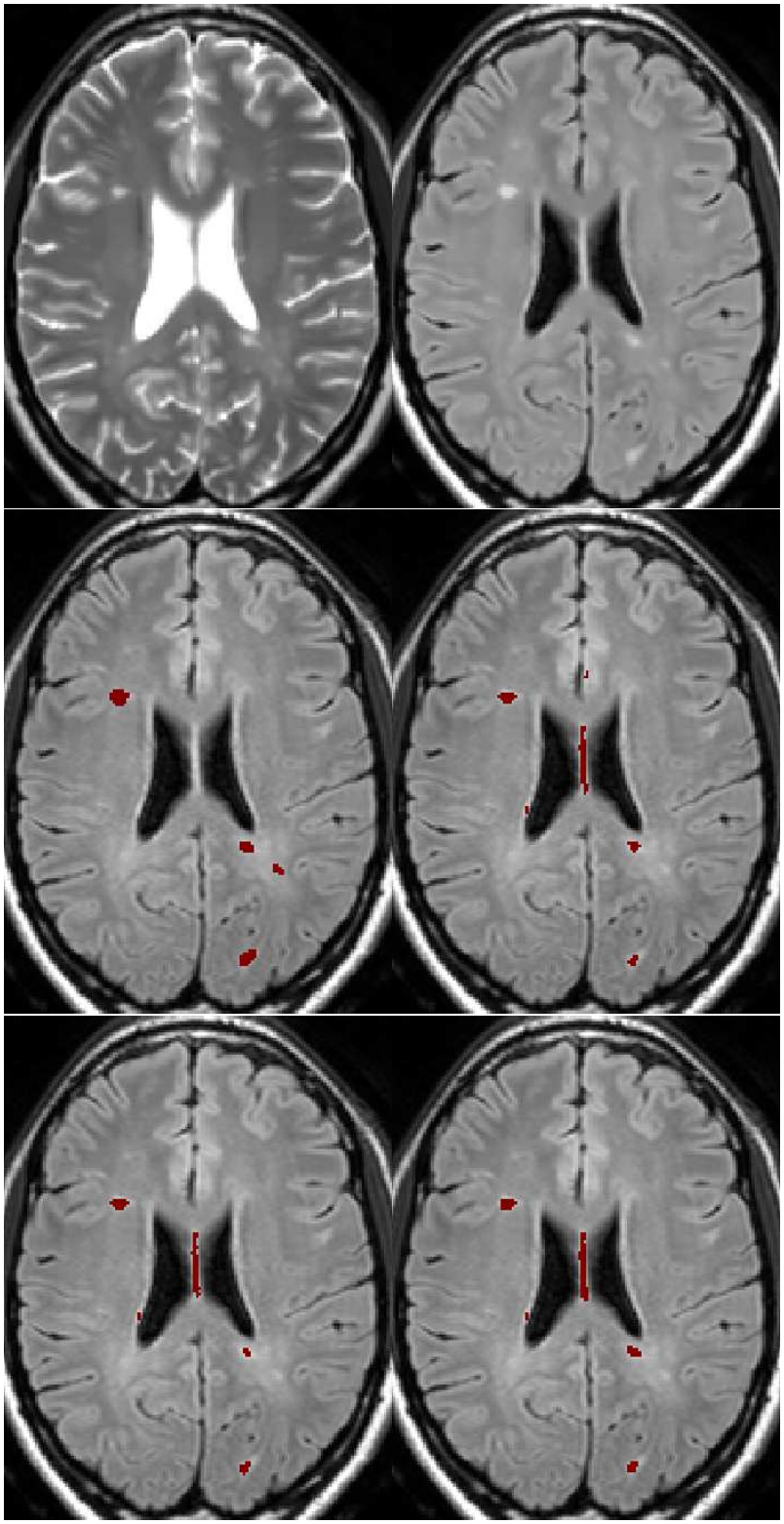


Figure 7.9: Segmentation results of patient 7 from the Rennes database. Top line, from left to right: processed T2-w and FLAIR. Middle line, form left to right: manual segmentaion and STREM. Bottom line, from left to right: MS4MS and GCEM.

The existence of only one rater limits the conclusions of this evaluation. More manual segmentations would give us information of the variability of the expert and enable the creation of a silver standard to compare our methods. In order to compare our methods to this database, some modifications are required to adapt the definition of lesions of our methods to the one of the expert.

7.4 MNI database

This database was created in the Montreal Neurological Institute (Montreal, Canada) to evaluate their segmentation methods. The advantage of this database is the existence of multiple manual segmentations that provides the possibility to employ the methods above mentioned for multiple experts.

Images in the database were already corrected for intensity inhomogeneity and registered in the same space, thus only the segmentation methods are evaluated and not the whole segmentation workflow.

7.4.1 Data

Images from ten relapsing-remitting MS patients were acquired on a GE 1.5T System (General Electrics, USA) at the Montreal Neurological Institute (MNI). The MR protocol included T1-w and Dual Echo (T2-w and PD) images with 3-mm axial slice thickness and with slice-resolution of $1mm$.

Images were rigidly registered [Collins et al. 1994] and corrected for intensity inhomogeneity [Sled et al. 1998]. Five experts manually segmented MS lesions on the processed images using the `display`[2] software developed at the McConnell Brain Imaging Center. The same images were segmented again by the same experts. To sum up, each patient has associated 10 manual segmentations of the MS lesions, two segmentations per expert.

Figure 7.10 shows the variation of the TLL computed using the manual segmentation from the experts, patients were ordered by TLL. Comparing the TLL of the clinical patients with the Brainweb phantom, we notice that the three last patients have more than twice the TLL of the severe Brainweb phantom and the distribution of TLL is similar to the Rennes database.

7.4.2 Results

We perform BET [Smith 2002] skull-stripping to restrict our segmentation to the brain, and segment MS lesions using our three segmentation methods. In addition, we compare our methods to one state-of-the-art MS lesion segmentation method called Expectation-Maximization Segmentation (EMS) [Van Leemput et al. 2001].

EMS[3] is an automatic segmentation method also based on the estimation of a GMM for the NABT. It is based on a modified EM algorithm where voxels are weighted in the estimation according to the probability that they are outliers. The

[2]http://www.bic.mni.mcgill.ca/software/Display/(MNI) Display.html
[3]http://www.medicalimagecomputing.com/downloads/ems.php

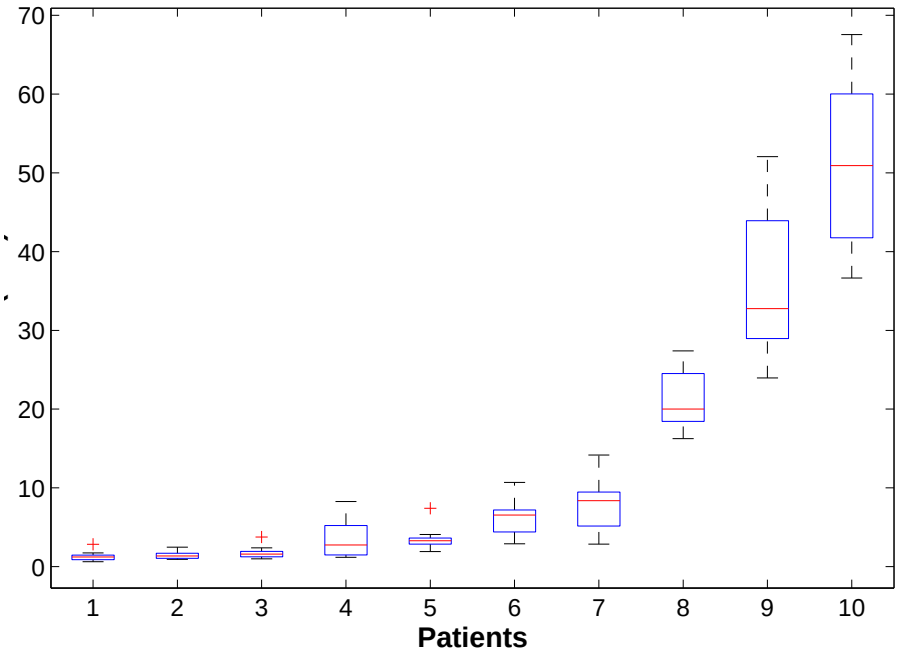


Figure 7.10: Total lesion load (TLL) computed from the manual segmentation of the experts. We observe high variability especially for the highest lesion loads.

modified EM algorithm includes the information from a brain atlas and also corrects for intensity inhomogeneity. EMS is employed with the default parameters and the resulting probabilistic map is thresholded at 0.5 to get the final binary segmentation.

7.4.2.1 STAPLE

We use the STAPLE approach to analyze the sensitivity and specificity of the experts and the automatic segmentation methods. Each image in this database have each been segmented twice by the experts. For this reason, we employ the third option described in Section 7.1.2.1, where the first set of segmentations is employed in the construction of the Staple silver standard (SSS) and the second set of segmentations is employed in the evaluation of the experts.

Specificity is not a very useful measure in the context of MS lesion segmentation. The volume of lesions is very small compared to the volume of the brain mask. In percentage, lesions occupy $0.1-4\%$ of the brain. A specificity of 99% means that 1% of voxels were uncorrectly segmented as lesions, which is equivalent to the lesions in some patients.

Figure 7.11 shows sensitivity and specificity for the five experts and the automatic methods. We observe great variations in sensitivity and specificity for the five experts. EMS shows higher sensitivity but lower specificity than the proposed automatic methods. GCEM shows a better sensitivity than MS4MS and STREM but again, lower specificity.

There is no expert that is clearly superior to the others. For example, expert 1 shows the highest sensitivity but also the lowest specificity. From this, we deduce

that he tended to oversegment in comparison to the other experts. Automatic segmentation methods have the same behavior, STREM obtains the lowest sensitivity and the best specificity and EMS obtains the opposite.

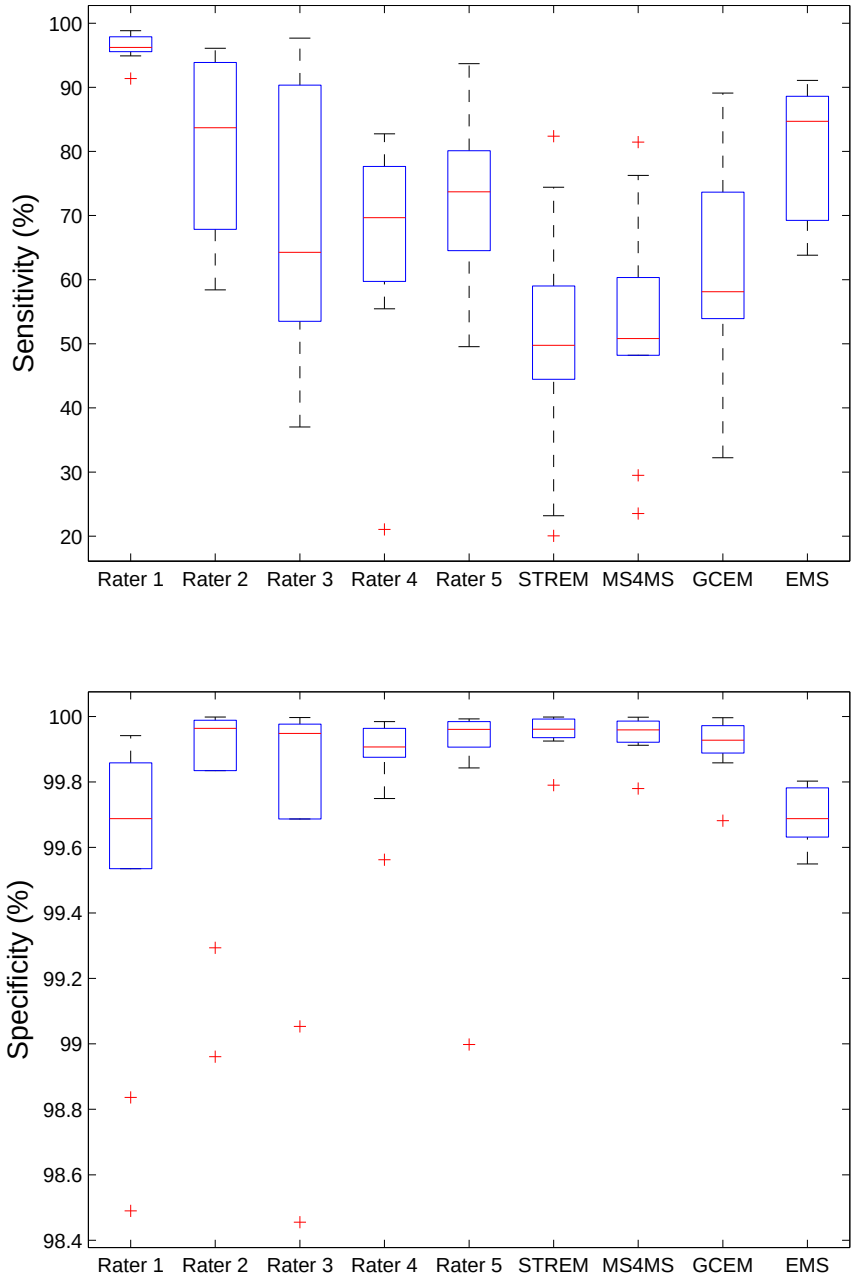


Figure 7.11: Boxplot of the sensitivity (top) and specificity (bottom) for each expert and automatic method on the MNI database.

Method	Average DSC
μ_{rDSC}	0.68
STREM	0.64
MS4MS	0.65
GCEM	0.65
EMS	0.57

Table 7.3: Average DSC values for the MNI database.

7.4.2.2 Voting Silver Standard

We construct a voting silver standard (VSS) using the 10 manual segmentations available. In the VSS, a voxel is considered to be a lesion if 6 out of 10 experts of the manual segmentations considered it a lesion. The objective is to create a binary image with less variability to compare the results of the automatic segmentation methods. We also compute μ_{rDSC} and σ_{rDSC} to measure the variability of the experts segmentation. We segment all images using the automatic methods and compute DSC using the VSS as reference.

Figure 7.12 shows results of the automatic segmentations against the variability of the manual segmentations. We observe that in 8 out of 10 patients the DSC of our three methods is within $\mu_{\text{rDSC}} \pm \sigma_{\text{rDSC}}$ while for the EMS it happens in 7 out of 10 patients. The low scores for our methods correspond to the two patients with highest TLL while the low scores for EMS are found in the three patients with lowest TLL.

If we compare our methods against EMS, we observe that in the first 7 patients (8 for GCEM) the DSC of our methods is higher than EMS. EMS has a higher DSC than our three methods in the two patients with the highest TLL only. If we compare our three methods, we observe three different patterns regarding the lesion load. For mild TLL (Patients 1,2 and 3), STREM is the best followed by MS4MS and GCEM. For moderate TLL (Patients 4, 5, 6 and 7), STREM and MS4MS are very similar and slightly better than GCEM. For severe TLL (Patients 8, 9 and 10), GCEM is the best method followed by MS4MS and then STREM.

The parameters of our methods were optimized using the Brainweb database, which have sequences are similar to the MNI database. This may explain the good results obtained for the 8 patients with lowest lesion load. Three patients of the MNI database have higher TLL (20, 34 and 48 cm^3) than the severe Brainweb phantom (10 cm^3), as such the parameters obtained for our methods are not suited for these patients.

7.4.3 Towards an Automatic Selection of Parameters

In the last section, we have pointed out the limitation of our segmentation methods in images from patients with high lesion loads. The Brainweb database does not provide images with TLL higher than 10 cm^3 which limits our previous search of parameters.

We observe that our methods under estimate the TLL of the patients with high

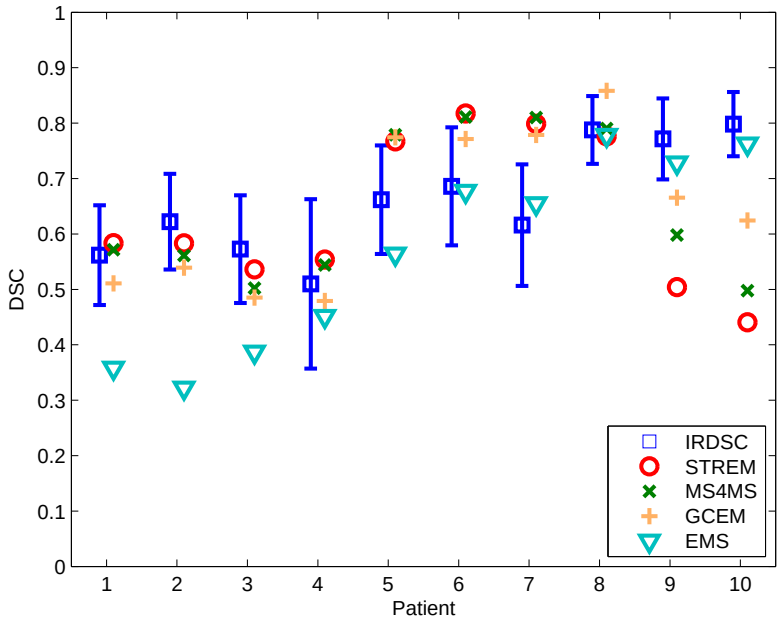


Figure 7.12: DSC results for MNI database.

Method	Fixed Parameters	High TLL parameters
STREM	$p_{\text{maha}} = 0.35,\ p_{\text{rules}} = 5e-4$	$p_{\text{maha}} = 0.40,\ p_{\text{rules}} = 5e-3$
MS4MS	$p_{\text{maha}} = 0.35,\ p_{\text{rules}} = 1e-3$	$p_{\text{maha}} = 0.45,\ p_{\text{rules}} = 1e-2$
GCEM	$S_b = 2.5,\ S_e = 3.5$	$S_b = 1.5,\ S_e = 3.0$

Table 7.4: Parameters of the three proposed segmentation methods for the two-step approach.

lesion loads (Figure 7.13). To obtain more accurate segmentation, we propose a two-step approach to adapt our parameters to the lesion load of each patient. In the first step, we process the patient with the proposed parameters and compute the TLL. If the estimated TLL is higher than a certain threshold TLL, the patient is recomputed with another set of parameters optimized for high lesion loads (Table 7.4).

We first perform the segmentation using STREM because is the fastest method. If the TLL estimated in this first segmentation is higher than $10cm^3$, we perform the segmentation a second time using the parameters for high lesion loads using the selected method (STREM, MS4MS or GCEM). In this approach, the estimation of the NABT is only performed the first time thus there is no significantly increase in the processing time.

The results obtained with this two-step approach are shown in Figure 7.14. We observe that the three patients with the highest lesion loads now obtain DSC values in the range of the experts scores.

A complete validation of this two-step approach requires a larger database to evaluate both the parameters for high lesion loads and the TLL threshold. Any other method to obtain an estimation of the lesion load could be integrated in this

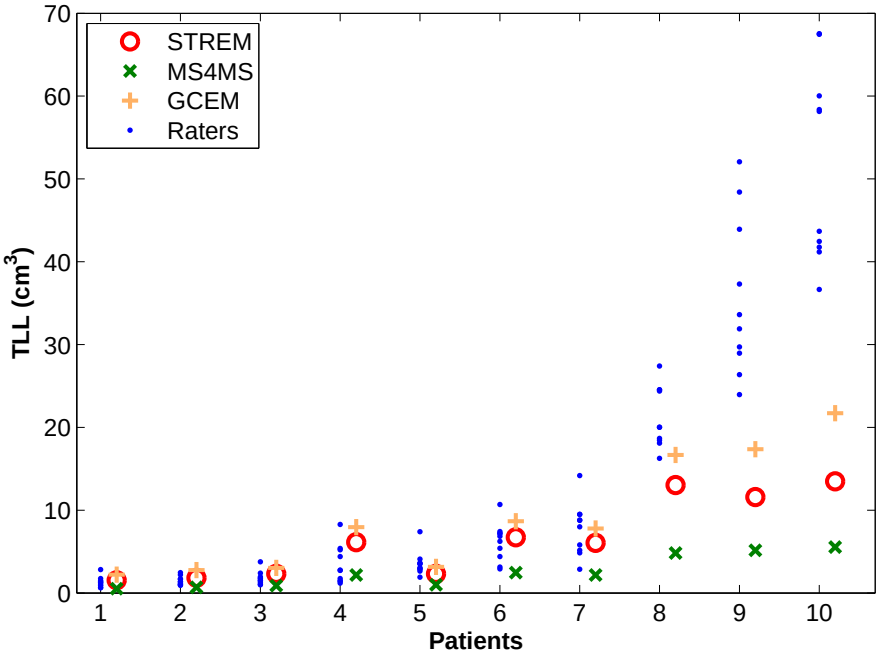


Figure 7.13: TLL obtained by the experts and by our automatic methods.

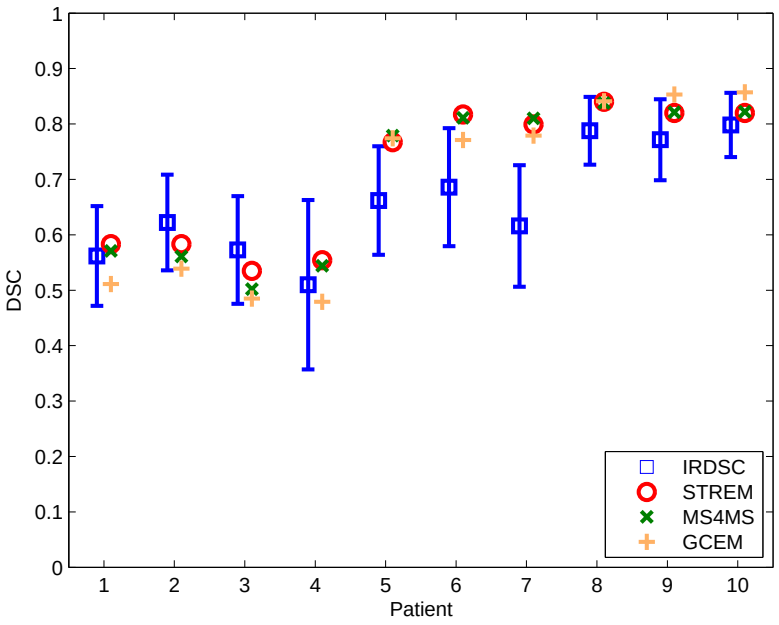


Figure 7.14: DSC results for MNI database using our two-step approach in order to adapt our parameter for the TLL.

approach.

7.4.4 Visual Comparison

Figure 7.15 shows a slice of the images of patient 4 and the results for the four automatic methods employed. We observe how the segmentation given by EMS has many false positives. There are many incorrectly detected lesions of one or two voxels even though the spatial information given by the Markov random field should limit this effect.

The three proposed method obtain similar segmentation results. We observe that there are some false positives due to the partial volumes found between the white matter and the CSF. We tried to model these partial volumes using series of Gaussians [Dugas-Phocion 2006], but this did not results in improvements in the segmentation.

STREM uses no spatial information in the segmentation and therefore we can observe in the lesion in the bottom left in Figure 7.15 how some voxels in the middle of the lesion were incorrectly segmented, contrary to MS4MS and GCEM.

7.4.5 Conclusion

The MNI database made it possible to a complete validation framework to evaluate the MS lesions segmentation. The multiple manual segmentations permitted the quantification of inter-expert variability and evaluate our segmentation methods.

We employed two different methods to evaluate our segmentation methods. Recently, STAPLE was developed to evaluate segmentation methods without a ground truth, but conclusions obtained using this method were limited. The specificity is MS lesions segmentation is not a very descriptive measure and all methods obtained higher specificities than 99%. The relation between sensitivity and specificity is not clear and in the majority of comparisons ended with one method being more specific but less sensitive than the other. In this respect, we observed that our methods were less sensitives but more specific than EMS.

We performed a second evaluation using a Voting Silver Standard. DSC is a well-known measure and we employed it in our comparison. This evaluation is more intuitive as there is only one measure for each method. Our methods obtain results comparable with experts' segmentations, when the lesion load is under $20cm^3$. We proposed a simple method to overcome with this limitation but we require more patients to evaluate this approach.

As shown in Table 7.5, our methods obtain similar DSC to those of other proposed methods in the literature. Each method was evaluated with a different database of clinical images and thus the comparison cannot be conclusive. In an effort to facilitate the comparison of algorithms, we participated in a segmentation challenge (Section 2.4.3) where STREM finished fourth of the nine methods proposed.

7.5 Conclusion

In this chapter, we described the evaluation metrics employed and evaluated our automatic methods.

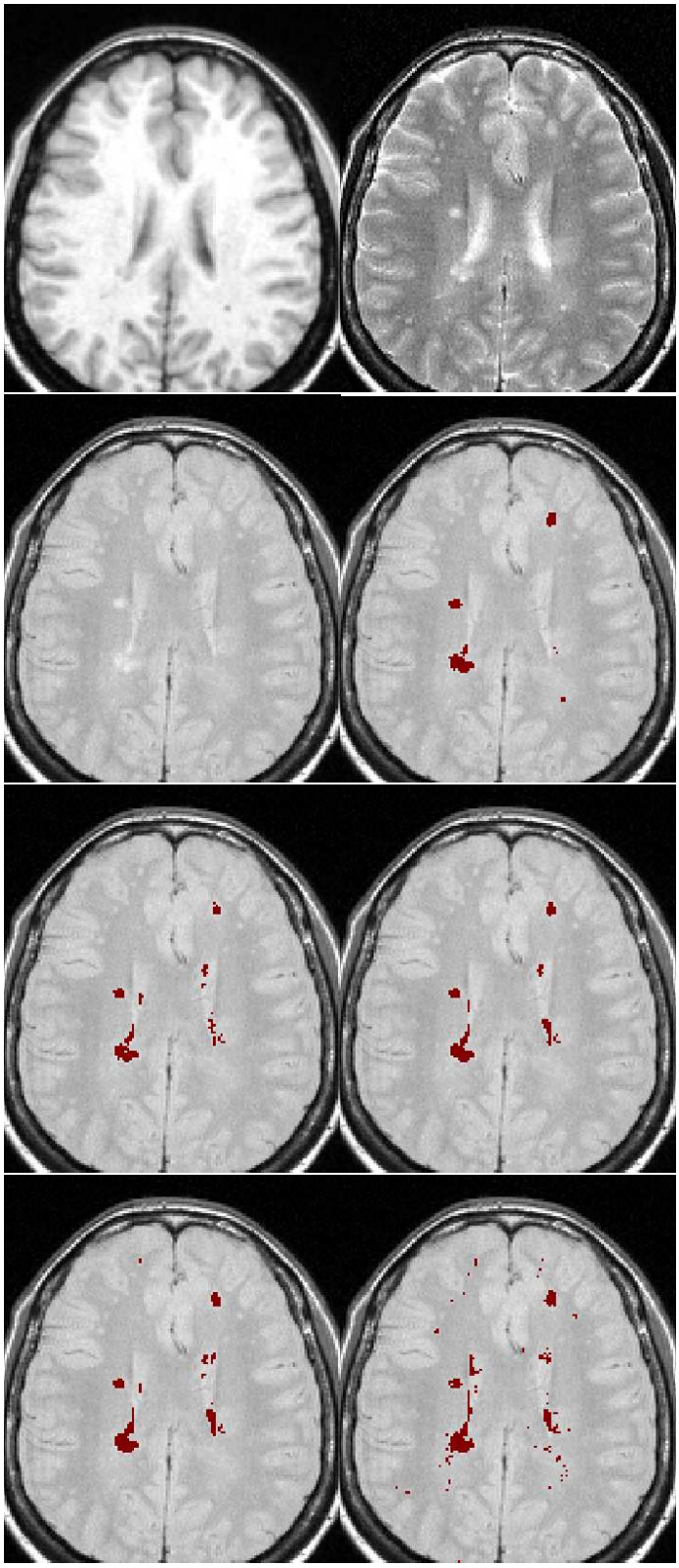


Figure 7.15: Segmentation results of patient 4 in the MNI database. From top to bottom and left to right: T1-w and T2-w; PD and voting silver standard (VSS); STREM and MS4MS; GCEM and EMS.

Method	Average DSC
STREM	0.64
MS4MS	0.65
GCEM	0.65
adaptive STREM	0.71
adaptive MS4MS	0.71
adaptive GCEM	0.69
[Van Leemput et al. 2001]	0.51
[Zijdenbos et al. 2002]	0.60
[Harmouche et al. 2006]	0.61*
[Shiee et al. 2009]	0.63
[Khayati et al. 2008]	0.75
[Sajja et al. 2006]	0.78

Table 7.5: Comparison of DSC for the lesion segmentation between different methods. (* The value is obtained using the κ index instead of DSC)

We compared our segmentation methods using synthetic images. MS4MS and GCEM obtained higher DSC than STREM when the intensity inhomogeneity and the noise was increased. This justify the inclusion of spatial information of those two methods.

We also evaluated our methods in two databases with completely opposite results. Our methods showed a good agreement with manual segmentation of the MNI database but low agreement with the manual segmentation of the Rennes database. Both databases have many differences, which may explain the differences obtained.

The definition of lesions in both databases was different. In the MNI database manual segmentation was performed in T2-w and PD images and in the Rennes database it was performed on FLAIR. The contours on T2-w images and FLAIR differ what explains the under segmentation found in the Rennes database. FLAIR images showed more artifacts than other images and these artifacts should be taken into account to reduce the number of false positives.

All three methods obtains similar results in the clinical images but the Brainweb database show some differences. In overall, GCEM is more robust in presence of noise and inhomogeneity than STREM and faster than MS4MS, and has the advantage that it can be semi-automatically refined if necessary.

Conclusion

This thesis studies the automatic methods used in the segmentation of lesions on patients with MS. The segmentation of MS lesions plays an important role in clinical trials and the automatic segmentation of these lesions will provide a robust and reproducible biomarker for clinical trials.

We began with a brief introduction to MS and the application of MRI in the diagnosis, treatment and follow-up of the disease. We showed the importance of segmenting MS lesions and described the methods employed in this task. Manual and semi-automatic segmentation methods are currently used in clinical trials, but automatic methods should eliminate user dependence, thus eliminating the variability of these methods.

We then presented our workflow for the segmentation of MS lesions and proposed various automatic segmentation methods. Methods are based on the work of Laure Aït-Ali [Aït-Ali 2006] who introduced the trimmed likelihood estimator (TLE) for the estimation of NABT model on the segmentation method STREM.

We proposed MS4MS, where voxels are first classified into homogeneous regions using spatial information, which are then classified into lesions or other tissues. Our assumption was that the classification of a region should be more robust and accurate than the classification of a single voxel.

GCEM was born of the collaboration with Jeremy Lecoeur, who worked in the semi-automatic segmentation of MRI images [Lecoeur 2010]. In practice, the automatic segmentation is always verified by an expert who validates the segmentation and edits it when necessary. Thanks to the graph cut approach, we have integrated both an automatic and a semi-automatic segmentation method in order to simplify verification by the expert.

To conclude, the three methods were validated using synthetic and clinical images from two different centers. It is necessary to validate methods in different conditions in order to measure the performance and evaluate the limitations of the algorithms used.

Contributions

We can differentiate two main contributions of this work:

- We proposed two new methods for the segmentation of MS lesions. First, we studied the behavior of STREM and proposed several improvements. Then we proposed to different approaches so as to include spatial information and we showed how those approaches are more robust to noise and inhomogeneity

than STREM. Another advantage of GCEM is that it offers the possibility of subsequently refining the segmentation.

- We proposed a common validation process for MS lesion segmentation methods consisting of three steps: studying the parameters of the algorithm, evaluating the results on synthetic images and validating the method on clinical images. This process has been applied for the three methods described in this document. In each chapter, we analyzed each parameter and studied how its value can be optimized. Then, all methods were compared using synthetic images to measure robustness to noise and inhomogeneity. Finally, we compared the methods with manual segmentation using two clinical databases.

Limitations

Definition of lesions

The validation showed contradictory results. Our methods obtained similar scores to manual segmentation on one clinical database, but the results for the other clinical database were poor. The two databases came from different sites, had different MR protocols and probably most importantly, had different definitions of lesions.

The variability shown at the MICCAI workshop (see Section 2.4.3) is an example of the high level of variability existing in the definition of lesions between MS centers. An international panel of neurologists and neuroradiologists must agree on a common definition of MS lesion in MRI in order to reduce this variability. Without agreement, the use of the same segmentation method across different centers will remain a problem.

Global *vs.* Local information

Two proposed methods merge global and local information. We demonstrated how the inclusion of local information by means of mean shift and graph cut algorithms improved robustness to noise and inhomogeneity. Lesions are heterogeneous and can be more or less hyper-intense; using the global definition of hyper-intensity causes some lesions to be missed. Using a local definition of hyper-intensity may reduce this problem.

Segmentation workflow

In order to use the segmentation methods, the latter must be included in a segmentation workflow. In Chapter 2, we proposed a segmentation workflow, but the validation of this workflow was limited to three subjects. Validating the workflow would require each step and its position in the workflow to be analyzed in a much more complex validation process.

Perspectives

Methodological perspectives

Our methods detected many false positives in FLAIR images due to hyper-intensities in the brain ventricles because the difference between these artifacts and periventricular lesions is subtle. Applying the inclusion rules concerning the shape of lesions may help to resolve this issue.

Although we performed some tests with partial volume modeling, we used three Gaussians only to model the NABT. The use of more complex models or local estimation approaches [Scherrer et al. 2007] may improve the estimation of the NABT model.

Validation perspectives

Performing the validation process using multiple manual segmentations provides more information about the accuracy of the algorithm. Our priority would therefore be to obtain more manual segmentations from the Rennes database in order to be able to offer a more comprehensive evaluation and understand the relatively poor results yielded by our methods.

We consider that the validation process also requires the comparison of segmentation methods. With this in mind, we would like to apply our two methods of segmentation the images of the MICCAI workshop mentioned in Section 2.4.3. Although there are some limitations in the comparison done by this workshop, we consider that this comparison will provide us with interesting and rather impartial information about our algorithms and will allow a comparison with other recent methods.

The next level of validation consists in the use of the segmentation methods in large clinical trials in order to test the robustness of the method and study the correlation of the lesions with the evolution of the disease. Another area requiring study is the reproducibility of the segmentation results, which would require a special protocol whereby patient images are acquired several times in a short period of time in order to measure the robustness of the segmentation.

Application perspectives

The University Hospital Pontchaillou (Rennes, France) is currently taking part in a multi-centric project where the use of the new contrast agent ultra-small superparamagnetic Iron oxide (USPIO) in MS is being evaluated. Our objective is to provide an automatic and reliable tool to segment MS lesions. In addition, new developments may be necessary in order to include the images where USPIO is used.

Another important aspect is the distribution of the segmentation methods. Our segmentation workflow is being included in NeuroLOG. NeuroLOG is an ANR project involving the development of a grid-based network in order to share clinical data and process workflows among several French centers, making the heterogeneity of data and workflows transparent to the user. This project will simplify the usability of our workflow in other MS research centers in France.

In addition, we are integrating our workflow on our own website. Similarly to other methods developed by our team, we will propose an online service[4] where any user can upload images that will be processed in our servers and sent back by email.

[4] http://www.irisa.fr/visages/benchmarks/

Publications

International Peer-reviewed Conferences

- García-Lorenzo, D., Lecoeur, J., Arnold, D. L., Collins, D. L., and Barillot, C. (2009). **Multiple Sclerosis Lesion Segmentation Using an Automatic Multimodal Graph Cuts.** *In Medical Image Computing and Computer-Assisted Intervention MICCAI 2009, volume 5762, pages 584–591.*

International Workshops

- García-Lorenzo, D., Prima, S., Collins, D. L., Arnold, D. L., Morrissey, S. P., and Barillot, C. (2008). **Combining Robust Expectation Maximization and Mean Shift algorithms for Multiple Sclerosis Brain Segmentation.** *In MICCAI Workshop in Medical Image Analysis for Multiple Sclerosis (MIAMS), pages 82–91, New York,USA.*

- García-Lorenzo, D., Prima, S., Morrissey, S. P., and Barillot, C. (2008). **A robust Expectation-Maximization algorithm for Multiple Sclerosis lesion segmentation**. *In IJ - 2008 MICCAI Workshop - MS Lesion Segmentation, New York,USA.*

- García-Lorenzo, D., Prima, S., Parkes, L., Ferré, J.-C., Morrissey, S. P., and Barillot, C.(2008). **The impact of processing workflow in performance of automatic white matter lesion segmentation in Multiple Sclerosis.** *In MICCAI Workshop in Medical Image Analysis for Multiple Sclerosis (MIAMS), pages 104–112, New York,USA.*

Abstract Conferences

- García-Lorenzo, D., Prima, S., Parkes, L., Ferré, J.-C., Roberts, N., Gauvrit, J.-Y., Edan, G., Morrissey, S. P., and Barillot, C. (2009). **MRI Brain segmentation combining robust Expectation Maximization (mEM) and Mean Shift (MeS) for Multiple Sclerosis (MS).** *In France-Germany Meeting on MS.*

- García-Lorenzo, D., Prima, S., Collins, L., Arnold, D., Barillot, C., and Morrissey, S. P. (2008). **Multiple Sclerosis (MS) Brain segmentation combining robust Expectation Maximization (EM) and Mean Shift (MeS).** *In 25th Annual Meeting of the ESMRMB (2008 Congress).*

- García-Lorenzo, D., Prima, S., Ferré, J.-C., Parkes, L., Gauvrit, J.-Y., Morrissey, S., and Barillot, C. (2008). **Quantitative evaluation of intensity inhomogeneity correction algorithms for Multiple Sclerosis.** *In Computer Assisted Radiology and Surgery 2008, volume 3, pages S402–S403.*

- García-Lorenzo, D., Prima, S., Parkes, L., Ferre, J.-C., Roberts, N., Gauvrit, J., Edan, G., Barillot, C., and Morrissey, S. (2008). **Performance of automatic white matter lesion segmentation in multiple sclerosis depends on choice of image processing workflow.** *In 18th European Neurological Society Meeting, Nice,France.*

- García-Lorenzo, D., Prima, S., Parkes, L., Ferré, J.-C., Roberts, N., Gauvrit, J.-Y., Edan, G., Morrissey, S. P., and Barillot, C. (2008). **Performance of automatic white matter MRI lesion segmentation in multiple sclerosis depends on choice of image processing workflow.** *In Franco-UK Meeting on MS.*

Acronyms

BET Brain Extraction Tool

BP Breakdown Point

BW Brainweb

CLV Coefficient of Lesion Variation

CNS Central Nervous System

CSF Cerebrospinal Fluid

CV Coefficient of Variation

CVJ Coefficient of Joint Variation

DSC Dice Similarity Coefficient

EDSS Extended Disability Status Scale

EM Expectation Maximization algorithm

EMS Expectation-Maximization Segmentation

FLAIR FLuid Attenuating Inversion Recovery

FSS Fuctional System Scale

GC Graph Cut

GCEM Graph Cut with an Expectation Maximization initialization

GM Gray Matter

GMM Gaussian Mixture Model

IIH Intensity Inhomogeneity

MICCAI Medical Image Computing and Computer-Assisted Intervention

MLE Maximum Likelihood Estimator

MNI Montreal Neurological Institute

MRI Magnetic Resonance Imaging

MS Multiple Sclerosis

MS4MS Mean Shift for Multiple Sclerosis

MeS Mean Shift

NABT Normal Appearing Brain Tissues

NAWM Normal Appearing White Matter

NLM Non-Local Means

NMR Nuclear Magnetic Resonance

OO Optimistic Overlap

PD Proton Density

PPMS Primary Progressive MS

PRMS Primary Relapsing MS

PV Partial Volumes

ROI Region of interest

RRMS Relapsing Remitting MS

SPMS Secondary Progressive MS

SSS STAPLE Silver Standard

STAPLE Simultaneous Truth Performance Level Experts

STREM Spatio Temporal Robust Expectation Maximization

TLE Trimmed Likelihood Estimator

VSS Voting Silver Standard

WM White Matter

Bibliography

Admasu, F., Al-Zubi, S., Toennies, K., Bodammer, N., and Hinrichs, H. (2003). Segmentation of multiple sclerosis lesions from MR brain images using the principles of fuzzy-connectedness and artificial neuron networks. In *International Conference on Image Processing.*, volume 2, pages II–1081–4 vol.3.

Admiraal-Behloul, F., van den Heuvel, D., Olofsen, H., van Osch, M., van der Grond, J., van Buchem, M., and Reiber, J. (2005). Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *NeuroImage*, 28(3):607–617.

Aït-Ali, L. (2006). *Analyse spatio-temporelle pour le suivi de structures évolutives en imagerie cérébrales multi–séquences.* PhD thesis, Université de Rennes I.

Aït-Ali, L. S., Prima, S., Hellier, P., Carsin, B., Edan, G., and Barillot, C. (2005). STREM: a robust multidimensional parametric method to segment MS lesions in MRI. *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'2005*, 8(Pt 1):409–416.

Akselrod-Ballin, A., Galun, M., Basri, R., Brandt, A., Gomori, M. J., Filippi, M., and Valsasina, P. (2006). An Integrated Segmentation and Classification Approach Applied to Multiple Sclerosis Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1122–1129.

Alonge, F., Ardizzone, E., and Pirrone, R. (2001). Neural Classification of Multiple Sclerosis Lesions in MR Images. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 5(4):228–233.

Anbeek, P., Vincken, K., and Viergever, M. (2008). Automated MS-Lesion Segmentation by K-Nearest Neighbor Classification. *MS Lesion Segmentation (MICCAI 2008 Workshop).*

Anbeek, P., Vincken, K. L., van Osch, M. J. P., Bisschops, R. H. C., and van der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage*, 21(3):1037–1044.

Ardizzone, E., Pirrone, R., and Gambino, O. (2003). Automatic segmentation of MR images based on adaptive anisotropic filtering. In *12th International Conference on Image Analysis and Processing*, pages 283–288.

Ardizzone, E., Pirrone, R., Gambino, O., and Peri, D. (2002). Two channels fuzzy c-means detection of multiple sclerosis lesions in multispectral MR images. In

International Conference on Image Processing, volume 2, pages II–345–II–348 vol.2.

Arnold, J. B., Liow, J. S., Schaper, K. A., Stern, J. J., Sled, J. G., Shattuck, D. W., Worth, A. J., Cohen, M. S., Leahy, R. M., Mazziotta, J. C., and Rottenberg, D. A. (2001). Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *Neuroimage*, 13(5):931–943.

ARSEP (2010). Website of the Assotiation pour la Recherche de la Esclerose en Plaques (ARSEP). [Online; accessed 03-February-2010].

Ashton, E. A., Takahashi, C., Berg, M. J., Goodman, A., Totterman, S., and Ekholm, S. (2003). Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. *Journal of Magnetic Resonance Imaging*, 17:300–308.

Atkins, M. S., Drew, M. S., and Tauber, Z. (2000). Towards automatic segmentation of MS lesions in PD/T2 MR images. In Hanson, K. M., editor, *Medical Imaging 2000: Image Processing*, volume SPIE Conference, pages 800–809.

Aubert-Broche, B., Evans, A. C., and Collins, L. (2006). A new improved version of the realistic digital brain phantom. *NeuroImage*, 32(1):138–145.

Barkhof, F., Filippi, M., Miller, D. H., Scheltens, P., Campi, A., Polman, C. H., Comi, G., Adèr, H. J., Losseff, N., and Valk, J. (1997). Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain*, 120:2059–2069.

Bermel, R. A. and Bakshi, R. (2006). The measurement and clinical relevance of brain atrophy in multiple sclerosis. *The Lancet Neurology*, 5(2):158 – 170.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575.

Blonda, P., Satalino, G., Baraldi, A., and De Blasi, R. (1998). Segmentation of multiple sclerosis lesions in MRI by fuzzy neural networks: FLVQ and FOSART. In *Fuzzy Information Processing Society - NAFIPS, 1998 Conference of the North American*, pages 39–43.

Boudraa, A.-O., Dehak, S. M. R., Zhu, Y.-M., Pachai, C., Bao, Y.-G., and Grimaud, J. (2000). Automated segmentation of multiple sclerosis lesions in multispectral MR imaging using fuzzy clustering. *Computers in Biology and Medicine*, 30(1):23–40.

Bouix, S., Martín-Fernández, M., Ungar, L., Nakamura, M., Koo, M.-S., McCarley, R. W., and Shenton, M. E. (2007). On evaluating brain tissue classifiers without a ground truth. *NeuroImage*, 36(4):1207–1224.

Boykov, Y. and Funka-Lea, G. (2006). Graph cuts and efficient N-D images segmentation. *International Journal of Computer Vision*, 70(2):109–131.

Boykov, Y. and Jolly, M.-P. (2000). Interactive organ segmentation using graph cuts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Lecture Notes in Computer Science, pages 276–286. Springer-Verlag.

Boykov, Y. and Jolly, M.-P. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *International Conference on Computer Vision*, volume 1, pages 105–112.

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137.

Bricq, S., Collet, C., and Armspach., J. (2008). MS Lesion Segmentation based on Hidden Markov Chains. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Buades, A., Coll, B., and Morel, J. M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530.

Cárdenes, R., Warfield, S. K., Macías, E. M., Santana, J. A., and Ruiz-Alzola, J. (2003). An Efficient Algorithm for Multiple Sclerosis Segmentation from Brain MRI. In *EUROCAST*, pages 542–551.

Carreira-Perpinan, M. (2006). Acceleration Strategies for Gaussian Mean-Shift Image Segmentation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 1160–1167.

Carreira-Perpinan, M. (2007). Gaussian Mean-Shift Is an EM Algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(5):767–776.

Charcot, J. (1868). Histologie de la sclérose en plaques. *Gazette des hopitaux*, 41:554–555.

Chard, D. T., Griffin, C. M., Parker, G. J. M., Kapoor, R., Thompson, A. J., and Miller, D. H. (2002). Brain atrophy in clinically early relapsing-remitting multiple sclerosis. *Brain*, 125(2):327–337.

Charil, A., Zijdenbos, A. P., Taylor, J., Boelman, C., Worsley, K. J., Evans, A. C., and Dagher, A. (2003). Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: application to 452 patient data sets. *Neuroimage*, 19(3):532–544.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799.

Cocosco, C., Kwan, K., and Evans, R. (1997). BrainWeb: online interface to a 3-D MRI simulated brain database. *Neuroimage*, 5(4):part 2/4, S425.

Collins, D., Neelin, P., Peters, T. M., and Evans, A. C. (1994). Automatic 3D Intersubject Registration of MR Volumetric Data in Standardized Talairach Space. *Journal of Computer Assisted Tomography*, 18:192–205.

Collins, D., Zijdenbos, A., Kollokian, V., Sled, J., Kabani, N., Holmes, C., and Evans, A. (Jun 1998). Design and construction of a realistic digital brain phantom. *Medical Imaging, IEEE Transactions on*, 17(3):463–468.

Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619.

Compston, A., McDonald, I. R., Noseworthy, J., Lassmann, H., Miller, D. H., Smith, K. J., Wekerle, H., and Confavreux, C. (2006). *McAlpine's Multiple Sclerosis*. Churchill Livingstone, fourth edition.

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., and Barillot, C. (2008). An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images. *Medical Imaging, IEEE Transactions on*, 27(4):425–441.

Cuadra, M., Cammoun, L., Butz, T., Cuisenaire, O., and Thiran, J.-P. (2005). Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *Medical Imaging, IEEE Transactions on*, 24(12):1548–1565.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Desolneux, A., Moisan, L., and More, J.-M. (2003). A grouping principle and four applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(4):508–513.

Dietrich, O., Raya, J. G., Reeder, S. B., Ingrisch, M., Reiser, M. F., and Schoenberg, S. O. (2008). Influence of multichannel combination, parallel imaging and other reconstruction techniques on mri noise characteristics. *Magnetic Resonance Imaging*, 26(6):754–762.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wuley, 2nd edition.

Dugas-Phocion, G. (2006). *Segmentation d'IRM Cérébrales Multi–Séquences et Application à la Sclérose en Plaques*. PhD thesis, École des Mines de Paris.

Dugas-Phocion, G., Gonzalez, M., Lebrun, C., Chanalet, S., Bensa, C., Malandain, G., and Ayache, N. (2004a). Hierarchical segmentation of multiple sclerosis lesions in multi-sequence MRI. In *Biomedical Imaging: Macro to Nano. IEEE International Symposium on*, volume 1, pages 157–160.

Dugas-Phocion, G., González Ballester, M. A., Malandain, G., Lebrun, C., and Ayache, N. (2004b). Improved em-based tissue segmentation and partial volume effect quantification in multi-sequence brain MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 3216/2004 of *Lecture Notes in Computer Science*, pages 26–33, Saint-Malo, France. Springer.

Dyment, D. A., Ebers, G. C., and Sadovnick, A. D. (2004). Genetics of multiple sclerosis. *The Lancet Neurology*, 3(2):104 – 110.

Ebers, G., Bulman, D., Sadovnick, A., Paty, D., Warren, S., Hader, W., Murray, T., Seland, T., Duquette, P., Grey, T., and et al. (1986). A population-based study of multiple sclerosis in twins. *N Engl J Med*, 315(26):1638–1642.

Esneault, S., Hraiech, N., Delabrousse, E., and Dillenseger, J.-L. (2007). Graph cut liver segmentation for interstitial ultrasound therapy. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 29, pages 5247–5250.

Fennema-Notestine, C., Ozyurt, I. B., Clark, C. P., Morris, S., Bischoff-Grethe, A., Bondi, M. W., Jernigan, T. L., Fischl, B., Segonne, F., Shattuck, D. W., Leahy, R. M., Rex, D. E., Toga, A. W., Zou, K. H., and Brown, G. G. (2006). Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. *Human Brain Mapping*, 27(2):99–113.

Filippi, M. and Rocca, M. A. (2005). *MR Imaging in White Matter Diseases of the Brain and Spinal Cord*, chapter Multiple Sclerosis: Other MR Techniques, pages 225–240. Springer Berlin Heidelberg.

Filippi, M., Rocca, M. A., Gasperini, C., Sormani, M. P., Bastianello, S., Horsfield, M. A., Pozzilli, C., and Comi, G. (1999). Interscanner Variation in Brain MR Lesion Load Measurements in Multiple Sclerosis Using Conventional Spin-Echo, Rapid Relaxation-Enhanced, and Fast-FLAIR Sequences. *Americal Journal of Neuroradiology*, 20(1):133–137.

Filippi, M., Rovaris, M., Sormani, M. P., Horsfield, M. A., Rocca, M. A., Capra, R., Prandini, F., and Comi, G. (1998). Intraobserver and interobserver variability in measuring changes in lesion volume on serial brain MR images in multiple sclerosis. *American Journal Neuroradiology*, 19(4):685–687.

Fischer, J. S., Rudick, R. A., Cutter, G. R., Reingold, S. C., and Force, N. M. S. C. O. A. T. (1999). The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment. *Multiple Sclerosis*, 5(4):244–250.

Freifeld, O., Greenspan, H., and Goldberger, J. (2007). Lesion Detection in Noisy MR Brain Images Using Constrained GMM and Active Contours. In *International Symposium on Biomedical Imaging: From Nano to Macro*, pages 596–599.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40.

García-Lorenzo, D., Lecoeur, J., Arnold, D. L., Collins, D. L., and Barillot, C. (2009). Multiple Sclerosis Lesion Segmentation Using an Automatic Multimodal

Graph Cuts. In *Medical Image Computing and Computer-Assisted Intervention*, volume 5762 of *Lecture Notes in Computer Science*, pages 584–591.

García-Lorenzo, D., Prima, S., Collins, D. L., Arnold, D. L., Morrissey, S. P., and Barillot, C. (2008a). Combining Robust Expectation Maximization and Mean Shift algorithms for Multiple Sclerosis Brain Segmentation. In *MICCAI Workshop in Medical Image Analysis for Multiple Sclerosis (MIAMS)*, pages 82–91, New York,USA.

García-Lorenzo, D., Prima, S., Ferré, J.-C., Parkes, L., Gauvrit, J.-Y., Morrissey, S., and Barillot, C. (2008b). Quantitative evaluation of intensity inhomogeneity correction algorithms for Multiple Sclerosis. In *Computer Assisted Radiology and Surgery 2008*, volume 3, pages S402–S403.

García-Lorenzo, D., Prima, S., Morrissey, S. P., and Barillot, C. (2008c). A robust Expectation-Maximization algorithm for Multiple Sclerosis lesion segmentation. In *IJ - 2008 MICCAI Workshop - MS Lesion Segmentation*, New York,USA.

García-Lorenzo, D., Prima, S., Parkes, L., Ferré, J.-C., Morrissey, S. P., and Barillot, C. (2008d). The impact of processing workflow in performance of automatic white matter lesion segmentation in Multiple Sclerosis. In *MICCAI Workshop in Medical Image Analysis for Multiple Sclerosis (MIAMS)*, pages 104–112, New York,USA.

Georgescu, B., Shimshoni, I., and Meer, P. (2003). Mean shift based clustering in high dimensions: a texture classification example. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 456–463vol.1.

Geurts, J. J. G., Bö, L., Pouwels, P. J. W., Castelijns, J. A., Polman, C. H., and Barkhof, F. (2005). Cortical lesions in multiple sclerosis: combined postmortem MR imaging and histopathology. *American Journal of Neuroradiology*, 26(3):572–577.

Geusebroek, J.-M., van den Boomgaard, R., Smeulders, A., and Dev, A. (2000). Color and space : The spatial structure of color images. In *IEEE European Conference on Computer Vision*, volume 1842 of *Lecture Notes in Computer Science*, pages 331–341.

Ghazel, M., Traboulsee, A., and Ward, R. (2006). Semi-Automated Segmentation of Multiple Sclerosis Lesions in Brain MRI using Texture Analysis. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, pages 6–10.

Gilden, D. H. (2005). Infectious causes of multiple sclerosis. *The Lancet Neurology*, 4(3):195–202.

Gold, R. and Hartung, H.-P. (2005). Towards individualised multiple-sclerosis therapy. *The Lancet Neurology*, 4(11):693–694.

Greenspan, H., Ruf, A., and Goldberger, J. (2006). Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *Medical Imaging, IEEE Transactions on*, 25(9):1233–1245.

Greig, D., Porteous, B., and Seheult, A. (1989). Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, 51(2):271–279.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G. J., Plummer, D. L., Tofts, P. S., McDonald, W. I., and Miller, D. H. (1996). Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. *Magnetic Resonance Imaging*, 14(5):495–505.

Groot, C. J. D., Bergers, E., Kamphorst, W., Ravid, R., Polman, C. H., Barkhof, F., and der Valk, P. V. (2001). Post-mortem MRI-guided sampling of multiple sclerosis brain lesions: increased yield of active demyelinating and (p)reactive lesions. *Brain*, 124(Pt 8):1635–1645.

Gudbjartsson, H. and Patz, S. (1995). The rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914.

Guttmann, C. R., Kikinis, R., Anderson, M. C., Jakab, M., Warfield, S. K., Killiany, R. J., Weiner, H. L., and Jolesz, F. A. (1999). Quantitative follow-up of patients with multiple sclerosis using MRI: reproducibility. *Journal of Magnetic Resonance Imaging*, 9(4):509–518.

Haacke, E. M., Brown, R. W., Thompson, M. R., and Venkatesan, R. (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss.

Harmouche, R., Collins, L., Arnold, D., Francis, S., and Arbel, T. (2006). Bayesian MS Lesion Classification Modeling Regional and Local Spatial Information. In *Pattern Recognition, International Conference on*, volume 3, pages 984–987, Los Alamitos, CA, USA. IEEE Computer Society.

Hou, Z. (2006). A Review on MR Image Intensity Inhomogeneity Correction. *International Journal of Biomedical Imaging*, pages Article ID 49515, 11 pages.

Hou, Z., Qian, W., Huang, S., Hu, Q., and Nowinski, W. L. (2007). Regularized fuzzy c-means method for brain tissue clustering. *Pattern Recognition Letters*, 28(13):1788–1794.

Jannin, P., Krupinski, E., and Warfield, S. (2006). Validation in medical image processing. *IEEE Trans Med Imaging*, 25(11):1405–1409.

Jimenez-Alaniz, J., Medina-Banuelos, V., and Yanez-Suarez, O. (2006). Data-driven brain MRI segmentation supported on edge confidence and a priori tissue information. *Medical Imaging, IEEE Transactions on*, 25(1):74–83.

Johnston, B., Atkins, M., Mackiewich, B., and Anderson, M. (1996). Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. *Medical Imaging, IEEE Transactions on*, 15(2):154–169.

Kamber, M., Shinghal, R., Collins, D., Francis, G., and Evans, A. (1995). Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. *Medical Imaging, IEEE Transactions on*, 14(3):442–453.

Keiper, M., Grossman, R., Hirsch, J., Bolinger, L., Ott, I., Mannon, L., Langlotz, C., and Kolson, D. (1998). MR identification of white matter abnormalities in multiple sclerosis: a comparison between 1.5 T and 4 T. *AJNR Am J Neuroradiol*, 19(8):1489–1493.

Kezele, I., Chen, J., Arnold, D., and Collins, D. (2007). The relation of focal white matter signal abnormality and focal volume loss in multiple sclerosis. *Multiple Sclerosis*, 13(6):809–813.

Khayati, R., Vafadust, M., Towhidkhah, F., and Nabavi, M. (2008). Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and markov random field model. *Computers in Biology and Medicine*, 38(3):379–390.

Kidd, D., Barkhof, F., McConnell, R., Algra, P. R., Allen, I. V., and Revesz, T. (1999). Cortical lesions in multiple sclerosis. *Brain*, 122 (Pt 1):17–26.

Kikinis, R., Guttmann, C. R., Metcalf, D., III, W. M. W., Ettinger, G. J., Weiner, H. L., and Jolesz, F. A. (1999). Quantitative follow-up of patients with multiple sclerosis using MRI: Technical aspects. *Journal of Magnetic Resonance Imaging*, 9(4):519–530.

Kroon, D., Oort, E. V., and Slump, K. (2008). Multiple Sclerosis Detection in Multispectral Magnetic Resonance Images with Principal Components Analysis. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology*, 33(11):1444–.

Lao, Z., Shen, D., Jawad, A., Karacali, B., Liu, D., Melhem, E., Bryan, R., and Davatzikos, C. (2006). Automated segmentation of white matter lesions in 3D brain MR images, using multivariate pattern classification. In *Biomedical Imaging: Macro to Nano, 2006. 3rd IEEE International Symposium on*, pages 307–310.

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., Bryan, R. N., and Davatzikos, C. (2008). Computer-Assisted Segmentation of White Matter Lesions in 3D MR Images Using Support Vector Machine. *Academic Radiology*, 15(3):300–313.

Lassmann, H. (2005). Multiple sclerosis pathology: evolution of pathogenetic concepts. *Brain Pathology*, 15(3):217–222.

Lecoeur, J. (2010). *Segmentation d'IRM cérébrales multidimensionnelles par coupes de graphes*. PhD thesis, Université de Rennes I.

Lecoeur, J., Morissey, S., Ferré, J.-C., Arnold, D., Collins, D., and Barillot, C. (2008). Multiple Sclerosis Lesions Segmentation using Spectral Gradient and Graph Cuts. In *Proceedings of MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)*, pages 92–103.

Likar, B., Maintz, J. B., Viergever, M. A., and Pernus, F. (2000). Retrospective shading correction based on entropy minimization. *Journal of Microscience*, 197(Pt 3):285–295.

Likar, B., Viergever, M. A., and Pernus, F. (2001). Retrospective correction of MR intensity inhomogeneity by information minimization. *Medical Imaging, IEEE Transactions on*, 20(12):1398–1410.

Lin, X., Tench, C. R., Turner, B., Blumhardt, L. D., and Constantinescu, C. S. (2003). Spinal cord atrophy and disability in multiple sclerosis over four years: application of a reproducible automated technique in monitoring disease progression in a cohort of the interferon β-1a (rebif) treatment trial. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(8):1090–1094.

Lublin, F. D., Reingold, S. C., and on Clinical Trials of New Agents in Multiple Sclerosis, N. M. S. S. U. A. C. (1996). Defining the clinical course of multiple sclerosis: Results of an international survey. *Neurology*, 46(4):907–911.

Lucchinetti, C., Brück, W., Parisi, J., Scheithauer, B., Rodriguez, M., and Lassmann, H. (2000). Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. *Annals of Neurology*, 47(6):707–717.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *Medical Imaging, IEEE Transactions on*, 16(2):187–198.

Mangin, J.-F. (2000). Entropy minimization for automatic correction of intensity nonuniformity. In *Mathematical Methods in Biomedical Image Analysis. IEEE Workshop on*, pages 162–169.

Manjón, J. V., Lull, J. J., Carbonell-Caballero, J., García-Martí, G., Martí-Bonmatí, L., and Robles, M. (2007). A nonparametric MRI inhomogeneity correction method. *Medical Image Analysis*, 11(4):336–345.

Mayer, A. and Greenspan, H. (2009). An Adaptive Mean-Shift Framework for MRI Brain Segmentation. *Medical Imaging, IEEE Transactions on*, 28(8):1238–1250.

McDonald, W., Compstonn, A., Edan, G., Goodkin, D., Hartung, H., Lublin, F., McFarland, H., Paty, D., Polman, C., Reingold, S., Sandberg-Wollheim, M., Sibley, W., Thompson, A., Van den Noort, S., Weinshenker, B., and Wolinsky, J. (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, 50:121–127.

McRobbie, D. W., Moore, E. A., Graves, M. J., and Prince, M. R. (2003). *MRI From Picture to Proton*. Cambridge University Press.

Meer, P. and Georgescu, B. (2001). Edge detection with embedded confidence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(12):1351–1365.

Meier, D. S., Weiner, H. L., Khoury, S. J., and Guttmann, C. R. G. (2004). Magnetic resonance imaging surrogates of multiple sclerosis pathology and their relationship to central nervous system atrophy. *Journal of Neuroimaging*, 14(s3):46S–53S.

Meilă, M. and Heckerman, D. (2001). An Experimental Comparison of Model-Based Clustering Methods. *Machine Learning*, 42(1-2):9–29.

Miller, D. H. (2004). Biomarkers and surrogate outcomes in neurodegenerative disease: lessons from multiple sclerosis. *NeuroRx*, 1(2):284–294.

Miller, D. H., Grossman, R. I., Reingold, S. C., and McFarland, H. F. (1998). The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain*, 121 (Pt 1):3–24.

Molyneux, P., Tubridy, N., Parker, G., Barker, G., MacManus, D., Tofts, P., Moseley, I., and Miller, D. (1998). The effect of section thickness on MR lesion detection and quantification in multiple sclerosis. *AJNR Am J Neuroradiol*, 19(9):1715–1720.

Molyneux, P. D., Miller, D. H., Filippi, M., Yousry, T. A., Radü, E. W., Adèr, H. J., and Barkhof, F. (1999). Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. *Neuroradiology*, 41(12):882–888.

Montillo, A., Udupa, J. K., Axel, L., and Metaxas, D. N. (2003). Interaction between noise suppression and inhomogeneity correction in MRI. In Sonka, M. and Fitzpatrick, M. J., editors, *Medical Imaging 2003: Image Processing*, volume 5032, pages 1025–1036. SPIE.

Morra, J., Tu, Z., Toga, A., and Thompson, P. (2008). Automatic Segmentation of MS Lesions Using a Contextual Model for the MICCAI Grand Challenge. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Müller, C. H. and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, 116(2):503–519.

Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308.

Pachai, C., Zhu, Y. M., Grimaud, J., Hermier, M., Dromigny-Badin, A., Boudraa, A., Gimenez, G., Confavreux, C., and Froment, J. C. (1998). A pyramidal approach for automatic segmentation of multiple sclerosis lesions in brain MRI. *Computerized Medical Imaging and Graphics*, 22(5):399–408.

Parodi, R. C., Sardanelli, F., Renzetti, P., Rosso, E., Losacco, C., Ferrari, A., Levrero, F., Pilot, A., Inglese, M., and Mancardi, G. L. (2002). Growing Region Segmentation Software (GRES) for quantitative magnetic resonance imaging of multiple sclerosis: intra- and inter-observer agreement variability: a comparison with manual contouring method. *European Radiology*, 12(4):866–871.

Paty, D. W. and Ebers, G. C., editors (1997). *Multiple Sclerosis*. Oxford University Press.

Pham, D. and Prince, J. (1999). Adaptive fuzzy segmentation of magnetic resonance images. *Medical Imaging, IEEE Transactions on*, 18(9):737–752.

Polman, C. H., Reingold, S. C., Edan, G., Filippi, M., Hartung, H.-P., Kappos, L., Lublin, F. D., Metz, L. M., McFarland, H. F., O'Connor, P. W., Sandberg-Wollheim, M., Thompson, A. J., Weinshenker, B. G., and Wolinsky, J. S. (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Annals of Neurology*, 58(6):840–846.

Poser, C. M., Paty, D. W., Scheinberg, L., McDonald, W. I., Davis, F. A., Ebers, G. C., Johnson, K. P., Sibley, W. A., Silberberg, D. H., and Tourtellotte, W. W. (1983). New diagnostic criteria for multiple sclerosis: Guidelines for research protocols. *Annals of Neurology*, 13(3):227–231.

Prastawa, M. and Gerig, G. (2008). Automatic MS Lesion Segmentation by Outlier Detection and Information Theoretic Region Partitioning. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Prima, S., Ayache, N., Barrick, T., and Roberts, N. (2001). Maximum likelihood estimation of the bias field in MR brain images: investigating different modelings of the imaging process. In *4th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'2001*, volume 2208, pages 811–819.

Rouaïnia, M., Medjram, M., and Doghmane, N. (2006). Brain MRI segmentation and lesions detection by EM algorithm. In *Proc of World Academy of Science, Engineering and Technology*, volume 17, pages 301–304.

Rousseau, F., Blanc, F., de Seze, J., Rumbach, L., and Armspach, J.-P. (2008). An a contrario approach for outliers segmentation: Application to Multiple Sclerosis in MRI. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 9–12.

Ruan., S., Jaggi, C., Xue, J., Fadili, J., and Bloyet, D. (Dec 2000). Brain tissue classification of magnetic resonance images using partial volume modeling. *Medical Imaging, IEEE Transactions on*, 19(12):1179–1187.

Rudick, R. A., Fisher, E., Lee, J.-C., Simon, J., and Jacobs, L. (1999). Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. *Neurology*, 53(8):1698–.

Sajja, B. R., Datta, S., He, R., Mehta, M., Gupta, R. K., Wolinsky, J. S., and Narayana, P. A. (2006). Unified approach for multiple sclerosis lesion segmentation on brain MRI. *Ann Biomed Eng*, 34(1):142–151.

Santago, P. and Gage, H. (Sep 1993). Quantification of MR brain images by mixture density and partial volume modeling. *Medical Imaging, IEEE Transactions on*, 12(3):566–574.

Scherrer, B., Dojat, M., F., F., and Garbay, C. (2007). LOCUS: LOcal cooperative Unified Segmentation of MRI Brain Scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 219–227.

Schumacher, G. A., Beebe, G., Kibler, R. F., Kurland, L. T., Kurtzke, J. F., McDowell, F., Nagler, B., Sibley, W. A., Tourtellotte, W. W., and Willmon, T. L. (1965). Problems of experimental trials of therapy in multiple sclerosis: Report by the panel on the evaluation of experimental trials of therapy in multiple sclerosis. *Annals of the New York Academy of Science*, 122:552–568.

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., and Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *Neuroimage*, 22(3):1060–1075.

Shahar, A. and Greenspan, H. (2004). A probabilistic framework for the detection and tracking in time of multiple sclerosis lesions. In *Biomedical Imaging: Macro to Nano, 2004. IEEE International Symposium on*, pages 440–443Vol.1.

Shen, C. and Brooks, M. (2005). Adaptive over-relaxed mean shift. In *Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on*, volume 2, pages 575–578.

Shiee, N., Bazin, P., and Pham, D. (2008). Multiple Sclerosis Lesion Segmentation Using Statistical and Topological Atlases. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2009). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, In Press, Corrected Proof:–.

Sijbers, J., den Dekker, A., Scheunders, P., and Van Dyck, D. (1998). Maximum-likelihood estimation of rician distribution parameters. *Medical Imaging, IEEE Transactions on*, 17(3):357–361.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *Medical Imaging, IEEE Transactions on*, 17(1):87–97.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155.

Soille, P. (2008). Constrained connectivity for hierarchical image decomposition and simplification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1132–1145.

Song, Z., Tustison, N., avants, B., and Gee, J. (2006). Adaptative Graph Cuts with tissue priors for brain MRI segmentation. In *International Symposium Biomedical Imaging*, pages 762–765.

Souplet, J., Lebrun, C., Ayache, N., and Malandain, G. (2008). An Automatic Segmentation of T2-FLAIR Multiple Sclerosis Lesions. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Souplet, J.-C. (2009). *Évaluation de l'atrophie et de la charge lésionelle sur des séquences IRM de patients atteints de sclérose en plaques*. PhD thesis, Université de Nice - Sophia Antipolis.

Souplet, J.-C., Lebrun, C., Clavelou, P., Camu, W., Chanalet, S., Ayache, N., and Malandain, G. (2007). A comparative study of skull stripping methods in relapsing-remitting multiple sclerosis: Emergence of a new automatic segmentation algorithm. Congress of the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS).

Stefano, N. D., Bartolozzi, M. L., Guidi, L., Stromillo, M. L., and Federico, A. (2005). Magnetic resonance spectroscopy as a measure of brain damage in multiple sclerosis. *Journal of the Neurological Sciences*, 233(1-2):203–208. Preserve the Neuron.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., and Warfield, S. (2008). 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. *MS Lesion Segmentation (MICCAI 2008 Workshop)*.

Suri, J. S., Singh, S., and Reden, L. (2002). Computer Vision and Pattern Recognition Techniques for 2-D and 3-D MR Cerebral Cortical Segmentation (Part I): A State-of-the-Art Review. *Pattern Analysis & Applications*, 5(1):46–76.

Tan, I. L., van Schijndel, R. A., van Walderveen, M. A. A., Quist, M., Bos, R., Pouwels, P. J. W., Desmedt, P., Adèr, H. J., and Barkhof, F. (2002). Magnetic resonance image registration in multiple sclerosis: comparison with repositioning error and observer-based variability. *J Magn Reson Imaging*, 15(5):505–510.

Tortorella, C., Viti, B., Bozzali, M., Sormani, M. P., Rizzo, G., Gilardi, M. F., Comi, G., and Filippi, M. (2000). A magnetization transfer histogram study of normal-appearing brain tissue in MS. *Neurology*, 54(1):186–.

Traboulsee, A., Li, D. K., Zhao, G., and Paty, D. W. (2005). *MR Imaging in White Matter Diseases of the Brain and Spinal Cord*, chapter Conventional MRI Techniques in Multiple Sclerosis, pages 211–223. Springer Berlin Heidelberg.

Udupa, J. and Samarasekera, S. (May 1996). Fuzzy Connectedness and Object Definition: Theory, Algorithms, and Applications in Image Segmentation. *Graphical Models and Image Processing*, 58:246–261(16).

Udupa, J. K., LeBlanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., Hirsch, B. E., and Woodburn, J. (2006). A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 30(2):75–87.

Udupa, J. K., Wei, L., Samarasekera, S., Miki, Y., van Buchem, M. A., and Grossman, R. I. (1997). Multiple sclerosis lesion quantification using fuzzy-connectedness principles. *Medical Imaging, IEEE Transactions on*, 16(5):598–609.

van Ginneken, B., Heimann, T., and Styner, M. (2007). 3D Segmentation in the Clinic: A Grand Challenge. In *3D Segmentation in the Clinic: A Grand Challenge. Miccai Workshop*, pages 7–15.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., and Suetens, P. (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. *Medical Imaging, IEEE Transactions on*, 20(8):677–688.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated model-based bias field correction of MR images of the brain. *Medical Imaging, IEEE Transactions on*, 18(10):885–896.

Van Waesberghe, J. H. T. M., Kamphorst, W., De Groot, C. J. A., Van Walderveen, M. A. A., Castelijns, J. A., Ravid, R., Lycklama à Nijeholt, G. J., Van Der Valk, P., Polman, C. H., Thompson, A. J., and Barkhof, F. (1999). Axonal loss in multiple sclerosis lesions: Magnetic resonance imaging insights into substrates of disability. *Annals of Neurology*, 46(5):747–754.

Vellinga, M., Geurts, J., Rostrup, E., Uitdehaag, B., Polman, C., Barkhof, F., and Vrenken, H. (2009). Clinical correlations of brain lesion distribution in multiple sclerosis. *Journal of Magnetic Resonance Imaging*, 29(4):768–773.

Vicente, S., Kolmogorov, V., and Rother, C. (2008). Graph cut based image segmentation with connectivity priors. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Vovk, U., Pernus, F., and Likar, B. (2007). A review of methods for correction of intensity inhomogeneity in MRI. *Medical Imaging, IEEE Transactions on*, 26(3):405–421.

Warfield, S., Dengler, J., Zaers, J., Guttmann, C. R., III, W. M. W., Ettinger, G. J., Hiller, J., and Kikinis, R. (1995). Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. *Journal of Image Guided Surgery*, 1(6):326–338.

Warfield, S., Zou, K., and Wells, W. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *Medical Imaging, IEEE Transactions on*, 23(7):903–921.

Warfield, S. K., Kaus, M., Jolesz, F. A., and Kikinis, R. (2000). Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analisys*, 4(1):43–55.

Wei, X., Warfield, S. K., Zou, K. H., Wu, Y., Li, X., Guimond, A., Mugler, J. P., Benson, R. R., Wolfson, L., Weiner, H. L., and Guttmann, C. R. G. (2002). Quantitative analysis of MRI signal abnormalities of brain white matter with high

reproducibility and accuracy. *Journal of Magnetic Resonance Imaging*, 15(2):203–209.

Weldeselassie, Y. and Harmaneh, G. (2007). DT-MRI segmentation using graph cuts. In *SPIE Conference on Medical Imaging : Image Processing*, volume 6512.

Wells III, W., Grimson, W., Kikinis, R., and Jolesz, F. (1996). Adaptive segmentation of MRI data. *Medical Imaging, IEEE Transactions on*, 15(4):429–442.

Welti, D., Gerig, G., Radü, E.-W., Kappos, L., and Székely, G. (2001). Spatio-Temporal Segmentation of Active Multiple Sclerosis Lesions in Serial MRI Data. In *Information Processing in Medical Imaging*, volume 2082/2001, pages 438–445.

Wicks, D. A. G., Tofts, P. S., Miller, D. H., de Boulay, G. H., Feinstein, A., Sacares, R. P., Harvey, I., and McDonald, R. B. W. I. (1992). Volume measurement of multiple sclerosis lesions with magnetic resonance images. *Neuroradiology*, 34(6):75–479.

Wiest-Daesslé, N. (2009). *Imagerie du tenseur de diffusion pour l'étude de pathologies cérébrales*. PhD thesis, Université de Rennes 1.

Wiest-Daesslé, N., Prima, S., Morrissey, S., and Barillot, C. (2007). Validation of a new optimisation algorithm for registration tasks in medical imaging. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'2007*, pages 41–44, Washington, USA.

Wu, Y., Warfield, S. K., Tan, I. L., Wells, W. M., Meier, D. S., van Schijndel, R. A., Barkhof, F., and Guttmann, C. R. G. (2006). Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. *Neuroimage*, 32(3):1205–1215.

Yang, F., Zhu, L., and Jiang, T. (2003). White matter lesion segmentation using robust parameter estimation algorithms. In Sonka, M. and Fitzpatrick, J. M., editors, *Medical Imaging 2003: Image Processing*, volume 5032, pages 1450–1454. SPIE.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*, 20(1):45–57.

Zhu, H. and Basir, O. (2003). Automated brain tissue segmentation and MS lesion detection using fuzzy and evidential reasoning. In *Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on*, volume 3, pages 1070–1073Vol.3.

Zijdenbos, A., Dawant, B., Margolin, R., and Palmer, A. (Dec 1994). Morphometric analysis of white matter lesions in MR images: method and validation. *Medical Imaging, IEEE Transactions on*, 13(4):716–724.

Zijdenbos, A. P., Forghani, R., and Evans, A. C. (2002). Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *Medical Imaging, IEEE Transactions on*, 21(10):1280–1291.

Zivadinov, R. and Leist, T. P. (2005). Clinical-magnetic resonance imaging correlations in multiple sclerosis. *Journal of Neuroimaging*, 15(4 Suppl):10S–21S.

List of Figures

Résumé

La sclérose en plaques (SEP) atteint autour de 80.000 personnes en France. L'imagerie par résonance magnétique (IRM) est un outil essentiel pour le diagnostic de la SEP. Plusieurs bio-marqueurs sont obtenus à partir des IRM, comme le volume des lésions, et sont utilisés comme mesure dans des études cliniques en SEP, notamment pour le développement des nouveaux traitements. La segmentation manuelle des lésions est une tâche encombrante et dont les variabilités intra- et inter-expert sont grandes.

Nous avons développé une chaîne de traitement automatique pour la segmentation des lesions focales en SEP. La méthode de segmentation est basée sur l'estimation robuste d'un modèle paramétrique des intensités du cerveau qui permet de détecter les lésions comme des données aberrantes. Nous avons aussi proposé deux méthodes pour ajouter de l'information spatiale avec les algorithmes mean shift et graph cut.

Nous avons validé quantitativement notre approche en utilisant des images synthétiques et cliniques, provenant de deux centres différents pour évaluer la précision et la robustesse.

Mots-clés: segmentation, imagerie par résonance magnétique, sclérose en plaques, modèles paramétriques

Abstract

Multiple sclerosis (MS) affects around 80.000 people in France. Magnetic resonance imaging (MRI) is an essential tool for diagnosis of MS and MRI-derived surrogate markers such as MS lesion volumes are often used as measures in MS clinical trials for the development of new treatments. The manual segmentation of these MS lesions is a time-consuming task that shows high inter- and intra-rater variability.

We developed an automatic workflow for the segmentation of focal MS lesions on MRI. The segmentation method is based on the robust estimation of a parametric model of the intensities of the brain; lesions are detected as outliers to the model. We proposed two methods to include spatial information in the segmentation using mean shift and graph cut.

We performed a quantitative evaluation of our workflow using synthetic and clinical images of two different centers to verify its accuracy and robustness.

Keywords: segmentation, magnetic resonance imaging, multiple sclerosis, parametric models