Statistical Analysis of Longitudinal MRI Data: Applications for Detection of Disease Activity in MS

Sylvain Prima[1], Nicholas Ayache[2], Andrew Janke[1], Simon J. Francis[1], Douglas L. Arnold[1], and D. Louis Collins[1]

[1] McConnell Brain Imaging Centre, Montreal Neurological Institute
3801 University Street, Montreal, Quebec, Canada
{prima,rotor,simon,doug,louis}@bic.mni.mcgill.ca
[2] INRIA Sophia Antipolis, EPIDAURE Project
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France
ayache@sophia.inria.fr

Abstract. We present a method to detect intensity changes in longitudinal volumetric MRI data from patients with multiple sclerosis (MS). Preprocessing includes spatial and intensity normalization. The intra-subject intensity normalization is achieved using a polynomial least trimmed squares method to match the histograms of all images in the series. Viewing the detection of disease activity in MRI as a *change-point problem*, we present two statistical tests and apply them to a patient's series of grey-level images on a voxel-by-voxel basis. Results are compared with manual lesion segmentation for one MS patient scanned approximately every 5 months for 5 years. Results are also shown for 12 MS patients with 30 monthly scans.

1 Introduction

Motivation: It has been difficult to evaluate the effect of therapy for patients with multiple sclerosis (MS) in clinical trials, since it is a complex disease with a high degree of variability in clinical signs and symptoms that vary over time and between individuals. The most standard clinical tool used to measure impairment and disability in MS is the Expanded Disability Status Scale (EDSS). The EDSS is highly weighted towards motor disability and is notoriously subject to high inter-rater variability [12,23]. The search for better measures of disease burden has turned to MRI, since it was the first method to allow direct visualization of MS plaques *in vivo*. MRI has had a major impact on the diagnosis [5,16,1] and understanding of MS [28,14]. Perhaps most importantly, MRI has shown that clinical measures (attacks) tend to grossly underestimate disease activity, since new lesions on MRI occur with roughly 10 times the frequency of clinical attacks [10]. Since T_2-weighted lesion volume has been used as a surrogate for disease burden in MS, and new lesions are indicative of disease activity, we have been interested in automated techniques for segmentation of active MS lesions.

T. Dohi and R. Kikinis (Eds.): MICCAI 2002, LNCS 2488, pp. 363–371, 2002.

Previous work: In the image processing literature, much effort has been dedicated to the development of techniques for the automatic segmentation of brain structures and lesions in individual MR scans [32,11,27]. For a given patient, a follow-up of this segmented data through time gives an insight into the course of the disease, and allows to monitor its evolution [3]. More specifically, focusing on the evaluation of disease activity, some authors have proposed to use non-rigid matching techniques and deformation field analysis to discriminate static tissues from evolving ones between two consecutive time points [26,19]. An alternative approach has been proposed by Gerig *et al.*, which takes into account the whole time series simultaneously, with a voxel-by-voxel analysis of the intensity profiles through time [8]. Scalar operators are devised and applied to these profiles, indicating the voxels where significant intensity changes occur, conveying an underlying biological process. Dempster-Shafer's theory is then used to fuse the information brought by each operator, leading to a 3D probability map of evolving regions. More recently, other authors have proposed to detect active lesions using an *a priori* knowledge about the evolution process, and taking into account the neighbourhood of each voxel for better rejection of false positives [29,18].

Overview of the paper: Following the idea of a voxel-by-voxel analysis of the intensity profiles, and viewing the detection of activity in time series of MR brain images as a *change point problem*, we introduce two statistical tests to discriminate voxels where an actual biological change occurs. One is very generic, whereas the other is more specific, detecting voxels where the intensity significantly *increases* through time. The latter is particularly suited for the follow-up of MS lesions and brain atrophy in T_2-weighted images, since both lesions and ventricles appear brighter than the surrounding brain structures. Before applying these tests, preprocessing of the MRI data is required to make the voxel intensities comparable over time. This includes bias field correction, registration and intensity normalization (Section 2.1). Subsequently, we devise the statistical tests in Section 2.2. In Section 3, we apply the whole analysis procedure to time series of real MR data of MS patients before concluding in Section 4.

2 Methods

2.1 Preprocessing Steps

Before longitudinal statistical analysis of intensity change, each subject-timepoint data volume must be spatially and intensity normalized to ensure that the intensities used in the statistical test come from the same anatomical point and that intensity variations due to imaging artefacts are minimized. There are a number of steps for both spatial and intensity normalization:

Intensity non-uniform correction: Intensity non-uniformity in MR images, or the so-called MR *shading artefact*, is due to a number of causes during the acquisition of the data. If left uncorrected, such an artifact precludes direct comparison of voxel intensities. The intensity artefact can be modelled as a slowly varying multiplicative bias field. We begin preprocessing by applying the

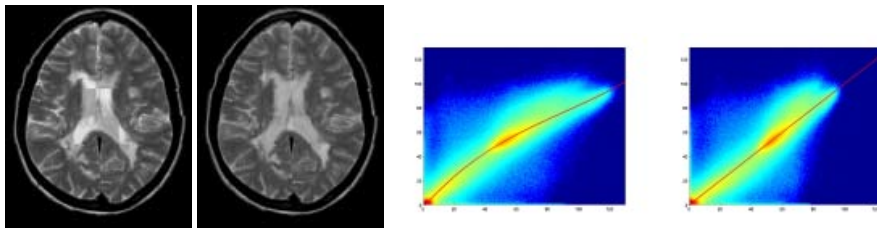
N3 correction algorithm of Sled *et al.* [24] to reduce the effect of this artefact. The N3 algorithm iteratively proceeds by computing the image histogram and estimating a smooth intensity mapping function that tends to sharpen peaks in the histogram.

Intensity normalization: In preparation for registration and subsequent analysis, each image is intensity normalized using a two pass procedure. In order to remove outlier intensity values, the middle 0.1 to 99.9% of the image histogram is mapped to the arbitrary range of 0 to 100, respectively. This procedure eliminates bright voxel values corresponding to blood flow artefacts that reduce the dynamic range of the tissues in the data. The second intensity normalization procedure is described below, after spatial normalization.

Spatial normalization: Each data set is transformed into a standard brain-based coordinate system so that no single data set have a preferred position in processing and so that similar anatomical structures from different data sets are mapped to same spatial position. We have selected the Talairach brain-based coordinate system known as *stereotaxic space* since it has become the defacto standard in the brain mapping community [6,13]. The registration algorithm proceeds with a coarse-to-fine approach by registering subsampled and blurred MRI volumes with an average MRI image, already registered in the stereotaxic coordinate system by optimizing a 9 parameters transform (3 translations, 3 rotations, 3 scalings) [4]. While this procedure is quite robust, we have found that there can be small (sub-millimeter) mis-registrations between the different time-points for a given subject. To address this problem, a second-phase registration is run using the average of all the subject's stereotaxically resampled time-points as a target. The use of the *patient-specific* stereotaxic target reduces the misregistration described above.

Refined intensity normalization: At this point, each data volume has been spatially normalized and an initial intensity normalization has been completed. This first pass intensity normalization is insufficient for comparison between different time points. Actually, even with the same protocol and the same scanner, the intensities within a given brain area are generally different between two acquisitions of the same patient, notably due to drift in receiver coil sensitivity. This is particularly true when the time frequency of the acquisitions is in the range of months or years, as in the case of MS patients. Thus, even though it is classical to assume affine dependence between the intensities of monomodal MR images of the same patient (as often done in registration algorithms [20]), a more complex relationship sometimes holds, making a simple affine mapping of the intensities fail (Figure 1).

Here, we propose to correct the intensities of each image J of the MRI time series with respect to a reference image I: the *patient-specific* stereotaxic target built in the spatial normalization step. As I is the average of all the images of the time series, its signal-to-noise ratio is higher than that of each of the individual MR image used for its computation. Rather than assuming that $J = f(I)$ with f affine ($f = a_1 x + a_0$), we suppose that f is a polynomial of a higher order $d > 1$ ($f = \sum_{k=0}^{d} a_k x^k$) and look for the Least Trimmed Squares (LTS) fit between J



(a) (b) (c) (d)

Fig. 1. Example of intensity normalization using LTS polynomial fitting. (a) Screen-door visualization of two transverse images from two successive time points at the level of the ventricles after the first pass intensity normalization (note the differing intensities in the lateral ventricles). (b) Same images after correction. Note contrast in ventricles is similar, but now the true shape change is evident between the two time points. (c) Joint intensity histogram of the images before correction with the estimated polynomial ($d = 3$). (d) Joint histogram of the two images after correction; it is now close to the line x=y, indicating that the intensities are now matched.

and $f(I)$. This has been previously proposed by Guimond *et al.* for multimodal registration [9]. Routinely, we have found that $d = 3$ yields satisfying results. The LTS regression [21] is far more robust to outliers than the classical Least Squares (LS) method. This is critical here, where biological changes are likely to occur (lesions and/or brain atrophy); these voxels do not fit the polynomial model and must be eliminated from its computation. No explicit solution exists for the computation of the LTS estimate of f. An iterative scheme described by Rousseeuw and Van Driessen [22] locates a (at least local) minimum of the LTS criterion, which amounts to numerous successive LS regressions (that can be solved by standard matrix algebra) on subsets of the image voxels. An example of the procedure is presented in Figure 1.

2.2 Statistical Analysis

In the following, we note $(x_1, \ldots, x_n)$ the series of values at a given voxel after the whole preprocessing has been completed. Among the operators proposed by Gerig *et al.* [8], some do not relate to the temporal pattern of the data, like the variance, or the difference between extremes values in the time series. Some others do, like those based on the fluctuations around the mean value $\bar{x}$ of the time series: high frequency fluctuations are likely to be noise, whereas low frequency variations probably convey actual biological changes. These operators closely relate to classical statistics such as the number of runs, the length of the longest run, the number of runs up and down, *etc.* The distributions of these statistics under the null hypothesis of no intensity change can be computed exactly [25,15], leading to a set of statistical tests termed as *randomness tests*, which are often used in cryptography.

Here, we propose an alternative approach, which consists in considering the problem of activity detection as a *change-point problem*. There is an extensive

literature about these problems [31], particularily met in the field of quality control. In the following, after stating some reasonable hypotheses about the time series, we show how to derive two simple statistics to test a null hypothesis of "no intensity change" against the alternative of "intensity change".

When there is no biological change underlying the considered voxel (null hypothesis H_0), the fluctuations of its intensity through time only relate to the acquisition noise, that can be supposed to be additive, stationary, spatially white and Gaussian with standard deviation σ, as usually stated. The values x_i can then be seen as realizations of n independent normal random variables X_i with distributions $N(\mu_i, \sigma^2)$, where $\mu_1 = \ldots = \mu_n = \mu$. In this framework, detecting an intensity change amounts to detecting a change in the mean value μ. A simple alternative hypothesis H_1 is to consider that there is only one change in the time series, between time points m and $m+1$, with m unknown. The problem can then be stated as follows:

- $H_0 : \mu_1 = \ldots = \mu_n$
- $H_1 : \mu_1 = \ldots = \mu_m$, $\mu_{m+1} = \ldots = \mu_n$, $1 \leq m \leq n-1$, $\mu_m \neq \mu_{m+1}$

A natural way to derive a statistic to test H_0 against H_1 is to compute the ratio of the likelihoods of the data under these two hypotheses respectively. Noting f_0 (resp. f_1) the density of $(X_1, \ldots, X_n)$ under the null (resp. the alternative hypothesis), this ratio is:

$$L(X_1, \ldots, X_n) = \frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)}$$

Further modelings of the unknown nuisance parameters m and $\delta = \mu_{m+1} - \mu_m$ (the amount of change) allow to simplify this expression. First, the possible biological change is *a priori* equally likely to happen at any time; we model m as the realization of a discrete random variable with uniform distribution. Second, we model δ as the realization of a normal distribution $N(0, \tau^2)$. If τ is small, it can be demonstrated that L is an increasing function of a statistic S with a very simple expression:

$$S = \sum_{k=1}^{n} \left(\sum_{i=k+1}^{n} (X_i - \bar{X}) \right)^2$$

Under H_0, and up to a multiplicative constant depending on σ, S converges in distribution to the limiting distribution of Smirnov's ω_n^2 criterion [7]. It is then straightforward to build a test for H_0 against H_1 based on S, high values of S indicating a statistically significant change. This statistic S is very generic, and is tailored to detect increasing as well as decreasing intensities. In the case of T_2-weighted MR images of MS patients, the changes in intensities are more likely to be unilateral ($\delta > 0$), since lesions appear brighter than the white matter where they are generally located. In the same way, brain atrophy implies the growing of the ventricles, which appear also brighter than the surrounding white matter. Thus, one could want to test H_0 against the more restrictive hypothesis H_2 defined as:

– $H_2 : \mu_1 = \ldots = \mu_m$, $\mu_{m+1} = \ldots = \mu_n$, $1 \leq m \leq n-1$, $\mu_m < \mu_{m+1}$

Still using the ratio of the likelihoods, a more specific and thus statistically more powerful test can be built for the purpose of detecting this unilateral change by modifying the hypotheses on the nuisance parameter δ. Modeling δ as the realization of a semi-normal (instead of normal) distribution, we can derive another statistic T:

$$T = \sum_{k=1}^{n} \sum_{i=k+1}^{n} (X_i - \bar{X})$$

Under H_0 and up to a multiplicative constant depending on σ, T follows a Gaussian law $N(0,1)$. Thus, a statistical test can be built based on T for H_0 against H_2, high values of T indicating a statistically significant increase in intensity [2]. In the preliminary set of experiments described in the next section, we give a qualitative comparison of the two tests based on S and T.

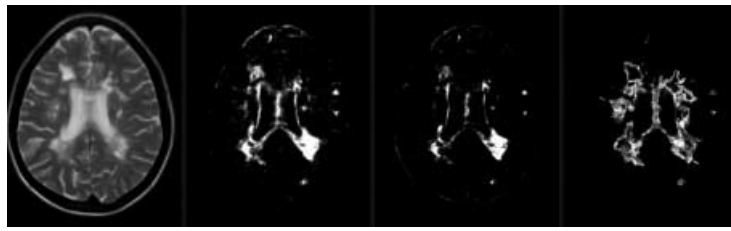
3 Experiments

Data for the first experiment consisted of a T_2-weighted sequence of MR images from a transverse dual-echo, turbo spin-echo sequence (256x256 matrix, 1 signal average, TR/TE1/TE2=2075/32/90 ms, 250mm field of view, 3mm slice thickness). Eleven image volumes over a four year period were acquired of a patient with very active disease. The *patient-specific stereotaxic target* (*i.e.*, the average of the first phase stereotaxic registration procedure of the 11 volumes) is shown in Fig. 2-a. Lesions were labelled manually in each image volume. A lesion difference map, created by subtracting lesions from time $n-1$ from n was used to approximate new lesion activity. The sum of the ten difference maps is shown in Fig. 2-d and represents disease activity over the 4 year period. The tests based on the statistics T and S were applied and the results are shown in Fig. 2-b and 2-c, respectively. (Note that while only 2D images are shown, all processing is completed in 3D.)

Data for the second experiment comes from the PRISMS Study Group [17]. This data consists in monthly MRI scanning for 30 months for 12 patients from placebo group. The T_2-weighted data was acquired with parameters similar to those above, but with a 5mm slice thickness and a 0.5mm slice gap. Both T and S tests were run for each subject after prepocessing all data as described above. Fig. 3 shows the result of the statistical test associated with S (*i.e.,* detection of a change in any direction). One can see that there are bright pixels due to lesions, brain atrophy and possible mis-registration. These patients have much less disease activity compared to that of experiment one, as demonstracted by fewer and smaller detected regions.

4 Discussion and Future Work

In this paper, we derive two statistics (one-sided, T and two-sided, S) for testing the hypothesis of no change against the hypothesis of one and only one change at



(a) (b) (c) (d)

Fig. 2. Results: (a) Transverse MR image of patient with MS at the level of the lateral ventricles. The image is from the *patient-specific stereotaxic target*, *i.e.*, the average of all time-points for the patient. We display the results of the test based on the statistic T (b), S (c), and the manual label difference image (d). The one-sided statistic T yields more evolving points than the two-sided statistic S. The power of the test based on T is higher, because most of the observed changes are unilateral (lesions intensities are higher than those of the white matter where they are generally located). Both results are qualitatively similar to that of the manual segmentation of the evolving voxels.

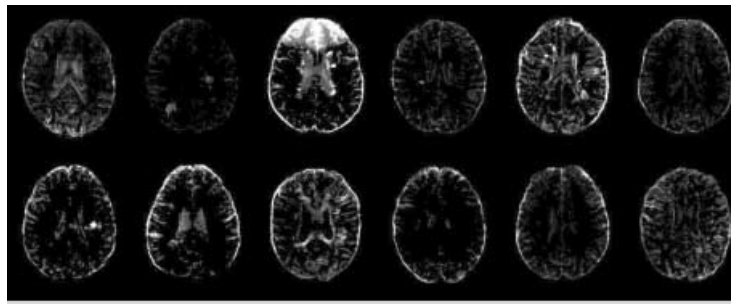


Fig. 3. Results: the images above show transverse slices (z=28mm) through the test result (statistic S) for 12 placebo patients from the PRISMS Study Group [17]. Small focal lesions are apparent in the 3rd, 4rth and 5th images of the first row and in the 1st, 2nd, 3rd and 6th images of the second row. One can also see false positives due to brain atrophy and possible image mis-registration. Note that the bright anterior region in the 3rd image of row 1 is due to a failed image intensity normalization in one of the 30 individual time-points, making these images appropriate for quality control in automated processing.

an unknown point in the time series. For the detection of activity in T_2-weighted images of MS patients, the test based on T is more specific. Other tests could be derived in the same way, to be more specific to the problem which is tackled. The more realistic the alternative hypothesis, the more powerful the corresponding test. In particular, due to the relapsing-remitting course of the disease for most MS patients, tests based on alternative hypotheses including the possibility of more than one change should be statistically more powerful. Non-parametric tests could be also investigated.

We envision four avenues for future work in longitudinal analysis of MRI data from MS patients. First, the statistical tests are applied only on a voxel by voxel basis. While a p-value can be computed from the test results, it must be corrected for multiple comparisons across all voxels of the volume. Bonferroni correction may be too strict; individual voxels are not independent in a statistical sense since their value is likely to depend on their neighbours. We plan to use Gaussian Random Field theory [30] to compute the proper corrections to identify statistically significant changes in intensity due to MS lesions. Second, The tests presented here have been applied to single modality data (*i.e.*, T_2-weighted images only). Since T_2-weighted data is often acquired in a dual echo sequence that also yields registered PD-weighted data, we plan to extend the tests to combine information from multiple modalities. Third, we plan to look into the sensitivity of the one-sided test based on T to detect brain atrophy around the ventricles, where T_2-weighted voxels will change from a medium intensity (tissue) to bright (CSF). Finally, we wish to investigate additional metrics of disease activity. Combination of global or local atrophy metrics with the lesion activity measure described here may result in a better surrogate of disease activity. Such a metric may lead to better understanding of MS pathology and will have important implications for disease prognosis, and monitoring treatment effect in clinical trials.

References

1. F. Barkhof, M. Filippi, D. H. Miller, P. Tofts, L. Kappos, and A. J. Thompson. Strategies for optimizing mri techniques aimed at monitoring disease activity in multiple sclerosis treatment trials. *J Neurol*, 244(2):76–84, Feb 1997.
2. H. Chernoff and S. Zacks. Estimating the current mean of a normal distribution which is subject to changes in time. *Ann. of Math. Stat.*, 35:999–1018, 1964.
3. D. Collins, J. Montagnat, A. Zijdenbos, and A. Evans. Automated Estimation of Brain Volume in Multiple Sclerosis with BICCR. In *IPMI'01*, volume 2082 of *LNCS*, pages 141–147, Davis, USA, June 2001. Springer.
4. D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D inter-subject registration of MR volumetric data in standardized Talairach space. *JCAT*, 18(2):192–205, March/April 1994.
5. F. Fazekas, H. Offenbacher, S. Fuchs, R. Schmidt, K. Niederkorn, S. Horner, and H. Lechner. Criteria for an increased specificity of MRI interpretation in elderly subjects with suspected multiple sclerosis. *Neurology*, 38(12):1822–5, Dec 1988.
6. P. T. Fox, J. S. Perlmutter, and M. E. Raichle. A stereotactic method of anatomical localization for positron emission tomography. *JCAT*, 9(1):141–153, 1985.
7. L. Gardner. On detecting changes in the mean of normal variates. *Ann. Math. Stat.*, 40:116–126, 1969.
8. G. Gerig, D. Welti, C. Guttmann, A. Colchester, and G. Székely. Exploring the discrimination power of the time domain for segmentation and characterization of active lesions in serial MR data. *MedIA*, 4(1):31–42, 2000.
9. A. Guimond, A. Roche, N. Ayache, and J. Meunier. Multimodal Brain Warping Using the Demons Algorithm and Adaptative Intensity Corrections. *IEEE TMI*, 20(1), 2001.
10. C. Isaac, D. K. Li, M. Genton, C. Jardine, E. Grochowski, M. Palmer, L. F. Oger, and D. W. Paty. Multiple sclerosis: A serial study using MRI in relapsing patients. *Neurology*, 38(10):1511–1515, 1988.
11. R. Kikinis, C. Guttmann, D. Metcalf, W. Wells III, G. Ettinger, H. Weiner, and F. Jolesz. Quantitative follow-up of patients with multiple sclerosis using MRI: Technical aspects. *JMRI*, 9(4):519–530, 1999.
12. J. F. Kurtzke. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale. *Neurology*, 33:1444–1452, 1983.
13. J. Mazziotta, A. Toga, A. Evans, P. Fox, and J. Lancaster. A probabilistic atlas of the human brain: theory and rationale for its development. the international consortium for brain mapping. *NeuroImage*, 2(2):89–101, 1995.

14. D. H. Miller, R. I. Grossman, S. C. Reingold, and H. F. McFarland. The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain*, 121(Pt 1):3–24, Jan 1998.
15. A. Mood. The distribution theory of runs. *Ann. Math. Stat.*, 11:367–392, 1940.
16. D. W. Paty. Magnetic resonance imaging in the assessment of disease activity in multiple sclerosis. *Canadian Journal of Neurological Sciences*, 15(3):266–72, Aug 1988.
17. PRISMS-4. Long-term efficacy of interferon-beta-1a in relapsing ms. *Neurology.*, 56(12):1628–1636, 2001.
18. D. Rey, J. Stoeckel, G. Malandain, and N. Ayache. Spatio-temporal Model-based Statistical Approach to Detect Evolving Multiple Sclerosis. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA'01*, Kauia, USA, Dec. 2001.
19. D. Rey, G. Subsol, H. Delingette, and N. Ayache. Automatic Detection and Segmentation of Evolving Processes in 3D Medical Images: Application to Multiple Sclerosis. In *IPMI'99*, LNCS, Visegrád, Hungary, June 1999. Springer.
20. A. Roche, G. Malandain, and N. Ayache. Unifying Maximum Likelihood Approaches in Medical Image Registration. *International Journal of Imaging Systems and Technology: Special Issue on 3D Imaging*, 11:71–80, 2000.
21. P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection.* Wiley Series in Probability and Mathematical Statistics, 1987.
22. P. Rousseeuw and K. Van Driessen. Computing LTS Regression for Large Data Sets. Technical report, Statistics Group, University of Antwerp, 1999.
23. R. Rudick, A. J, C. Confavreux, G. Cutter, G. Ellison, J. Fischer, F. Lublin, A. Miller, J. Petkau, S. Rao, S. Reingold, K. Syndulko, A. Thompson, J. Wallenberg, B. Weinshenker, and E. Willoughby. Clinical outcomes assessment in multiple sclerosis. *Ann. Neurol.*, 40(3):469–79, Sep 1996.
24. J. G. Sled and G. B. Pike. Standing-wave and rf penetration artifacts caused by elliptic geometry: an electrodynamic analysis of mri. *IEEE TMI*, 17(1):87–97, 1998.
25. W. Stevens. Distributions of groups in a sequence of alternatives. *Ann. Eugen.*, 9:10–17, 1939.
26. J.-P. Thirion and G. Calmon. Deformation Analysis to Detect and Quantify Active Lesions in Three-Dimensional Medical Image Sequences. *IEEE TMI*, 18(5), May 1999.
27. K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE TMI*, 20(8):677–688, Aug. 2001.
28. C. J. Wallace, T. P. Seland, and T. C. Fong. Multiple sclerosis: The impact of MR imaging. *AJR Am. J. Roentgenol.*, 158:849–857, 1992.
29. D. Welti, G. Gerig, E.-W. Radue, L. Kappos, and G. Szekely. Spatio-temporal Segmentation of Active Multiple Scleroris Lesions in Serial MRI Data. In *IPMI'01*, volume 2082 of *LNCS*, pages 438–445, Davis, USA, June 2001. Springer.
30. K. Worsley, S. Marrett, P. Neelin, A. Vandal, K. Friston, and A. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73, 1996.
31. S. Zacks. Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation. In M. Rizvi, J. Rustagi, and D. Siegmund, editors, *Recent advances in statistics, Herman Chernoff Festschrift*, pages 245–269, New York, USA, 1983. Academic Press.
32. A. Zijdenbos, A. Jimenez, and A. Evans. Pipelines: Large scale automatic analysis of 3d brain data sets. In A. Evans, editor, *4th International Conference on Functional Mapping of the Human Brain*, Montreal, June 1998. Organization for Human Brain Mapping. abstract no. 783.