Automated Segmentation of Multiple Sclerosis Lesions by Model Outlier Detection

Koen Van Leemput*, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens

***Abstract*—This paper presents a fully automated algorithm for segmentation of multiple sclerosis (MS) lesions from multispectral magnetic resonance (MR) images. The method performs intensity-based tissue classification using a stochastic model for normal brain images and simultaneously detects MS lesions as outliers that are not well explained by the model. It corrects for MR field inhomogeneities, estimates tissue-specific intensity models from the data itself, and incorporates contextual information in the classification using a Markov random field. The results of the automated method are compared with lesion delineations by human experts, showing a high total lesion load correlation. When the degree of spatial correspondence between segmentations is taken into account, considerable disagreement is found, both between expert segmentations, and between expert and automatic measurements.**

***Index Terms*—Digital brain atlas, MRI, multiple sclerosis, tissue classification.**

I. Introduction

MULTIPLE sclerosis (MS) is a common disease of young adults that primarily affects the white matter (WM) of the central nervous system. Magnetic resonance (MR) imaging is increasingly being used to assess the progression of the disease and to evaluate the effect of drug therapy, supplementing traditional neurological disability scales such as the extended disability status scale (EDSS) [1]. EDSS is heavily weighted toward locomotor disability and has substantial intrarater and interrater variability [2], [3]. Although MR measurements may show significant variability as well, they are far more sensitive and clearly reveal one important aspect of the underlying pathological process. They are, therefore, nowadays the primary outcome of preliminary clinical trials to evaluate whether a new therapy might favorably modify the evolution of the disease [2], [3]. A landmark study in this respect was the interferon beta-1b trial [4] that showed reduction in disease progression as assessed by MRI-based findings.

In clinical trials, the large number of MR images to be analyzed makes manual analysis by human experts extremely time-consuming. Furthermore, the intraobserver and interobserver variability associated with manual delineations complicates the analysis of the results, as demonstrated in the beta-interferon study [4] where there was a significant reduction in measured lesion loads in the third year due to a systematic change in the manual tracings. Also, it is not clear how a human rater combines information from multiple images when multispectral MR data are examined. Therefore, there is a need for fully automated methods for MS lesion segmentation that can analyze large amounts of multispectral MR data in a reproducible way which correlates well with expert analyses.

In this paper, we present such a method and we demonstrate its performance on MS data sets that consist of T1-, T2-, and PD-weighted MR images. The approach presented here differs from existing work in one or more of the following ways. First, rather than attempting to model MS lesions explicitly, we detect them as outliers with respect to a statistical model for normal brain MR images. Second, the method is fully automated due to the use of a brain atlas that contains information about the expected location of the major tissue types. And third, the method retrains itself automatically on each individual scan, making it adaptive to changes in pulse sequence or voxel size.

The paper is organized as follows. In Section II, we apply concepts borrowed from the robust statistics literature to our previously published method for automated bias field correction and tissue classification of normal brain MR images [5], [6]. This results in an iterative algorithm that interleaves statistical classification of the data into a number of normal tissue types, assessment of the belief for each voxel that it is an MS lesion based on its intensity and on the classification of its neighbors, and re-estimation of tissue and bias field parameters whereby MS lesions are down-weighted. In Section III, we apply the method to images drawn from an ongoing clinical trial. Section IV discusses the strengths and weaknesses of the method compared with existing work. Section V summarizes our conclusions.

II. Method

A. Background

Suppose that N L-channel samples $\boldsymbol{y}_i = [y_{i_1} \ldots y_{i_L}]^t$ with $i = 1, 2, \ldots, N$ are drawn independently from a multivariate

Manuscript received July 12, 2000; revised May 15, 2001. This work was supported by the EC-funded BIOMED-2 program under Grant BMH4-CT96-0845 (BIOMORPH) and Grant BMH4-CT98-6048 (QAMRIC), and by the Research Fund KULeuven under Grant GOA/99/05 (VHS+). The work of K. Van Leemput was supported in part by a grant from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). The work of D. Vandermeulen was supported in part by the F.W.O.-Vlaanderen under Grant 1.5.397.97. F. Maes is a Postdoctoral fellow of the Fund for Scientific Research—Flanders (F.W.O.-Vlaanderen, Belgium). The Associate Editor responsible for coordinating the review of this paper and recommending its publication was T. Taxt. *Asterisk indicates corresponding author.*

*K. Van Leemput is with the Medical Image Computing (Radiology-ESAT/PSI), Faculties of Medicine and Engineering, University Hospital Gasthuisberg, Herestraat 49, B-3000 Leuven, Belgium (e-mail: koen.vanleemput@uz.kuleuven.ac.be).

F. Maes, D. Vandermeulen, and P. Suetens are with the Medical Image Computing (Radiology-ESAT/PSI), Faculties of Medicine and Engineering, University Hospital Gasthuisberg, B-3000 Leuven, Belgium.

A. Colchester is with the Neurosciences Medical Image Analysis Group, Electronic Engineering Laboratory, University of Kent at Canterbury, Canterbury, Kent CT2 7NT, U.K.

Publisher Item Identifier S 0278-0062(01)06580-6.

normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$f(\boldsymbol{y}_i \mid \Phi) = \frac{1}{\sqrt{(2\pi)^L|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right)$$

with $\Phi = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ for notational convenience. The maximum-likelihood (ML) parameters are found by maximizing $\prod_{i=1}^{N} f(\boldsymbol{y}_i \mid \Phi)$ or, equivalently, the log-likelihood

$$\sum_{i=1}^{N} \log f(\boldsymbol{y}_i \mid \Phi) \tag{1}$$

yielding $\boldsymbol{\mu} = (1/N)\sum_{i=1}^{N} \boldsymbol{y}_i$ and $\boldsymbol{\Sigma} = (1/N)\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^t$.

In most practical applications, however, the assumed normal model is only an approximation to reality, and estimation of the model parameters Φ should not be severely affected by the presence of a limited amount of outliers. Considerable research efforts in the field of robust statistics [7] have resulted in a variety of methods for robust estimation of model parameters in the presence of outliers, from which the so-called M-estimators [7] present the most popular family. Since $\lim_{f(\boldsymbol{y}|\Phi)\to 0} \log f(\boldsymbol{y} \mid \Phi) = -\infty$, the contribution to the log-likelihood in (1) of an observation that is atypical for the normal distribution is high. The idea behind M-estimators is to alter (1) slightly in order to reduce the impact of such outliers. A simple way to do this, which has recently become very popular in image processing [8]–[12], is to model a small fraction $\epsilon \in [0, 1]$ of the data as being drawn from a rejection class that is assumed to be uniformly distributed

$$f_\epsilon(\boldsymbol{y}_i \mid \Phi) = (1-\epsilon) f(\boldsymbol{y}_i \mid \Phi) + \epsilon\delta.$$

As can easily be verified, assessing the ML parameters is now equivalent to maximizing

$$Q(\Phi) = \sum_{i=1}^{N} \log(f(\boldsymbol{y}_i \Phi) + \lambda), \quad \lambda \geq 0 \tag{2}$$

with respect to the parameters Φ with $\lambda = \epsilon\delta/(1-\epsilon)$[8]. Since $\lim_{f(\boldsymbol{y}|\Phi)\to 0} \log(f(\boldsymbol{y} \mid \Phi) + \lambda) = \log(\lambda)$, the contribution of atypical observations is reduced compared with (1).

The ML parameters Φ should satisfy $(\partial Q(\Phi)/\partial\Phi) = 0$. Since

$$\frac{\partial Q(\Phi)}{\partial \Phi} = \sum_{i=1}^{N} \left(\frac{f(\boldsymbol{y}_i \mid \Phi)}{f(\boldsymbol{y}_i \mid \Phi) + \lambda}\right) \frac{\partial}{\partial \Phi} \log f(\boldsymbol{y}_i \mid \Phi)$$

as shown in [8], one possibility to numerically maximize (2) is to calculate iteratively the weights

$$t_i^{(m)} = \frac{f\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)}{f\left(\boldsymbol{y}_i \Phi^{(m-1)}\right) + \lambda} \tag{3}$$

based on the parameter estimation $\Phi^{(m-1)}$ in iteration $(m-1)$, and subsequently update the parameters $\Phi^{(m)}$ that maximize

$$\sum_{i=1}^{N} t_i^{(m)} \log f(\boldsymbol{y}_i \mid \Phi) \tag{4}$$

which yields for a multivariate normal distribution

$$\begin{aligned}
\boldsymbol{\mu}^{(m)} &= \frac{\sum_{i=1}^{N} t_i^{(m)} \boldsymbol{y}_i}{\sum_{i=1}^{N} t_i^{(m)}} \\
\boldsymbol{\Sigma}^{(m)} &= \frac{\sum_{i=1}^{N} t_i^{(m)} \left(\boldsymbol{y}_i - \boldsymbol{\mu}^{(m)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}^{(m)}\right)^t}{\sum_{i=1}^{N} t_i^{(m)}}.
\end{aligned}$$

Solving an M-estimator by iteratively re-calculating weights and updating the model parameters based on these weights, is commonly referred to as the W-estimator [13]. The weight $t_i^{(m)} \in [0, 1]$ reflects the typicality of sample i with respect to the normal distribution. For typical samples, $t_i^{(m)} \simeq 1$, whereas $t_i^{(m)} \simeq 0$ for samples that deviate far from the model. Comparing (4) with (1), it can be seen that the M-estimator effectively down-weights observations that are atypical for the normal distribution, making the parameter estimation more robust against such outliers.

B. Robust Estimation of MR Model Parameters

In previous work [5], [6], we described a model-based method for fully automated segmentation of normal brain MR images that interleaves tissue classification with estimation of tissue class specific intensity distribution parameters and correction for so-called bias field inhomogeneities. We now outline how this algorithm can be made robust with respect to model outliers, such as MS lesions.

Suppose that there are K tissue types present in an L-channel MR image of the brain. Suppose further that each voxel $i = 1, 2, \ldots, N$ in the image is drawn independently from one of the tissue types $k = 1, 2, \ldots, K$, with some a priori known spatially varying probability π_k. Finally, suppose that the intensity probability distribution of each class k can be modeled by a multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, and that the spatially smoothly varying bias fields in each of the L channels can be modeled by a linear combination of J polynomial basis functions $\phi_j(x_i), j = 1, 2, \ldots, J$ where x_i denotes the spatial position of voxel i. Denoting $\boldsymbol{c_l}$ as the vector of bias field parameters of channel $l \in \{1, 2, \ldots, L\}$, and $\Phi = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K, \boldsymbol{c}_1, \ldots \boldsymbol{c}_L\}$ as the total set of model parameters, the probability density for an MR image with intensities $Y = [\boldsymbol{y}_1\, \boldsymbol{y}_2 \ldots \boldsymbol{y}_N]$ is given by

$$\begin{aligned}
f(Y \mid \Phi) &= \prod_{i=1}^{N} \left(\sum_{k=1}^{K} f_k(\boldsymbol{y}_i \mid \Phi)\pi_k\right) \\
f_k(\boldsymbol{y}_i \mid \Phi) &= \frac{1}{\sqrt{(2\pi)^L|\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{u}_i - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{u}_i - \boldsymbol{\mu}_k)\right) \\
\boldsymbol{u}_i &= \boldsymbol{y}_i - [\boldsymbol{c}_1 \ldots \boldsymbol{c}_L]^t \begin{bmatrix} \phi_1(x_i) \\ \vdots \\ \phi_J(x_i) \end{bmatrix}
\end{aligned}$$

with $\boldsymbol{u}_i$ the intensity of voxel i after bias field correction.

The parameters Φ that maximize the log-likelihood $L(\Phi) = \log f(Y \mid \Phi)$ can be assessed with the so-called expectation-maximization (EM) algorithm [14], that iteratively performs a statistical classification of the voxels based on the current parameter estimation (expectation step), and subsequently updates the parameters based on this classification (maximization step). As shown in [6], the expectation step based on the parameter estimation $\Phi^{(m-1)}$ of the $(m-1)$-th iteration yields, up to a constant that is independent of Φ

$$Q\left(\Phi \mid \Phi^{(m-1)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}^{(m)} \log f_k(\boldsymbol{y}_i \mid \Phi) \tag{5}$$

where the classification

$$p_{ik}^{(m)} = \frac{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)\pi_k}{\sum_q f_q\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)\pi_q} \tag{6}$$

is the probability that voxel i belongs to tissue type k based on the parameters $\Phi^{(m-1)}$. The subsequent maximization step involves searching for the parameters $\Phi^{(m)}$ that maximize (5), yielding closed-form expressions for $\Phi^{(m)}$ that can be found in [5][1] . Interleaving this parameter estimation step with the classification step (6) guarantees iteratively better estimations of Φ[14].

Unfortunately, maximizing (5) with respect to Φ is not robust against outliers in the data, such as MS lesions. The weights p_{ik} represent the degree to which voxel i belongs to tissue type k. Since $\sum_{k=1}^{K} p_{ik} = 1$, an observation that is atypical for each of the tissue classes cannot have a small membership value for all classes simultaneously. Based on the concepts explained in Section II-A we, therefore, introduce a second type of weight that reflects the degree of typicality of each voxel in the K tissue classes. As shown below, these membership values are not constrained to sum to unity for each voxel and, therefore, allow down-weighting model outliers for the model parameter estimation.

Similar to the approach described in Section II-A, where (1) was replaced by the more robust (2), we replace (5) by

$$Q'\left(\Phi \mid \Phi^{(m-1)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}^{(m)} \log(f_k(\boldsymbol{y}_i \mid \Phi) + \lambda). \tag{7}$$

Maximizing (7) with respect to Φ implies $(\partial Q'(\Phi \mid \Phi^{(m-1)})/\partial\Phi) = 0$. Since

$$\frac{\partial Q'\left(\Phi \mid \Phi^{(m-1)}\right)}{\partial \Phi} = \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}^{(m)} \frac{f_k(\boldsymbol{y}_i \mid \Phi)}{f_k(\boldsymbol{y}_i \mid \Phi) + \lambda} \frac{\partial \log f_k(\boldsymbol{y}_i \mid \Phi)}{\partial \Phi}$$

we define the typicality weights

$$t_{ik}^{(m)} = \frac{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)}{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right) + \lambda} \tag{8}$$

and maximize

$$Q_W\left(\Phi \mid \Phi^{(m-1)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}^{(m)} t_{ik}^{(m)} \log(f_k(\boldsymbol{y}_i \mid \Phi) \tag{9}$$

instead. The difference with (5) lies herein, that in (9) the weights p_{ik} are replaced by a combination of two weights $p_{ik}t_{ik}$. Since $\sum_k p_{ik}t_{ik}$ is not constrained to be unity, model outliers can have a small degree of membership in all tissue classes simultaneously. Therefore, observations that are atypical for each of the K tissue types, have a reduced weight on the parameter estimation, thereby robustizing the EM procedure.

Maximization of (9) with respect to Φ results in closed-form expressions for the tissue class specific parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, as well as for the bias field parameters $\boldsymbol{c}_l$. The exact equations are identical to the ones derived in [5], provided that p_{ik} is replaced by $p_{ik}t_{ik}$. For the sake of completeness, we reproduce them in the Appendix.

To summarize, the robustized EM algorithm interleaves classification (6), assessment of typicality (8), estimation of tissue intensity distribution parameters [Appendix, (17), and (18)] and bias field correction [Appendix, (19)]. It is clear that the presented algorithm can be viewed as a generalization of the W-estimator of Section II-A to the case of multiple classes. Furthermore, choosing $\lambda = 0$ results in $t_{ik} = 1, \forall i, k$ and, therefore, reduces the method to the original algorithm described in [5].

C. From Typicality Weights to Outlier Belief Values

So far, we have only been concerned with robust parameter estimation in the presence of model outliers. Our main interest, however, lies in the identification of these outliers, as they are candidate MS lesions. In this section, we take a closer look at the typicality weights t_{ik} calculated in (8) and show how, after a slight alteration, they can be used to assess the belief that voxel i is an outlier.

Referring back to the W-estimator described in Section II-A, (3) classifies the fraction t_i of voxel i as belonging to the normal distribution. The remaining fraction

$$1 - t_i = \frac{\lambda}{f(\boldsymbol{y}_i \mid \Phi) + \lambda}$$

is, therefore, a measure for the belief that voxel i is a model outlier. In a similar way

$$1 - \sum_k p_{ik}t_{ik} = \sum_k p_{ik}(1 - t_{ik}) = \sum_k p_{ik} \frac{\lambda}{f_k(\boldsymbol{y}_i \mid \Phi) + \lambda} \tag{10}$$

reflects the belief that voxel i is not generated by the model described in Section II-B. However, as discussed below, the dependence of (10) through $f_k(\boldsymbol{y}_i \mid \Phi)$ on the determinant of the covariance matrices prevents its direct interpretation as a true outlier belief value.

In statistics, an observation $\boldsymbol{y}$ is said to be abnormal with respect to a given normal distribution if its so-called Mahalanobis-distance $d = \sqrt{(\boldsymbol{y}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}$ exceeds a predefined threshold. Regarding (8), the belief that voxel i is an outlier exceeds the belief that it is a regular sample from class k if $t_{ik} < 0.5$ or $f_k(\boldsymbol{y}_i \mid \Phi) < \lambda$, which is equivalent to $d_{ik}^2 > -2\log(\lambda\sqrt{(2\pi)^L|\boldsymbol{\Sigma}_k|})$. Because of its dependence on $|\boldsymbol{\Sigma}_k|$, the Mahalanobis-distance threshold above which i is considered abnormal with respect to class k changes over the iterations of the EM algorithm as $\boldsymbol{\Sigma}_k$ is updated. Even worse, this threshold varies over the different classes, in such a way that

[1] In fact, these equations only guarantee that $Q(\Phi^{(m)} \mid \Phi^{(m-1)}) \geq Q(\Phi^{(m-1)} \mid \Phi^{(m-1)})$, resulting in a generalized EM algorithm [14].

observations are more easily rejected from classes with a broad distribution than from classes with a narrow one. Because of these problems, it is not clear how λ should be chosen.

Ideally, a voxel i should be considered abnormal with respect to class k if $d_{ik} > \kappa$, where $\kappa \geq 0$ is an explicit Mahalanobis-distance threshold that is equal for all classes alike. We investigated modifications to (8) that result in this behavior, and we have selected a method which replaces (8) by

$$t_{ik}^{(m)} = \frac{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)}{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right) + \frac{1}{\sqrt{(2\pi)^L |\boldsymbol{\Sigma}_k^{(m-1)}|}} \exp\left(-\frac{1}{2}k^2\right)} \tag{11}$$

λ is now made class-dependent because $|\boldsymbol{\Sigma}_k|$ is taken into account, and re-parameterized using the more easily interpretable κ. The actual choice of κ can be regarded as the choice of a statistical significance level and will be dealt with in Section III.

D. Application to MS Lesion Segmentation

Outlier voxels also occur outside MS lesions. This is typically true for partial volume (PV) voxels that, in contravention to the assumptions made in Section II-B, do not belong to one single tissue type but are rather a mixture of more than one tissue. Since they are perfectly normal brain tissue, though, we prevent them from being detected as MS lesion by introducing constraints on intensity and context on the weights t_{ik} calculated in (11).

In our MR images, we have also noticed the presence of gross model outliers in the cerebro-spinal fluid (CSF) that appear dark on the PD- and T2-weighted images and that we attribute mainly to the falx cerebri of the dura mater and to blood vessels. While we are only interested in MS lesions, we have experienced that these CSF outliers can impede accurate estimations of the tissue-specific intensity models. Therefore, we explicitly look for these CSF outliers as well, so that they are rejected from the model parameter estimation too.

1) Additional Intensity Constraints: Since MS lesions appear hyper-intense on both the PD- and the T2-weighted images, we define

$$l_i^{(m)} = \begin{cases} 1, & \text{if } \left[\boldsymbol{u}_i^{(m)}\right]_{\text{T2}} > \left[\boldsymbol{\mu}_{GM}^{(m-1)}\right]_{\text{T2}} \\ & \text{and } \left[\boldsymbol{u}_i^{(m)}\right]_{\text{PD}} > \left[\boldsymbol{\mu}_{GM}^{(m-1)}\right]_{\text{PD}} \\ 0, & \text{otherwise} \end{cases}$$

as an indicator of whether voxel i has the correct intensity to be a candidate MS lesion, based on the estimated mean intensity of grey matter (GM) in the T2- and the PD-weighted channel. Here, $\boldsymbol{u}_i^{(m)}$ denotes $\boldsymbol{y}_i$ after bias field correction [Appendix, (16)]. Similarly

$$v_i^{(m)} = \begin{cases} 1, & \text{if } \left[\boldsymbol{u}_i^{(m)}\right]_{\text{T2}} < \left[\boldsymbol{\mu}_{GM}^{(m-1)}\right]_{\text{T2}} \\ & \text{and } \left[\boldsymbol{u}_i^{(m)}\right]_{\text{PD}} > \left[\boldsymbol{\mu}_{GM}^{(m-1)}\right]_{\text{PD}} \\ 0, & \text{otherwise} \end{cases}$$

indicates candidacy to belong to the dark outliers in the CSF. We now replace (11) by

$$t_{ik}^{(m)} = \frac{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)}{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right) + \frac{\exp\left(-\frac{1}{2}k^2\right)}{\sqrt{(2\pi)^L |\boldsymbol{\Sigma}_k^{(m-1)}|}} \left(l_i^{(m)} + v_i^{(m)}\right)} \tag{12}$$

so that voxels that do not meet the intensity conditions for lesion or CSF outliers are not allowed to have a reduced weight t_{ik}. Because l_i and v_i cannot simultaneously attain the unity value

$$l_i \left(1 - \sum_k p_{ik} t_{ik}\right) \tag{13}$$

is the belief that voxel i belongs to an MS lesion, whereas this is $v_i\,(1 - \sum_k p_{ik}t_{ik})$ for CSF outliers.

2) Additional Contextual Constraints: Around 90%–95% of the MS lesions are WM lesions, and the gross dark outliers are located inside the CSF. We, therefore, add the contextual constraint that MS lesions and CSF outliers should be located in the vicinity of WM and normal CSF, respectively. To this end, we define

$$q_{ik} = \begin{cases} p_{ik}t_{ik} + l_i(1 - \sum_h p_{ih}t_{ih}), & \text{if } k = \text{WM} \\ p_{ik}t_{ik} + v_i(1 - \sum_h p_{ih}t_{ih}), & \text{if } k = \text{CSF} \\ p_{ik}t_{ik}, & \text{otherwise} \end{cases}$$

in which the classification map of what is assumed healthy WM is fused with the map of MS lesions, yielding a mask that covers the total WM. In the same way, the map of CSF outliers is added to the classification map of CSF. Inspired by the application of Markov random fields (MRFs) using the mean-field approximation in the EM procedure as presented in [6], we introduce a spatially varying prior for the K tissue types using a simple Potts model, i.e., the extension of the binary Ising [15] model to more than two classes. Let the $[K \times K]$ matrices G and H denote the costs associated with class transitions among neighboring voxels in the plane and out of the plane, respectively

$$G = \begin{bmatrix} 0 & \zeta & \zeta & \cdots \\ \zeta & 0 & \zeta & \cdots \\ \zeta & \zeta & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad H = \begin{bmatrix} 0 & \eta & \eta & \cdots \\ \eta & 0 & \eta & \cdots \\ \eta & \eta & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Let $\boldsymbol{g}_i$ denote a vector whose k-th component $g_{ik} = \sum_{i'} q_{i'k}$ counts the number of its four nearest neighbors i' in the plane that are classified to tissue type k. Similarly, let $\boldsymbol{h}_i$ represent the same entity based on its two nearest neighbors out of the plane. The prior probability $\omega_{ik}^{(m)} = [\boldsymbol{\omega}_i^{(m)}]_k$ that voxel i belongs to tissue type k given the classification of its neighbors during the previous iteration $(m-1)$, is, up to a normalization factor, modeled by [6]

$$\boldsymbol{\omega}_i^{(m)} = \exp\left(-G\boldsymbol{g}_i^{(m-1)} - H\boldsymbol{h}_i^{(m-1)}\right).$$

In other words, the higher ζ and η, the lower the prior probability that a voxel belongs to a class that is different from its neighbors.

Because MS lesions can usually be considered as abnormal WM, we assume that the prior probability that voxel i belongs to a lesion equals the prior probability that it is WM. The same reasoning holds for the dark outliers and CSF. Therefore, the addition of the contextual constraints results in

$$t_{ik}^{(m)} = \frac{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)\omega_{ik}^{(m)}}{f_k\left(\boldsymbol{y}_i \mid \Phi^{(m-1)}\right)\omega_{ik}^{(m)} + \frac{\exp\left(-\frac{k^2}{2}\right)\left(l_i^{(m)}\omega_{i,\mathrm{WM}}^{(m)} + v_i^{(m)}\omega_{i,\mathrm{CSF}}^{(m)}\right)}{\sqrt{(2\pi)^L \left|\sum_k^{(m-1)}\right|}}}. \tag{14}$$

In the absence of neighboring WM, the prior probability for WM $\omega_{i,\mathrm{WM}}$ will be lower than the prior ω_{ik} for other tissue types, thereby discouraging voxel i from being classified as MS lesion.

As explained in [6], the MRF parameters ζ and η are derived from the data itself. During each iteration m, a neighborhood configuration histogram is set up, that counts how many times each possible combination of different tissue types in neighboring voxels occurs in the current classification of nonlesion tissue $p_{ik}^{(m)} t_{ik}^{(m)}$. From this histogram, an estimation of $\zeta^{(m)}$ and $\eta^{(m)}$ is derived [6]. The larger the voxel size, the more the tissue types will be mixed in each other's neighborhood, and the lower the class transition costs ζ and η. Therefore, the contextual constraints in (14) automatically adapt to the voxel size of the data.

To summarize, the algorithm iteratively interleaves statistical classification of the voxels into normal tissue types using (6), assessment of the belief for each voxel that it is not part of an MS lesion or a CSF outlier based on its intensity and on the classification of its neighboring voxels using (14), and, only based on what is considered as healthy tissue, estimation of the multivariate normal intensity distributions and bias correction (Appendix, (17),(18) and (19)). Upon convergence, the belief that voxel i is part of an MS lesion is obtained from (13).

E. Initialization and Convergence Criterion

We initialize the iterative algorithm by providing it with a rough prior estimation of the classification p_{ik}. With the typicality weights t_{ik} and the bias field parameters $\boldsymbol{c}_l$ initialized to unity and zero, respectively, a first estimation of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ is made (Appendix, (17) and (18)), allowing in its turn a preliminary estimation of the bias field parameters $\boldsymbol{c}_l$ (Appendix, (19)). These model parameter estimations then provide a new estimation of the classification p_{ik}(6) and typicality weights t_{ik}(14), etc.

The prior classification is derived from a digital brain atlas that contains information about the expected location of WM, GM, and CSF (see [5] for more details). We use the atlas that is distributed with the SPM99 software package [16], that is the average of a large number of affinely co-registered manual segmentations of healthy brain MR images [17]. We use the spatially varying prior probability maps of the atlas not only for initialization but also for the tissue type priors π_k in (6). As will be explained in Section II-F, the atlas can be fully automatically brought into spatial correspondence with the MR data. Therefore, the EM algorithm works without any user interaction, so that its calculations are fully reproducible.

Convergence of the original EM algorithm that iteratively improves the log-likelihood $L(\Phi)$ of the MR model of Section II-B means that the relative change of the log-likelihood $|L(\Phi^{(m)}) - L(\Phi^{(m-1)})|/|L(\Phi^{(m)})|$ becomes negligible. It can be shown [14] that, $\forall\Phi$

$$L(\Phi) - L\left(\Phi^{(m-1)}\right) \geq Q\left(\Phi \mid \Phi^{(m-1)}\right) - Q\left(\Phi^{(m-1)} \mid \Phi^{(m-1)}\right)$$

where $Q(\Phi \mid \Phi^{(m-1)})$ is given by (5), which proves that iterative maximization of $Q(\Phi \mid \Phi^{(m-1)})$ indeed consistently improves $L(\Phi)$. In this paper, however, we iteratively maximize the modified version $Q_W(\Phi \mid \Phi^{(m-1)})$ given by (9) rather than the original $Q(\Phi \mid \Phi^{(m-1)})$, so that $L(\Phi)$ is no longer guaranteed to increase. We, therefore, detect convergence when

$$\xi^{(m)} = \frac{\left|Q_W\left(\Phi^{(m)} \mid \Phi^{(m-1)}\right) - Q_W\left(\Phi^{(m-1)} \mid \Phi^{(m-1)}\right)\right|}{\left|Q_W\left(\Phi^{(m)} \mid \Phi^{(m-1)}\right)\right|} \tag{15}$$

becomes negligible.

F. Implementation

In our implementation we use four classes: WM, GM, CSF, and "other," where the "other" class is modeled by two normal distributions for nonbrain tissues and a Rayleigh distribution for MR background noise as explained in [5]. Voxels where the atlas indicates a prior probability of unity for "other" are of no interest and are discarded. Because of the atlas, p_{ik} is only nonzero for $k =$ "other" in regions far away from the expected location of MS lesions. We, therefore, fix $t_{ik} = 1$ for $k =$ "other," thereby not allowing voxels to be rejected from the "other" class as these could never be true MS lesions.

We have implemented the method in Matlab-code [18] on top of the SPM99 package [16], except for the MRF related parts that were coded in C for efficiency purposes. We first bring the multispectral MR data of the same subject into spatial correspondence using a fully automated affine registration technique based on maximization of mutual information [19]. With the same registration program, we drive a T1-weighted image that is associated with the digital brain atlas into correspondence with the MR data. The prior probability maps of the atlas are subsequently resampled to the image grid of the MR data, and are used for initialization of the EM algorithm as explained in Section II-E. To speed up the iterative process, the model parameters are updated based on only those voxels that lie on the subgrid of the full image grid that best approximates 4 × 4 × 4 mm^3. As in [5] and [6], we use fourth-order polynomial models for the bias fields, and we stop the iterations as soon as $\xi^{(m)}$(15) drops below 10^{-4}.

To summarize, the method is fully automated, with only a single parameter κ in (14) that needs to be experimentally tuned. The choice of κ significantly affects the quality of the MS lesion segmentation, as will be discussed in detail in Section III.

III. Results

As part of the BIOMORPH project [20], we analyzed MR data acquired during a clinical trial in which 50 MS patients were repeatedly scanned with an interval of approximately one month over a period of about one year. The serial image data were acquired on a Philips T5 1.5-T MR scanner, and consisted

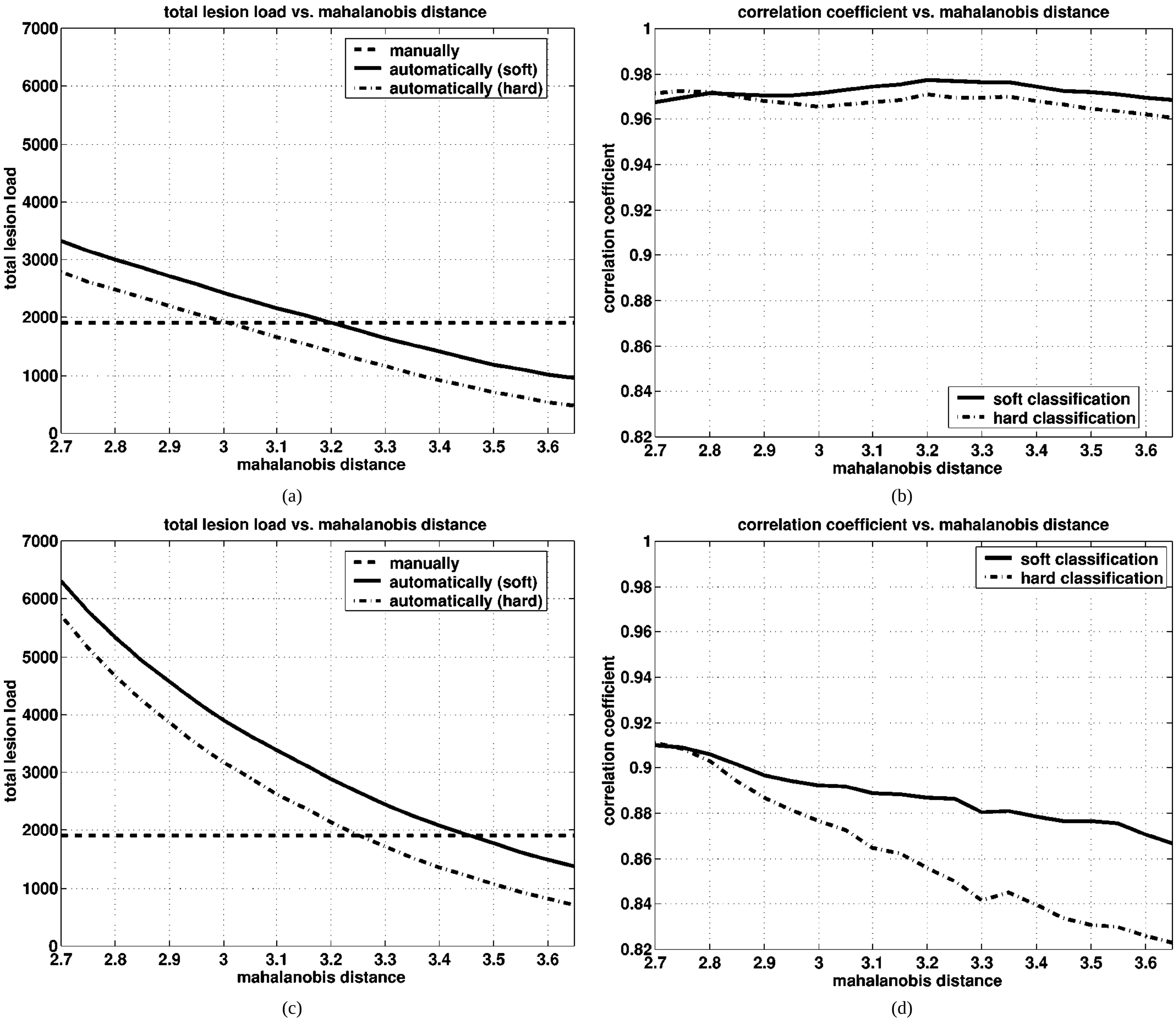


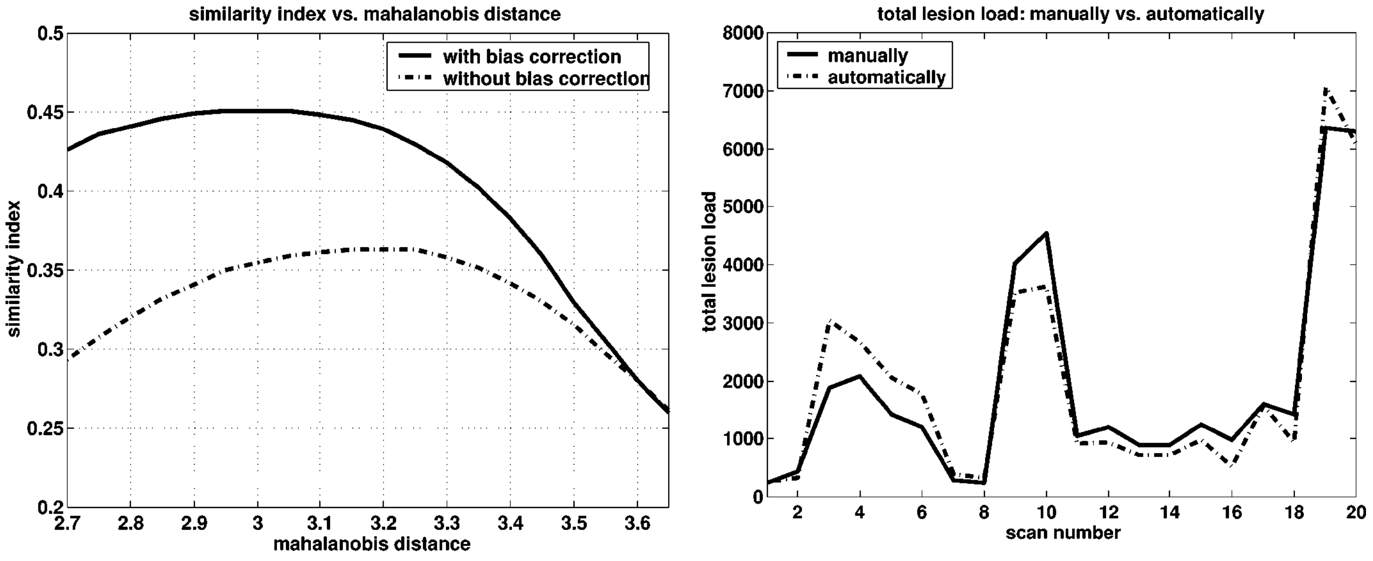
Fig. 1. (a) TLL over all 20 scans of ten patients at two time points, segmented by manual delineation of the MS lesions by a human expert and by the automated method with varying values of κ. (b) Correlation coefficient between the TLL values of these 20 scans individually obtained by manual delineation and by the automated method with varying values of κ. (c) and (d) Same as (a) and (b), respectively, but with the bias correction component of the automatic algorithm disabled.

at each time point of a double echo spin-echo PD/T2-weighted image pair (TR 2816 ms and TE 20/80 ms) and a spin-echo T1-weighted image (TR 425 ms and TE 15 ms) that contained 24 axial slices with a pixel size of 0.9 × 0.9 mm^2, 5-mm slice thickness, and an interslice gap of 0.5 mm. For each patient, at least one additional scan of the same modalities but with a higher spatial resolution was acquired, consisting of a double echo turbo spin-echo PD/T2 weighted image pair (TR 3300 ms and TE 23.5/120 ms, 52 axial slices, pixel size 0.9 × 0.9 mm^2, slice thickness 2.4 mm, interslice gap 0.1 mm) and a fast field echo T1-weighted image (TR 28.3 ms and TE 6.9 ms, 60 continuous 2.4-mm-thick axial slices with pixel size 0.9 × 0.9 mm^2).

A. Validation on Low-Resolution Images

MS lesions were manually traced by a human expert based only on the T2-weighted images for ten patients at two consecutive time points. Segmentations obtained with the automatic algorithm with varying values of κ were compared with the expert segmentations by the total lesion load (TLL) on these 20 scans, measured as the number of voxels classified as MS lesion. TLL was calculated in two different ways for the automated segmentations. The first, based on a partial occupancy or "soft" classification, computed TLL as the integration of the estimated lesion fraction (13) over all voxels, whereas the second, based on a "hard" classification, measured TLL as the number of voxels in which this fraction exceeds the value of 0.5.

Fig. 1(a) shows the average TLL over the 20 scans for the automated method for the Mahalanobis distance κ varying from 2.7 (corresponding to a significance level of $p = 0.063$) to 3.65 ($p = 0.004$), in steps of 0.05. As expected, both TLL values calculated by the automated method decrease when κ is increased, since the higher κ, the less easily voxels are rejected from the



(a) (b)

Fig. 2. (a) Similarity index between the automatic and the expert lesion delineations on 20 data sets for varying values of κ, with and without the bias field correction component enabled in the automated method. (b) TLL of each of the 20 data sets obtained by manual delineation and using the automated method with $\kappa = 3$.

model. It can also be seen that the TLL based on a hard classification is significantly smaller than the TLL calculated directly from the soft classification. Referring to (14), this can be explained by the fact that the lesion fraction is small but nonzero for voxels that are well explained by the model. Summation of this small fraction over all voxels results in a global offset that is not negligible. When the TLL is calculated after a hard classification, the contribution of the nonlesion voxels disappears, yielding a better indication of the TLL calculated by the algorithm. Fig. 1(a) shows that varying κ from 2.7 to 3.65 results in an automatic TLL ranging from 150% to only 25% of the expert TLL, and that $\kappa \simeq 3$ ($p = 0.029$) results in an automatic TLL that is very close to the manual TLL.

Despite the strong influence of κ on the absolute TLL values, the linear correlation between the automated TLLs of the 20 scans and the expert TLLs is remarkable insensitive to the choice of κ. Fig. 1(b) depicts the correlation coefficient for κ varying from 2.7 to 3.65. Over this wide range, the correlation coefficient varies between 0.96 and 0.98 for both the hard and the soft TLLs. For values of κ where the automated method under-segments the lesions according to Fig. 1(a), the effect of converting the soft classification into a hard one introduces random noise in the measurements, resulting in a slightly lower correlation coefficient. In contrast, for values of κ where the method over-segments, eliminating voxels with a small lesion fraction seems to help in reducing the number of false positives.

To investigate the need for bias field correction, we re-applied the algorithm on the same data for the same range of κ, but with the bias field coefficients c_l fixed to zero throughout the iterations. The average TLL and the correlation coefficients with varying κ in this situation are depicted in Fig. 1(c) and (d), which need to be compared with Fig. 1(a) and (b), respectively. Clearly, bias field correction has a tremendous effect on the quality of the automated MS lesion segmentation. Whereas previously the correlation coefficient varied between 0.96 and 0.98, it varies between 0.82 and 0.91 without bias field correction.

Comparing the TLL of two raters does not take into account any spatial correspondence of the segmented lesions. We, therefore, also compared the different segmentations using the similarity index [21], [22]. With V_a and V_e the number of voxels rated as MS lesion by the automated algorithm after hard classification and by the expert, respectively, and V_{ae} the number of voxels rated as lesion by both the automated method and the expert, the similarity index is defined as $2V_{ae}/(V_a + V_e)$, which is simply the volume of intersection of the two segmentations divided by the mean of the two segmentation volumes. Fig. 2(a) depicts the value of this index over all 20 scans for varying κs, both with and without bias correction, again demonstrating the need for bias field corrections. The best correspondence, with a similarity index of 0.45, was found for $\kappa \simeq 3$. For this value of κ, the automatic TLL was virtually equal to the expert TLL, as can be verified from Fig. 1(a). Therefore, a similarity index of 0.45 means that less than half of the voxels labeled as lesion by the expert were also identified by the automated method, and vice versa.

For illustration purposes, we depict the expert TLLs of the 20 scans along with the automatic ones for $\kappa = 3$ in Fig. 2(b). A paired t-test did not reveal a significant difference between the manual and these automatic TLL ratings ($p = 0.94$). Scans 1 and 2 are two consecutive scans from one patient, 3 and 4 from the next, and so on. Note that in nine out of ten cases, the two ratings agree over the direction of the change of the TLL over time. Fig. 3 displays the MR data of what is called scan 19 in Fig. 2(b) and the automatically calculated classification along with the lesion delineations performed by the human expert.

B. Validation on Higher-Resolution Images

In addition to the 20 manual delineations on low-resolution images, three of the higher resolution scans were also analyzed by two human experts, trained at a different institute, by tracing MS lesions based on the T2-weighted images alone. One of the

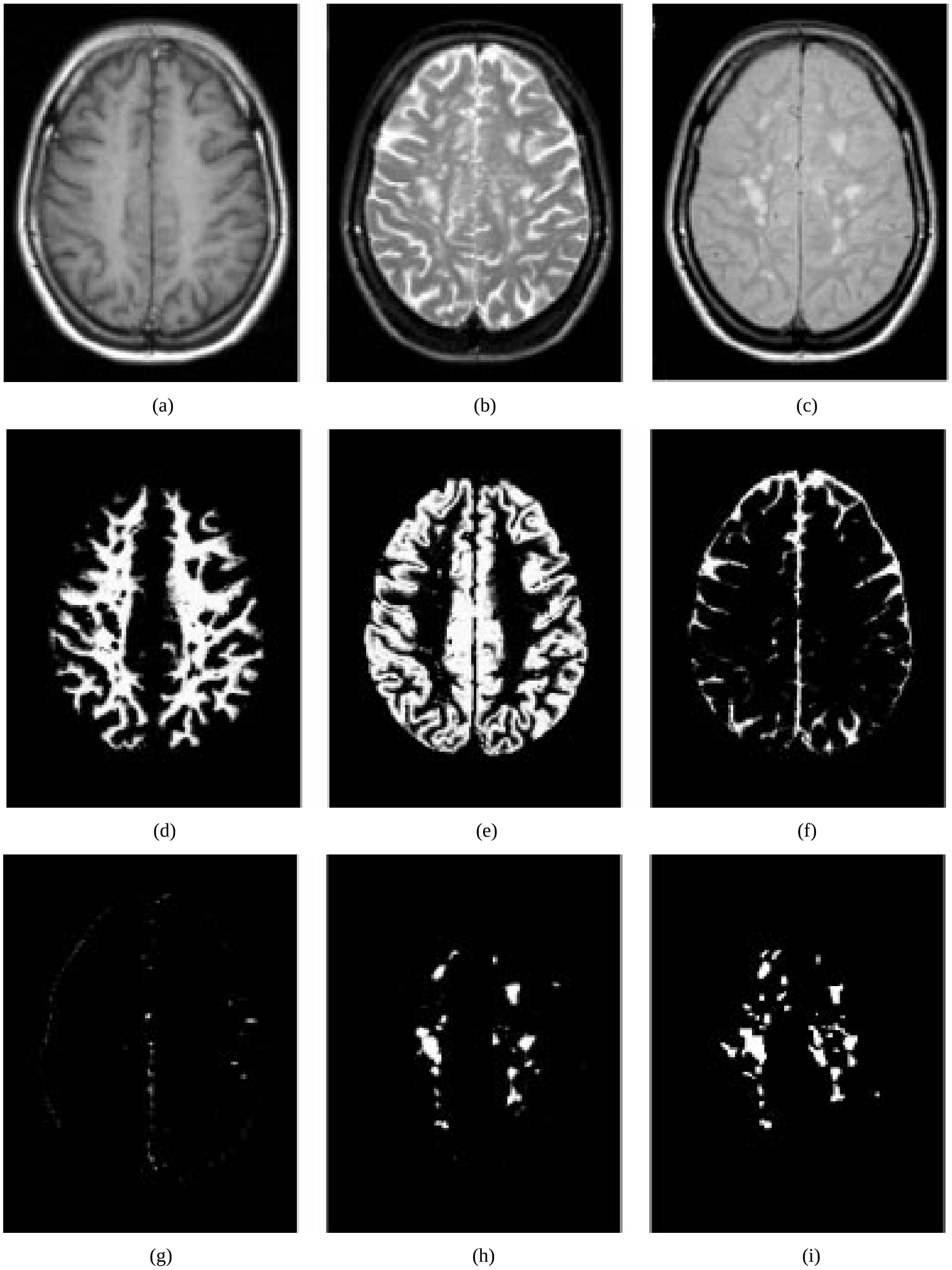


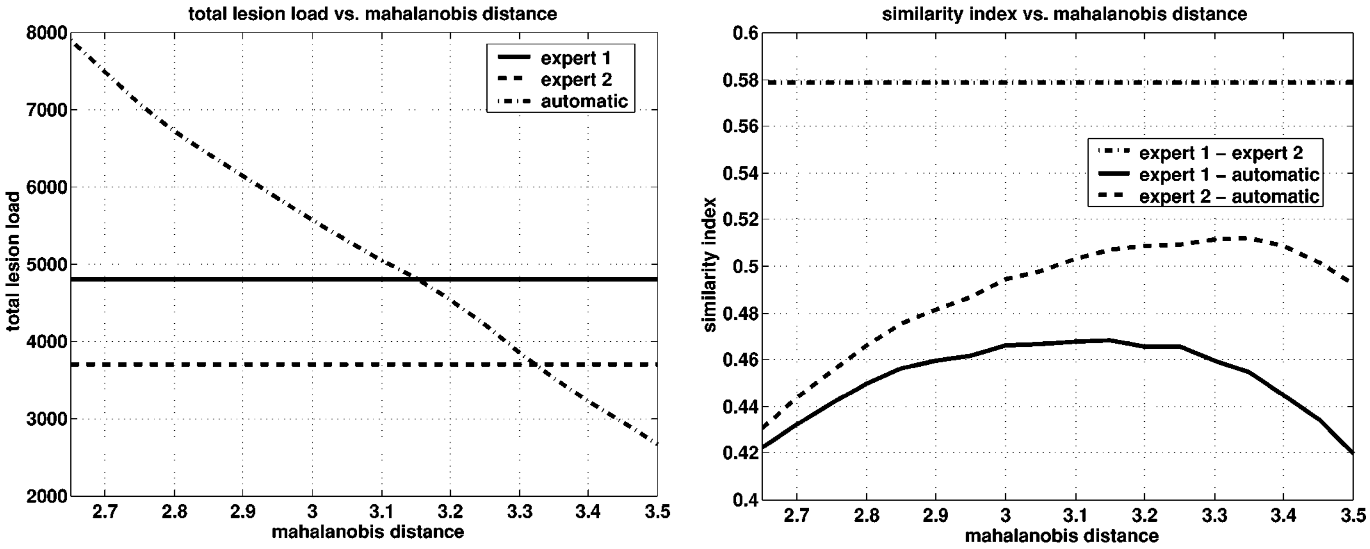
Fig. 3. Automatic classification of one of the 20 data sets that were also analyzed by a human expert. (a) T1-weighted image. (b) T2-weighted image. (c) PD-weighted image. (d) WM classification. (e) GM classification. (f) CSF classification. (g) dark outliers in the CSF. (h) MS lesion classification. (i) Expert delineation of the MS lesions.

experts, hereafter referred to as expert 2, was the same expert who delineated the lesions on the low-resolution images. As can be seen from Table I, expert 1 consistently labeled more voxels as MS lesion than expert 2, with an average of 30% more volume. Regarding the spatial correspondence of the lesion delineations, the similarity index between the two experts was 0.58. Only 51% of the voxels labeled by expert 1 were also indicated by expert 2, whereas this was 66% percent conversely, indicating that the voxels which both experts labeled as lesion had a closer correspondence with the delineations of expert 2 than with the delineations of expert 1.

TABLE I
TOTAL LESION LOAD OF LESIONS SEGMENTED BY TWO HUMAN EXPERTS ON THREE HIGHER-RESOLUTION SCANS

TLL	scan 1	scan 2	scan 3
expert 1	5303	939	8172
expert 2	3772	598	6745

For these high resolution images, Fig. 4(a) shows the average TLL computed by the automated algorithm for κ varying from 2.65 to 3.5 (corresponding to $p = 0.071$ and $p = 0.007$, respectively), based on hard classifications. The similarity index



(a)

(b)

Fig. 4. Automated lesion segmentation with varying values of κ on three higher resolution data sets compared with manual delineation by two human experts: (a) TLL; (b) similarity index between each pair of raters.

between the automatic method and either of the experts for the same range of κ is depicted in Fig. 4(b). As was the case with the low-resolution data in Section III-A, the κ for which the average automatic TLL is most similar to the average expert TLL, is also the one that yields the best spatial correspondence. Since expert 1 indicated more voxels as MS lesion than expert 2, the κ for which the automated method best mimics the labeling performed by expert 1 is smaller than the one that best approaches the segmentations done by expert 2. Whereas in Section III-A, the optimal κ with regard to expert 2 was 3 ($p = 0.029$), it is now 3.35 ($p = 0.011$), which is a remarkably higher value. This can be explained by the fact that in Section III-A, the slice thickness was more than double the slice thickness in the current data and, therefore, the current data contain far less PV voxels. Therefore, the different tissue types are more tightly clustered in feature space, yielding smaller estimates of the variances $\mathbf{\Sigma}_k$, thereby reducing the absolute distance to the means $\boldsymbol{\mu}_k$ at which voxels are considered as abnormal. The best similarity index for the automatic method compared with expert 1 is 0.47, and to expert 2 is 0.51, indicating that the automatic method shows a better spatial correspondence with expert 2 than with expert 1. When compared with the similarity index of 0.58 between the two experts, it can be seen that the two experts are in better agreement with each other than with the automated algorithm.

IV. Discussion

We have described a model-based method for automated MS lesion segmentation that iteratively interleaves statistical classification of the image voxels into a number of healthy tissue types, assessment of the belief for each voxel that it actually belongs to healthy tissue, and estimation of intensity distribution parameters and MR bias field parameters only based on healthy tissue voxels. The main characteristics of our algorithm are the detection strategy of MS lesions as model outliers, the absence of any human interaction due to the use of a probabilistic brain atlas, and the automatic adaptation of the method to changes in MR pulse sequence and voxel size.

MS lesions are detected as voxels that are not well explained by a statistical model for normal brain MR images. This approach circumvents explicit lesion modeling, which is difficult because of their widely varying appearance in MR images, and because not every individual scan contains a sufficient number of lesions for estimating the model parameters. The core of our method is a clustering algorithm that is made robust against model outliers, which is a research topic that has recently received much attention (see [23] and [24] for an overview). From an algorithmic point of view, our method bears close resemblance to an adaptation of the EM classifier described by Schroeter *et al.* [9], who iteratively classified normal brain MR images into a small number of Gaussian distributions, each time rejecting voxels that exceed a predefined Mahalanobis distance to each of the Gaussians, and updating the model parameters only based on nonrejected voxels. In contrast to their method that either accepts or rejects voxels, our method uses a soft rejection scheme and also takes the classification of the voxels and their neighbors into account.

Most of the methods for MS lesion segmentation described in the literature are semi-automated rather than fully automated methods, designed to facilitate the tedious task of manually outlining lesions by human experts, and to reduce the interrater and intrarater variability associated with such expert segmentations. Typical examples of user interaction in these approaches include accepting or rejecting automatically computed lesions [25], or manually drawing regions of pure tissue types for training an automated classifier [22], [25]–[27]. While these methods have proven to be useful, they remain impractical when hundreds of scans need to be analyzed as part of a clinical trial, and the variability of manual tracings is not totally removed. In contrast, our method is fully automated as it uses a probabilistic brain atlas to train its classifier. Furthermore, the atlas provides spatial information that avoids nonbrain voxels from being classified as MS lesion, making the method work without the often-used tracing of the intracranial cavity in a preprocessing step [22], [25]–[29]. A limitation of this approach is that the method cannot be directly applied to regions for which no such atlas is currently available, such as the spinal cord.

A unique feature of our algorithm is that it automatically adapts its intensity models and contextual constraints when analyzing images that were acquired with a different MR pulse sequence or voxel size, such as for instance the scans of Section III-A and B. Zijdenbos *et al.* described [30] and validated [31] a fully automated pipeline for MS lesion segmentation based on an artificial neural network classifier. Similarly, Kikinis[28] *et al.* and Guttmann *et al.* [32] have developed a method with minimal user intervention that is built on the EM classifier of Wells *et al.* [33] with dedicated preprocessing and postprocessing steps. Both methods use a fixed classifier that is only trained once and that is subsequently used to analyze hundreds of scans. In clinical trials, however, interscan variations in cluster shape and location in intensity space cannot be excluded, not only because of hardware fluctuations of MR scanners over a period of time, but also because different imagers may be used in a multicenter trial [32]. In contrast to the methods described above, our algorithm retrains its classifier on each individual scan, making it adaptive to such contrast variations. Also, due to the multispectral nature of the approach, additional MR data that may be available, such as fluid attenuated inversion recovery (FLAIR) images, can be immediately exploited by our method to facilitate tissue classification and lesion discrimination without prior re-training of the classifier.

Often, a post-processing step is applied to automatically segmented MS lesions, in which false positives are removed based on a set of experimentally tuned morphologic operators, connectivity rules, size thresholds, etc [22], [26], [28]. Since such rules largely depend on the voxel size, they may need to be re-tuned for images with a different voxel size. Alternatively, images can be re-sampled to a specific image grid before processing, but this introduces partial voluming that can reduce the detection of lesions considerably, especially for small lesion loads [32]. To avoid these problems, we have added explicit contextual constraints on the iterative MS lesions detection that automatically adapt to the voxel size. Similar to other methods [26], [27], [29], [30], we exploit the knowledge that the majority of MS lesions is situated inside WM. Johnston *et al.* [26], [34] fused the segmentation maps of normal WM and MS lesions obtained with a MRF-based statistical classifier, producing a mask that covers the total WM, and subsequently re-classified the voxels inside that mask to either WM or lesion. Similarly, our method fuses the normal WM with the lesions in each iteration, producing, in combination with MRF constraints, a prior probability mask for WM that is automatically updated during the iterations. Since the MRF parameters are re-estimated for each individual scan, the contextual constraints automatically adapt to the voxel size of the images.

A number of authors have explored the use of the time domain for MS lesion segmentation in serial MR data. After segmentation of each three-dimensional (3-D) data set individually in a time series, Metcalf *et al.* [35] and Kikinis *et al.* [28] used a four-dimensional (4-D) connected component labeling as a post-processing step to remove lesions that appeared isolated in time or that had a 4-D volume below a predefined threshold. Gerig *et al.* [36] only considered the voxel intensity changes over time, without segmentation of spatial structures. In a similar vein, Rey *et al.* [37] and Thirion and Calmon [38] analyzed the deformation field computed by Thirion's nonrigid registration algorithm [39] between two consecutive time points. Extending the MRF constraints from 3-D to 4-D in our approach would yield an algorithm that takes both spatial and temporal contextual information into account, with possibly a better discrimination between lesion and nonlesion voxels. However, a major drawback of using the time domain as an additional feature, is the need for coregistration and resampling of all images of a serial scan sequence. As mentioned previously, this resampling introduces partial voluming that decreases the spatial discrimination power [32] for which the additional discrimination power of the time domain might not compensate.

Although the algorithm we present is fully automatic, an appropriate Mahalanobis distance threshold κ has to be chosen in advance. When evaluating the role of κ, a distinction has to be made between the possible application areas of the method. In clinical trials, the main requirement for an automated method is that its measurements change in response to a treatment in a manner proportionate to manual measurements, rather than having an exact equivalence in the measurements [2], [3]. In Section III-A we, therefore, performed a linear regression analysis between the TLLs estimated by the automatic algorithm and the TLLs derived from human expert segmentations on 20 data sets for a wide range of κ. Although the average TLL produced by the automated method varied from only 25% up to 150% of the average expert TLL estimation, the automatic measurements always kept changing proportionately to the manual measurements, with high correlation coefficients between 0.96 and 0.98. Therefore, the actual choice of κ is fairly unimportant for this type of application.

The role of κ is much more critical when the goal is to investigate the basic MS mechanisms or time correlations of lesion groups in MS time series, as these applications require that the lesions are also spatially correctly detected. To assess the spatial correspondence between automated segmentations and manual lesion tracings performed by two human experts, we calculated the so-called similarity index between each pair of raters on three data sets in Section III-B. By varying κ, the automated method could be tuned toward the segmentation behavior of each expert. However, despite the high interexpert variability, the agreement between the two experts was still clearly better than between either of the experts and the automated algorithm. Of particular concern was the fact that the optimal κ that best brought the automatic segmentations into agreement with the labelings done by one of the experts, differed for the different scan types of Section III-A and B. We believe that this effect is caused by the presence of far more PV voxels in the data of Section III-A, due to the considerably larger slice thickness. In general, the higher the resolution and the better the contrast between lesions and unaffected tissue in the images, the easier MS lesions are detected by the automatic algorithm and the higher κ should be chosen. Therefore, the algorithm presumably needs to be tuned for different studies, despite the automatic adaptation of the tissue models and the MRF parameters to the data.

While the method yields TLL measurements that clearly correlate with manual estimates, its spatial localization of MS lesions does not seem sufficiently accurate compared with segmentations performed by two human experts, trained at a different institute, on three data sets. However, the automatic algorithm worked on multispectral data, while the manual segmentations were only based on T2-weighted images, which might explain part of the disagreement. For a more thorough validation and assessment of intraobserver and interobserver variability associated with manual delineation, multiple human experts trained at different institutions should perform numerous manual segmentations of volume data sets, having available multispectral MR to assist in lesion identification. However, even then, validation of the automated method remains difficult. The fundamental problem is the lack of any method to measure accuracy: that is, there is no reliable method to identify which portions of the image truly represent MS lesions. Well designed studies such as [31] have shown that there is very wide variation in lesion volumes estimated by different observers, especially when these were trained at different institutions. There is no reason to assume that the mean of a large number of human expert estimations represents the true result.

To our knowledge, the spatial distribution of automatically segmented lesions has not received much attention in the literature, most methods concentrating only on the TLL [30]–[32] or on the reduction of human inter and intraobserver variability in TLL estimation with semi-automated methods [22], [25], [27]. Bello and Colchester [40] reviewed different measures for spatial correspondence and introduced a mutual information-based index which took account of the probability of chance correspondence. In the present paper we have used the simpler "similarity index" which is the intersection divided by the mean.

To date, we have successfully analyzed over 300 scans with the method presented in this paper. In the future, we plan to extend our work by subdividing and analyzing automatically segmented lesions based on their spatial location relative to the atlas and based on their appearance in the T1-, T2-, and PD-weighted images. As part of the clinical trial described in this paper, we ultimately intend to correlate the automatic measurements with the clinical data of the patients.

V. Conclusion

This paper describes a fully automated atlas-based approach for MS lesion segmentation from multispectral MR images. The method simultaneously estimates the parameters of a stochastic tissue intensity model for normal brain MR images, and detects MS lesions as voxels that are not well explained by the model. The results of the automated method were compared with lesions delineated by human experts, showing a high TLL correlation, but an average overall spatial correspondence that is lower than that between the experts.

Appendix

We here reproduce the closed-form expressions for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{c}_l$ that maximize (9). The interested reader is referred to [5] for more information. Let

$$\boldsymbol{u}_i^{(m)} = \boldsymbol{y}_i - \left[\boldsymbol{c}_1^{(m-1)} \ldots \boldsymbol{c}_L^{(m-1)}\right]^t \begin{bmatrix} \phi_1(x_i) \\ \vdots \\ \phi_J(x_i) \end{bmatrix} \tag{16}$$

represent the intensity of voxel i after correction for the bias field that was estimated during the previous iteration $(m-1)$. The intensity distribution parameters are then given by

$$\boldsymbol{\mu}_k^{(m)} = \frac{\sum_i p_{ik}^{(m)} t_{ik}^{(m)} \boldsymbol{u}_i^{(m)}}{\sum_i p_{ik}^{(m)} t_{ik}^{(m)}} \tag{17}$$

$$\boldsymbol{\Sigma}_k^{(m)} = \frac{\sum_i p_{ik}^{(m)} t_{ik}^{(m)} \left(\boldsymbol{u}_i^{(m)} - \boldsymbol{\mu}_k^{(m)}\right)\left(\boldsymbol{u}_i^{(m)} - \boldsymbol{\mu}_k^{(m)}\right)^t}{\sum_i p_{ik}^{(m)} t_{ik}^{(m)}} \tag{18}$$

and the bias field parameters by

$$\begin{bmatrix} A^t W_{11}^{(m)} A & A^t W_{12}^{(m)} A & \cdots \\ A^t W_{21}^{(m)} A & A^t W_{22}^{(m)} A & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_1^{(m)} \\ \boldsymbol{c}_2^{(m)} \\ \vdots \end{bmatrix} = \begin{bmatrix} A^t \left(W_{11}^{(m)} \boldsymbol{r}_{11}^{(m)} + W_{12}^{(m)} \boldsymbol{r}_{12}^{(m)} + \cdots\right) \\ A^t \left(W_{21}^{(m)} \boldsymbol{r}_{21}^{(m)} + W_{22}^{(m)} \boldsymbol{r}_{22}^{(m)} + \cdots\right) \\ \vdots \end{bmatrix} \tag{19}$$

with

$$\begin{aligned} A &= \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \phi_3(x_1) & \cdots \\ \phi_1(x_2) & \phi_2(x_2) & \phi_3(x_2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \\ W_{qr}^{(m)} &= (\text{diag})\left((w_i^{qr})^{(m)}\right), \\ \boldsymbol{r}_{qr}^{(m)} &= [\boldsymbol{y}]_r - \frac{\sum_k (w_{ik}^{qr})^{(m)} \left[\boldsymbol{\mu}_k^{(m)}\right]_r}{\sum_k (w_{ik}^{qr})^{(m)}} \\ (w_i^{qr})^{(m)} &= \sum_k (w_{ik}^{qr})^{(m)}, \\ (w_{ik}^{qr})^{(m)} &= p_{ik}^{(m)} t_{ik}^{(m)} \left[\left(\boldsymbol{\Sigma}_k^{(m)}\right)^{-1}\right]_{qr} \end{aligned}$$

Acknowledgment

The manual segmentations were performed by Dr. B. Sharrack (Department of Neurology, United Medical & Dental Schools of Guy's & St Thomas's) and C. Rush (Kent Institute of Medicine and Health Sciences, University of Kent).

References

[1] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS)," *Neurology*, vol. 33, pp. 1444–1452, Nov. 1983.

[2] A. C. Evans, J. A. Frank, J. Antel, and D. H. Miller, "The role of MRI in clinical trials of multiple sclerosis: Comparison of image processing techniques," *Ann. Neurol.*, vol. 41, no. 1, pp. 125–132, Jan. 1997.

[3] M. Filippi, M. A. Horsfield, P. S. Tofts, F. Barkhof, A. J. Thompson, and D. H. Miller, "Quantitative assessment of MRI lesion load in monitoring the evolution of multiple sclerosis," *Brain*, vol. 118, pp. 1601–1612, 1995.

[4] D. W. Paty and D. K. B. Li, "Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial," *Neurology*, vol. 43, pp. 662–667, Apr. 1993.

[5] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, pp. 885–896, Oct. 1999.

[6] ——, "Automated model-based tissue classification of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, pp. 897–908, Oct. 1999.

[7] P. J. Huber, "Wiley series in probability and mathematical statistics," in Robust Statistics. New York: Wiley, 1981.

[8] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian mixture density modeling, decomposition, and applications," *IEEE Trans. Image Processing*, vol. 5, pp. 1293–1302, Sept. 1996.

[9] P. Schroeter, J.-M. Vesin, T. Langenberger, and R. Meuli, "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 17, no. 2, pp. 172–186, Apr. 1998.

[10] R. Guillemaud and M. Brady, "Estimating the bias field of MR images," *IEEE Trans. Med. Imag.*, vol. 16, pp. 238–251, June 1997.

[11] D. L. Wilson and J. A. Noble, "An adaptive segmentation algorithm for time-of-flight MRA data," *IEEE Trans. Med. Imag.*, vol. 18, pp. 938–945, Oct. 1999.

[12] A. C. S. Chung and J. A. Noble, "Statistical 3D vessel segmentation using a Rician distribution," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1999, vol. 1679, Proceedings of Medical Image Computing and Computer-Assisted Intervention—MICCAI'99, pp. 82–89.

[13] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds., "Understanding robust and explanatory data analysis," in *Wiley series in probability and mathematical statistics*. New York: Wiley, 1983.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[15] E. Ising, "Beitrag zur theorie des ferromagnetismus," *Zeitschrift für Physik*, vol. 31, pp. 253–258, 1925.

[16] J. Ashburner, K. Friston, A. Holmes, and J.-B. Poline, *Statistical Parametric Mapping*. London, U.K.: Wellcome Dept. Cogn. Neurol., Univ. College London.

[17] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," *Proc. IEEE Nuclear Science Symp. Medical Imaging Conf.*, pp. 1813–1817, 1993.

[18] *Matlab*. Natick, MA, USA: The MathWorks Inc..

[19] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, pp. 187–198, Apr. 1997.

[20] "European project on brain morphometry,", BIOMORPH, EU-BIOMED2 project nr. BMH4-CT96-0845, 1996–1998.

[21] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[22] A. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imag.*, vol. 13, pp. 716–724, Dec. 1994.

[23] G. J. McLachlan and D. Peel, "Robust cluster analysis via mixtures of multivariate t-distributions," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1998, vol. 1451, Proceeding of Joint IAPR International Workshops SSPR'98 and SPR'98, pp. 658–666.

[24] R. Davé and R. Krishnapuram, "Robust clustering methods: A unified view," *IEEE Trans. Fuzzy Systems*, vol. 5, pp. 270–293, May 1997.

[25] J. K. Udupa, L. Wei, S. Samarasekera, Y. Miki, M. A. van Buchem, and R. I. Grossman, "Multiple sclerosis lesion quantification using fuzzy-connectedness principles," *IEEE Trans. Med. Imag.*, vol. 16, pp. 598–609, Oct. 1997.

[26] B. Johnston, M. S. Atkins, B. Mackiewich, and M. Anderson, "Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI," *IEEE Trans. Med. Imag.*, vol. 15, pp. 154–169, Apr. 1996.

[27] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, and A. C. Evans, "Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images," *IEEE Trans. Med. Imag.*, vol. 14, pp. 442–453, Sept. 1995.

[28] R. Kikinis, C. R. G. Guttmann, D. Metcalf, W. M. Wells III, G. J. Ettinger, H. L. Weiner, and F. A. Jolesz, "Quantitative follow-up of patients with multiple sclerosis using MRI: Technical aspects," *J. Magn. Reson. Imag.*, vol. 9, no. 4, pp. 519–530, Apr. 1999.

[29] S. Warfield, J. Dengler, J. Zaers, C. R. G. Guttmann, W. M. Wells III, G. J. Ettinger, J. Hiller, and R. Kikinis, "Automatic identification of grey matter structures from MRI to improve the segmentation of white matter lesions," *J. Image-Guided Surg.*, vol. 1, no. 6, pp. 326–338, 1995.

[30] A. Zijdenbos, A. Evans, F. Riahi, J. Sled, J. Chui, and V. Kollokian, "Automatic quantification of multiple sclerosis lesion volume using stereotaxic space," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1996, vol. 1131, Proceedings of Visualization in Biomedical Computing—VBC'96, pp. 439–448.

[31] A. Zijdenbos, R. Forghani, and A. Evans, "Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT," in *Lecture Notes in Computer Science* Berlin, Germany, 1998, vol. 1496, Proceedings of Medical Image Computing and Computer-Assisted Intervention—MICCAI'98, pp. 439–448.

[32] C. R. G. Guttmann, R. Kikinis, M. C. Anderson, M. Jakab, S. K. Warfield, R. J. Killiany, H. L. Weiner, and F. A. Jolesz, "Quantitative follow-up of patients with multiple sclerosis using MRI: Reproducibility," *J. Magn. Reson. Imag.*, vol. 9, no. 4, pp. 509–518, Apr. 1999.

[33] W. M. Wells III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, pp. 429–442, Aug. 1996.

[34] B. Johnston, M. S. Atkins, and K. S. Booth, "Partial volume segmentation in 3-D of lesions and tissues in magnetic resonance images," in *Proc. Medical Imaging 1994: Image Processing*, vol. 2167, 1994, pp. 28–39.

[35] D. Metcalf, R. Kikinis, C. Guttmann, L. Vaina, and F. Jolesz, "4D connected component labeling applied to quantitative analysis of MS lesion temporal development," in *Proc. 14th Annu. Int. Conf. IEEE-EMBS*, vol. 3, 1992, pp. 945–946.

[36] G. Gerig, D. Welti, C. Guttmann, A. Colchester, and G. Székely, "Exploring the discrimination power of the time domain for segmentation and characterization of lesions in serial MR data," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1998, vol. 1496, Proceedings of Medical Image Computing and Computer-Assisted Intervention—MICCAI'98, pp. 469–480.

[37] D. Rey, G. Subsol, H. Delignette, and N. Ayache, "Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1999, vol. 1613, Proceedings of Information Processing in Medical Imaging—IPMI'99, pp. 154–167.

[38] J.-P. Thirion and G. Calmon, "Deformation analysis to detect and quantify active lesions in three-dimensional medical image sequences," *IEEE Trans. Med. Imag.*, vol. 18, pp. 429–441, May 1999.

[39] J.-P. Thirion, "Non-rigid matching using demons," in *Proc. Computer Vision and Pattern Recognition, CVPR'96*, June 1996, pp. 245–251.

[40] F. Bello and A. C. F. Colchester, "Measuring global and local spatial correspondence using information theory," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention—MICCAI'98*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 1998, vol. 1496, pp. 964–973.