Automatic Change Detection in Multi-Modal Serial MRI: Application to Multiple Sclerosis Lesion Evolution

Marcel Bosc\textsuperscript{a,b,1}, Fabrice Heitz\textsuperscript{a}, Jean-Paul Armspach\textsuperscript{b}, Izzie Namer\textsuperscript{b}, Daniel Gounot\textsuperscript{b} Lucien Rumbach\textsuperscript{c}

\textsuperscript{a}Laboratoire des Sciences de l’Image et de la Télédetection (LSIIT) UMR-7005 CNRS, 67400 Illkirch, France
\textsuperscript{b}Institut de Physique Biologique UMR-7004 CNRS, 67085 Strasbourg, France
\textsuperscript{c}Centre Hospitalier Universitaire, Service de Neurologie, 25030 Besancon, France
Abstract

The automatic analysis of subtle changes between MRI scans is an important tool for assessing disease evolution over time. Manual labeling of evolutions in 3D data sets is tedious and error prone. Automatic change detection, however, remains a challenging image processing problem. A variety of MRI artifacts introduce a wide range of unrepresentative changes between images, making standard change detection methods unreliable. In this paper we describe an automatic image processing system that addresses these issues. Registration errors and undesired anatomical deformations are compensated using a versatile multi-resolution deformable image matching method that preserves significant changes at a given scale. A non-linear intensity normalization method is associated with statistical hypothesis test methods to provide reliable change detection. Multi-modal data is optionally exploited to reduce the false detection rate. The performance of the system was evaluated on a large database of 3D multi-modal, MR images of patients suffering from relapsing remitting Multiple Sclerosis (MS). The method was assessed using Receiver Operating Characteristics (ROC) analysis, and validated in a protocol involving two neurologists. The automatic system outperforms the human expert, detecting many lesion evolutions that are missed by the expert, including small, subtle changes.

Key words: Change detection, Deformable matching, Statistical tests, Serial MRI, Multiple Sclerosis

---

Email address: bosc@ipb.u-strasbg.fr (Marcel Bosc).

1 This research was supported by the Ligue Française Contre la Sclérose en Plaques, and the CNRS (multilaboratory project IRMC).
1 Introduction

The constant improvement in spatial resolution and the availability of multi-modal imaging hardware yields an increasing amount of three dimensional image data. Manual interpretation of huge amounts of data is tedious and error prone. A major challenge for medical imaging is the development of reliable techniques for automatic detection of changes. Early detection of anatomical modifications enables faster diagnosis. Once a pathology has been diagnosed, long-term monitoring of disease evolution is also essential.

A direct comparison between successively scanned images is generally not possible. Patient position is never identical, acquisition parameters may drift between scans, and a variety of complex global or local deformations of anatomical structures may be observed.

During manual change labeling, the expert implicitly uses prior high-level knowledge to compensate for some of these undesired changes. He can visually correct for repositioning errors, and uses anatomical knowledge to identify and reject certain artifacts. Techniques for identifying and correcting each important artifact source are thus necessary.

This work is centered on image processing techniques that correct these artifacts and decide whether or not changes are significant. The proposed approach is comprehensive and does not require any knowledge of the acquisition process. Raw MRI data that could be useful to eliminate certain artifacts are thus not necessary. The approach may be applied to other change detection problems, the artifacts addressed here being common to many medical imaging devices.

The system was evaluated in the specific case of Multiple Sclerosis (MS) lesion evolutions analysis. MRI has become a sensitive, objective marker of MS evolution over time. Serial multi-modal RARE, T1-weighted (GE3D), and FLAIR 3D MR images were obtained from 8 patients suffering from relapsing remitting MS (Fig. 1). The performance of the system was evaluated on a large (over 200) database of $128^3$ 3D MR images, using Receiver Operating Characteristics (ROC) analysis. The performances of the automatic system compared favorably with manual labellings provided by two neurologists. The automatic system outperformed the human expert for several cases exhibiting small, subtle changes (see for instance Fig. 2).
Fig. 1. Large evolution of an MS lesion in 3 different modalities (FLAIR, RARE, GE3D). The images labeled “2” were taken two months after the images labeled “1”. Lesion evolutions of this size are easily detected by manual or automatic methods (lesion is visible at cross hairs).

Fig. 2. Automatically detected periventricular MS lesion evolution. This subtle evolution was missed (but validated \emph{a posteriori}) by the expert (view enlarged for visualization).
2 Related Work

A variety of different approaches to change detection in serial MRI may be found in the literature. Existing approaches, often devoted to particular issues with specific tools, may be divided into two major categories. A first category corresponds to approaches looking for large structural changes. Methods from the second category focus on smaller, more localized changes. An example of large structural change detection is Smith et al. (2000), where global brain volume changes are determined by estimating the motion of extracted brain/CSF (Cerebro-Spinal Fluid) boundaries. Here, we will consider the detection of small (eventually subtle) changes. Subtle changes are changes that are hard to detect by the human expert at first sight.

Another major divide in change detection techniques, is whether one is considering long time sequences, like in fMRI, or short (typically two frames) sequences. Change detection on long time sequences tends to focus on signal processing of the time series at a given spatial location of the 3D image set. Such an approach is employed in Gerig et al. (1999) for MS monitoring, using statistical tests for change detection over the time series. Hereafter we will consider methods that are able to process short time series of two images.

Most change detection systems use registration as a preliminary step to obtain accurate geometrical alignment of the images before image comparison. This first (optional) step is generally followed by an inter-image comparison step.

2.1 Inter-image registration

Rigid (rotational and translational) registration has been used in Hajnal et al. (1995); Curati et al. (1996); Rey et al. (2002), but it does not compensate for the significant voxel size changes that occur in commercial scanners. Affine (scale and shear) registration gives more accurate results (Freeborough et al., 1996; Lemieux et al., 1998), but is not sufficient to compensate for more complex deformations that may occur. These complex deformations can be of pathological nature (global brain atrophy (Smith et al., 2000)), may come from natural causes (gravity (Hajnal et al., 1995), hydrostatic pressure changes, dehydration...) or be due to acquisition artifacts (incorrect shim, gradient calibration drift (Freeborough et al., 1996; Smith et al., 2000)). To properly compensate for these complex phenomena, it is necessary to resort to deformable matching methods, which are able to correct most large scale artifacts, while preserving small significant changes. Deformable matching is commonly used for inter-subject matching (Toga and Thompson, 1999; Thompson and Toga, 1999; Thirion, 1998) and computed deformation fields have also been used for
change detection (see section 2.2.2). However, deformable matching has seldom been considered as a realignment step for intra-patient change-detection, since deformable registration may cancel the small significant inter-image changes. A solution, adopted here, is to consider hierarchical matching techniques, which are able to register image structures up to a given controlled resolution. In section 3.5 we describe an iterative registration procedure, developed by the authors (Musse et al., 2001, 2002), which enables such an accurate and careful compensation of artifacts.

2.2 Inter-image comparison

The second step in the design of a change detection method is to devise a technique for comparing the two images. Three main families of approaches are considered.

2.2.1 Segmentation-Based Approaches

In the specific case of lesion monitoring, a first approach is to manually, semi-automatically, or fully-automatically segment lesions and then compare them to determine changes.

Manual outlining was used a few years ago, but is an extremely time consuming process; semi-automated methods were then developed (Kohn et al., 1991; Lim and Pfefferbaum, 1989; Mitchell et al., 1996; Vannier et al., 1985). More recently, automated procedures have been proposed (Kamber et al., 1995; Kikinis et al., 1999; Goldberg-Zimring et al., 1998; Grimaud et al., 1996; Udupa et al., 1997; Van Leemput et al., 2001).

Automatic segmentation of brain structures and MS lesions remains a challenging task. Large lesions may be adequately extracted, but small lesions are difficult to distinguish from noise and other structures, particularly in modalities where lesions have the same intensity values as other brain structures. Once lesions are segmented, their evolution may be determined by comparing lesion load. These methods generally do not use prior inter-image registration, therefore precise evaluation of intensity evolution within a lesion may be difficult.

2.2.2 Deformation Field-Based Approaches

This second family of approaches is based on the estimation and analysis of inter-image deformation fields. Deformable registration (Musse et al., 2001; Thirion, 1998; Thompson and Toga, 1999), when performed at a sufficiently
fine resolution level, will warp one image onto another, effectively canceling out most inter-image changes. The computed transformation then contains interesting information on the relative changes between both images. Specifically, lesion evolutions may appear as contractions or dilations in the inter-image deformation field. This approach is akin to optical flow interpretation techniques used in video image motion analysis (Mitiche and Bouthemy, 1996). Several different variants to this approach have been considered (Thirion and Calmon, 1999; Rey et al., 2002; Freeborough and Fox, 1998; Ashburner and Friston, 2000; Davatzikos et al., 2001). For instance, both Thirion et al. (Thirion and Calmon, 1999; Rey et al., 2002) and Freeborough et al. (Freeborough and Fox, 1998) combine non-rigid deformation computation and flow field analysis to detect evolving lesions or to quantify measurement of volume variations. The Jacobian operator is used to characterize local volume changes that are related to evolutions. Another related example is VBM (Ashburner and Friston, 2000; Davatzikos et al., 2001), where deformations of segmented regions are analyzed to determine regional atrophy.

2.2.3 Direct Intensity Comparison

This last family of approaches relies on a direct comparison of intensities at the voxel level or on small regions. These approaches are comprehensive in that they do not focus on a particular type of evolution.

A first step consists in compensating for global intensity changes (intensity normalization). This step may be conducted before and/or after inter-image registration. Existing methods for intensity normalization apply linear intensity correction functions. We found, however, that intensity differences may be highly non-linear for successive acquisitions (Fig. 7). In section 3.6, we develop an original joint histogram-based technique for compensating such non-linear intensity differences.

Once images are realigned and intensities corrected, they must be compared. This may be done using either a deterministic or a probabilistic approach. Most methods use simple deterministic image subtraction followed by manual or automatic thresholds (O’Riordan et al., 1998). Probabilistic methods (Hsu et al., 1984; Lemieux et al., 1998) build a statistical model of intensities and noise that is used to determine if a valid change is observed.

The approach proposed here is based on a statistical change detection scheme which extends previous work (Hsu et al., 1984; Lemieux et al., 1998) by handling multi-modal image data, using multidimensional Gaussian models of image intensities. A significant decrease of the false detection rate is observed, with respect to standard single-modal approaches.
3 Materials and Methods

3.1 Subjects

Image data was obtained from 8 patients suffering from relapsing remitting MS. Images were taken once every two months during a two year period. Each exam consisted of images in three modalities.

3.2 Imaging

Isotropic 3D MR images were acquired with a standard 2 Tesla whole body imager (Tomikon S200 Bruker, Karlsruhe, Germany) using a head transmitter and receiver. Three scans were performed at each exam. The first scan was a transverse 3D RARE image (FOV = 25.6 cm ; matrix size 128 x 128 ; slice thickness 2 mm ; inversion time 1.3 seconds), the second, a 3D fluid attenuated inversion recovery (FLAIR) image (FOV = 25.6 cm ; matrix size 128 x 128; slice thickness 2 mm) and the third, a T1-weighted 3D image (GE3D) (FOV = 25.6 cm ; matrix size 128 x 128 x 128). No attempt was made to achieve precise repositioning between scans or between exams.

3.3 Image Processing : Overview

Our change detection approach can be split into four main steps:

- Brain extraction (section 3.4).
- Repositioning and correction of deformation artifacts (section 3.5):
  - Iterative affine registration (section 3.5.3).
  - Deformable registration (section 3.5.4).
  - Image resampling (section 3.5.5).
- Intensity normalization (section 3.6).
- Single or multi-modal statistical change detection test (section 3.8).

Figure 3 summarizes the different image processing steps, which are described in the following paragraphs. All processing is fully automatic: no user intervention is required at any moment.

Each patient has been imaged at several (approximately 12) time-points. At each time-point three images of different modalities have been taken (see section 3.1). All of the images go through the brain extraction and registration steps. Time points are then coupled into successive pairs: (exam-1,exam-2)
Fig. 3. Flowchart showing the different steps of the change detection method (see text). For simplicity, only one modality is shown.

(exam-2,exam-3) ... (exam-11,exam-12). Exam pairs go through the remaining processing steps and are finally compared. An image database mechanism was implemented in order to manage the large amount of images and processing steps involved.

3.4 Brain Extraction

The goal is to compare image intensities at identical anatomical positions. We therefore explicitly choose to ignore and compensate for global brain volume change. Besides, we have observed significant brain movement within the skull on animations of rigidly-registered heads. We must therefore register brains and not the surrounding skull and background noise, which would perturb the registration.

The first step is thus to extract the head from background noise. The extraction does not need to be extremely accurate, but needs to be reliable, given the large amount of images that will be processed. This rough segmentation of the head is obtained using low level image processing techniques: Otsu thresholding (Otsu, 1979), erosion and hole filling.
Fig. 4. Atlas-based brain extraction. The transformation field between the reference image and the source (patient) image is estimated using deformable matching. This transformation field is then used to map the reference brain segmentation (atlas) onto the source (patient) image.

The second step consists in extracting the brain (figure 4). An atlas-based segmentation is used to this end (Musse et al., 2002). The atlas is here a reference image that has been previously semi-automatically segmented using a watershed-based approach (Bueno et al., 2001). This reference image is registered onto the patient image we wish to segment using deformable registration (section 3.5.4). Applying the resulting deformation field on the atlas segmentation provides the desired extraction of the brain (Musse et al., 2002) (see figure 4).

3.5 Repositioning and Correction of Deformation Artifacts

3.5.1 Overview

Automatic change detection is extremely sensitive to mis-registration errors. As noted by Lemieux et al. (1998), in regions of high intensity gradient (such as lesion edges), positional shifts of less than .05 voxels may produce spurious detections. Therefore we are clearly working in a sub-voxel context, which implies the need for high quality interpolation (section 3.5.5) and a sophisticated model for registration.

Conventional rigid (rotational and translational) or affine (scale and shear) registration are insufficient for modeling the complex deformations that appear between MRI scans (see section 2.1 for a description of these deformations). To correct these complex deformations, we use an iterative affine image matching method (section 3.5.3), followed by hierarchical deformable registra-
3.5.2 Reference Positions for Registration

As noted in Smith et al. (2000), it is important that all images undergo equivalent processing steps. If a given image of a patient was to be chosen as a positional reference for registering his own images, some images would undergo geometrical transforms, while the reference image would not. Therefore, a common reference position is chosen, and care is taken so that the actual
registration of the images of the dataset is done with reliable intra-patient single-modal registration.

Images, in three modalities, belonging to a patient not in the database are selected. These images are then carefully registered on each other using affine multi-modal registration (figure 5 step 1). These three images will be used as absolute reference positions for the whole database. Note that the intra-patient multi-modal affine registration assures an accurate multi-modal fit. Particularly, it compensates for voxel size errors that would not be addressed by rigid registration.

The registration of a given patient time series is done in two steps. First a set of three patient images of all three modalities are arbitrarily chosen as the patient references. These patient reference images are registered (using inter-patient single-modal affine registration, step 2) onto their respective absolute reference images (figure 5). Inter-patient affine registration may be inaccurate, so we ensure a tight inter-modal fit, with a further multi-modal intra-patient affine registration (figure 5 step 3).

The advantage of this scheme is that we are now able to register the patient images using accurate and reliable intra-patient single-modal affine and deformable registration (figure 5 step 4). Moreover, all registered images in the multi-patient database are in the same (absolute) geometrical position.

In our current implementation, all registration algorithms are voxel-based and rely on the minimization of global energy (cost) functions, with standard sequential simplex optimization. Quadratic cost-functions are considered for single-modal registration and the mutual information criterion is used for multi-modal registration.

3.5.3 Robust Iterative Affine Registration

The deformable registration step (figure 5 step 4) requires a good initial fit. This initial fit is obtained using affine registration, which corrects errors due to voxel size drift. However, differences in the brain extractions between two images can lead to an important bias in registration. To cope with this problem an iterative adjustment of the segmentation is embedded in the registration scheme. Each iteration step is composed of an affine registration followed by a recomputation of the brain extractions. The new brain extraction is the intersection of the brains of the registered images. In practice, it was observed that brain extractions did not vary significantly after two iterations.
3.5.4 Deformable Registration

The deformable matching method (Musse et al., 2001, 2002), is a hierarchical (coarse-to-fine) registration process, that estimates a parametric displacement field decomposed on multiresolution B-spline basis functions. At a given resolution level, the method minimizes a quadratic energy function using modified quasi Newton optimization. The final resolution level may be chosen according to the desired scale of the image structures to be matched. Stopping the algorithm at a low resolution level will correct large scale deformations whereas continuing to higher resolutions will progressively compensate for smaller differences between the registered images. The choice of the adequate resolution level is discussed in section 5.

3.5.5 Resampling

As noted previously, high quality, sub-voxel registrations are required. We therefore need an efficient interpolation strategy. The theoretically perfect sinc interpolator must, in practice, be truncated. This gives incorrect results and is computationally inefficient. Many interpolation methods have been proposed (Lehmann et al., 1999). Recently, Thevenaz and Unser (Thevenaz et al., 2000) have proposed new innovative and efficient interpolation techniques. Traditionally, interpolation is expressed as the reconstruction of a continuous domain image \( I(x) \) from a sampled image \( I_k \) using an interpolating function \( \phi_{int} \) (sinc for example):

\[
I(x) = \sum_{k \in Z^3} I_k \phi_{int}(x - k)
\]  

(1)

Here, \( I_k \) are the values of \( I(x) \) on a discrete grid of points. The problem was reformulated in Thevenaz et al. (2000) as:

\[
I(x) = \sum_{k \in Z^3} c_k \phi(x - k)
\]  

(2)

where the \( c_k \) are computed coefficients that are no longer the exact values of the image on the discrete grid, and therefore \( \phi \) no longer needs to be a strictly interpolating function. This change of perspective leads to new, fast and high quality interpolation schemes. The interpolation is decomposed into two stages, a fast recursive prefiltering stage, and a conventional interpolation phase using a non-interpolating kernel. In our change detection framework, a global geometric transform is computed by composing the transforms computed during the affine and deformable registration steps. A 5 degree box spline kernel (Thevenaz et al., 2000) is used in the last deformable registration step.
3.6 Intensity Normalization

Image intensity of identical tissues may change between successive scans. To correct this problem previous work (Lemieux et al., 1998) used linear mapping between image intensities: the corrected intensity $I_0'(s) = f(I_1(s))$ was a linear function $f(x) = ax + b$ of the original intensity $I(s)$. However, we found that a linear model was insufficient. We propose, here, a non-linear normalization method relying on the estimation of joint probability distributions.

3.6.1 Estimation of Joint Probability Distributions

Histogram, as well as joint histogram computation is conventionally done using standard parametric or nonparametric techniques for estimating probability density functions (pdfs) (Duda et al., 2001). These techniques consider separate independent samples of the distribution. A large number of samples is needed for an adequate estimation, especially for multidimensional pdfs. Conventional histogram estimation techniques do not use the fact that we are working with discrete images, and that a discrete grid sampled image $I_k, k \in Z^3$ is actually a dual representation of a continuous-domain function $I(s), s \in R^3$ (through the Shannon sampling theorem). This means that an infinitely large number of samples are available for accurate histogram evaluation (figure 6). It also means that conventional image histogram computation is incorrect, especially when high frequencies are present. Theoretically, using conventional pdf estimation techniques, with highly supersampled images will give good approximations of the underlying probability distribution. In practice this requires a large amount of computation time. Another approach is to analytically compute the histogram of the continuous-domain image. This is feasible for one dimensional linear interpolation schemes. Locally, in the one-dimensional case $I_1$ and $I_1'$ are both linear with respect to the spatial coordinate $x$: $I_1(x) = ax + b$, $I_1'(x) = cx + d$ and therefore $I_1' = \frac{c}{a}I_1 + d - \frac{b}{a}$. The local contribution to the joint histogram $j(u, v)$ is therefore a line segment. Note that in the conventional (discrete) approach, where no interpolation is used, the local contribution to the joint histogram would have been a single point instead of a line segment. This operation is done in all three dimensions and results are averaged. Resulting histograms agree with the histograms of the supersampled images, and are of much higher quality than conventional histograms.

3.6.2 Non-linear Normalization

The joint histogram (Fig. 7) is an estimate of the joint probability distribution $j(u, v) = p(I_1(s) = u, I_2(s) = v) : s \in \Omega$ of intensities in two images. For
similar images \( p(I_1(s), I_2(s)) \) is high for points near \( u = v \). If the intensity of a tissue changes over time, due for instance to sensor inconsistencies, the joint histogram of two images taken at separate time points will undergo a non-linear transformation. We have repeatedly observed this artifact (Fig. 7) that can be a considerable problem in later stages of change detection. To correct it we estimate an intensity correction function \( f \) that maximizes the probability \( f(u) = \arg \max_v j(u, v) \). The intensity corrected image is then \( I'_1(s) = f(I_1(s)) \). In practice, a robust estimator, that is less sensitive to outliers, is used. The median value of the \( v \rightarrow j(u, v) \) gives good results.

### 3.7 RF Artifact

Remaining artifacts include image intensity non-uniformities, which result from inhomogeneities in the RF field. The major problem in eliminating this artifact is that artifactual intensity variations are hard to distinguish from real image intensity variations (Cohen et al., 2000). In our framework, however, a simple approach is possible. Once images are registered and intensity normalized, a simple subtraction gives us an image that is composed of noise,
residual artifacts and, eventually real image changes. Residual artifacts in
the brain region are essentially movement and pulsation artifacts, which have
zero mean and medium-high spatial frequency. Real image changes are small
scale compared to the RF artifact. Applying a large low-pass Gaussian filter
($\sigma = 15mm$) on this difference image, easily isolates the RF artifact from the
previously described changes.

3.8 Change Detection

Once the major artifacts have been removed, the images must be compared.
That is, we must decide at each point ($s$) whether or not there has been a
genuine change. The simplest approach is direct point by point subtraction.
However, this gives noisy results, and does not provide an objective decision
criterion. Statistical change detection theory provides a sound framework for
this purpose (Kay, 1998). We consider a fixed size 3D window $W_s$ centered
on a voxel $s$ (figure 8). In window $W_s$, a simple statistical model of intensity
distribution is considered. We chose a window size of 3x3x3 voxels (see section
5 for a discussion on the effects of this parameter). The generalized likelihood
ratio test (GLRT) (Hsu et al., 1984; Kay, 1998) computes the ratio of the
probabilities of two hypotheses:

$H_0$: there is no change between $I_1$ and $I_2$ inside $W_s$.
$H_1$: there is a significant change between $I_1$ and $I_2$ inside $W_s$.  

Fig. 8. Change detection considers a statistical model of intensities in a 3D window
$W_s$ centered on $s$
The probabilities are expressed in terms of parametrical pdfs. $H_0$ is interpreted using the pdf of $I_1$ and $I_2$ with the same parameter $\theta_0$ (no change), $H_1$ is interpreted using the pdf of $I_1$ and $I_2$ with two different parameters $\theta_1 \neq \theta_2$.

In the GLRT, the unknown parameters $\theta$ of the distributions are replaced with their maximum likelihood estimates $\hat{\theta}$ (Kay, 1998). The generalized likelihood ratio is then:

$$R_{GLRT} = \frac{p(I_1; \hat{\theta}_1)p(I_2; \hat{\theta}_2)}{p(I_1; \hat{\theta}_0)p(I_2; \hat{\theta}_0)}$$  \hspace{1cm} (3)$$

where $I_1$ and $I_2$ are respectively the values of $I_1$ and $I_2$ in $W_s$.

### 3.8.1 Single-Modal Change Detection

We chose a simple model for the parametric pdfs: the intensity within $W_s$ is modeled as a constant value $\mu$ plus i.i.d. zero mean additive Gaussian noise. Several studies, in particular in motion detection from video image sequences, have shown that simple intensity models (such as the constant plus Gaussian i.i.d. noise model) are far more robust than more specific models involving more parameters (Mitiche and Bouthemy, 1996). Higher order (i.e. linear or non linear) intensity models bring only marginal improvement in the detection results. For example, this simple intensity model is effective even if the noise is not perfectly Gaussian. The noise variance is supposed constant throughout the image and is considered to be identical in both images. It is estimated with a standard estimator by using the difference image described in section 3.7, that contains residual noise. Therefore the only pdf parameter $\theta$ that must be estimated is the average $\theta = \mu$ value. The maximum likelihood estimate of $\mu$ is

$$\hat{\mu} = \frac{\sum_{p \in W_s} I_p}{n}$$

where $n$ is the number of voxels in $W_s$.

Taking the log of equation 3 gives the log-likelihood ratio:

$$l = \frac{1}{2\sigma^2} \sum_{p \in W_s} \left[ -(I_1(p) - \hat{\mu}_1)^2 - (I_2(p) - \hat{\mu}_2)^2 + (I_1(p) - \hat{\mu}_0)^2 + (I_2(p) - \hat{\mu}_0)^2 \right]$$
which simplifies to:

\[
H_1 \\
\frac{\sqrt{n}}{2\sigma} |\hat{\mu}_2 - \hat{\mu}_1| > \lambda \\
< \lambda
\]

(4)

\(H_1\) is chosen if \(l > \lambda\), \(H_0\) is chosen if \(l < \lambda\) where \(\lambda\) is a threshold.

3.8.2 Multi-Modal Change Detection

The single-modal change detection framework presented above may be extended to the multi-modal case. For \(m\) modalities, each image value is considered as an \(m\)-dimensional vector. The probability of observing intensity values \(I(p), p \in W_s\) for a given model average \(\mu\) is

\[
p(I; \mu) = \left( \frac{1}{\sqrt{(2\pi)^m|C|}} \right)^n e^{-\frac{1}{2} \sum_{p \in W_s} (I(p) - \mu)^T C^{-1} (I(p) - \mu)}
\]

(5)

where \(n\) is the number of voxels in \(W_s\), and \(C\) is the covariance matrix of the noise. The covariance matrix is also assumed constant over the whole image and estimated using standard ergodic estimators (Duda et al., 2001; Kay, 1998). The multi-modal log likelihood ratio becomes:

\[
l = -\frac{1}{2} \sum_{p \in W_s} \left( (I_1(p) - \hat{\mu}_1)^T C^{-1} (I_1(p) - \hat{\mu}_1) + (I_2(p) - \hat{\mu}_2)^T C^{-1} (I_2(p) - \hat{\mu}_2) - (I_1(p) - \hat{\mu}_0)^T C^{-1} (I_1(p) - \hat{\mu}_0) - (I_2(p) - \hat{\mu}_0)^T C^{-1} (I_2(p) - \hat{\mu}_0) \right)
\]

(6)

Please note that the relative “weight” given to each modality by this method depends on the noise covariance matrix. Noisy modalities will therefore be down-weighted by this approach.

3.9 Post-Processing

The change detection step labels individual voxels that have changed. In our change detection system we focus on change events which might involve sev-
eral adjacent voxels (see section 4.2). Therefore, neighboring changed voxels are grouped together using the following simple clustering scheme: An high initial likelihood value $x$ is chosen and is iteratively decreased. Each time new voxels are found at value $x$, they are either added to an existing cluster of a neighboring voxel, or, if no neighboring voxel is in a cluster, a new cluster is created. Each group is assigned a value. This value is the largest likelihood ratio of all the voxels in the group. Groups are then sorted in decreasing likelihood order. To the end user is presented a set of evolutions, of decreasing confidence, through which he may interactively cycle (Figure 9). Evaluating, and rejecting false detections is very fast, typically under a minute for an image pair.

4 Results

The automatic detection system has been evaluated by two neurologists on a real-world base of over 200 images as well as on simulated evolutions. We present, here, the rationale behind our methodology.

Evaluating the results of change detection is a difficult task, as there is no solid ground truth available. Indeed, many subtle changes are subject to debate after close examination by experts. Important total lesion load inter-rater variability has, for instance, been reported in Mitchell et al. (1996), and we have also observed significant disagreement between experts.

Two evaluation approaches were considered:

1. using simulated evolutions (section 4.1) on scan pairs from healthy subjects.
2. comparing the results of change detection by an expert with those of automatic change detection (section 4.2)

The use of simulated data has the advantage of providing a well controlled ground truth, which allows the study of the influence of different parameters. However, simulated data is well known to be easily biased. It often does not effectively reproduce all of the complex factors involved in real data. Furthermore, real brain changes may occur in healthy patients (see section 2.1). Simulated lesion evolutions were therefore used to determine how the characteristics of an evolution affect our detection system. In particular, the influence of the size of an evolution was evaluated.

The experimental protocol described in section 4.2, evaluates the detection system in a real-world application. Since determining a strict ground truth is not possible, we opted for an approach where: a first expert manually detects
Lesions were constructed using a Gaussian profile that was added to real brain MR images in all modalities. The intensity of the Gaussian profile was determined by using intensities observed in real-world MS lesions. Several values were used for the standard deviation of the Gaussian $\sigma_r$ ($\sigma_r$ is the radius of the simulated lesion at half of its intensity). Evolutions were modeled as lesions that appear completely between two exams: the lesions were added to images of normal control subjects, and these images were compared to previous images of the normal control subjects taken two months earlier. All of the images went through the complete processing chain. Several experiments were performed, each one involving one or more simulated lesions. Simulated lesions were placed sufficiently far apart from each other so as to avoid interference.

Table 1 shows how the detection probability $p_{\text{det}}$ of simulated lesions depends on $\sigma_r$ (in voxels: 1 voxel = 2mm). Lesions are correctly detected if they are larger than $\sigma_r > 0.6$ voxels. Here, a simulated lesion evolution is considered to be detected if its change detection likelihood $l$ is higher than the likelihood of all false-alarms.

Experiments on the size of the detection window $W_s$ show that it is the primary factor influencing the minimal size of detectable evolutions. The detection window has a smoothing effect: small windows enable the detection of small evolutions but yield noisy results, large windows mask out small evolutions but provide more reliable results. The window size of $3\times3\times3$ voxels was considered to be a good trade-off for MS-lesion evolution.

The system was also tested on large lesion evolutions to test its breakdown point. We found that lesions up to $\sigma_r = 7$ voxels were correctly dealt with.
For larger sizes small spurious detections may appear outside of the evolution, mainly due to the low-pass RF correction filter and to over-compensation during deformable registration. The breakdown point, $\sigma_r = 7$ voxels, corresponds to very large evolutions, of a size comparable to the ventricles, that are well outside of the scope of the type of approach we are considering here (see section 2). We have therefore validated that the deformable registration involved in the realignment step does not affect the evolutions we are considering.

4.2 Experimental Evaluation

4.2.1 Experimental Protocol

The evaluation is based on the detection of “evolving lesions”: we do not address the problem of estimating the volume of detected lesion evolutions (see also section 5). For the expert, the precise definition of a “lesion evolution” is subject to interpretation, and is based on his prior knowledge of lesion configurations. For our automatic detection system, a “lesion evolution” is a cluster of voxels resulting from the statistical change detection step (section 3.9).

We chose to consider detected clusters rather than detected voxels for two reasons. First, the expert interprets the results by labeling lesions as having changed or not, he does not consider individual voxels. It would thus be difficult to compare the expert evaluation with automatic change detection if this was done on a voxel change basis. Second, for the expert, detection of individual voxel changes would mean manual labeling of each detected voxel, which implies a tedious (in practice not feasible) segmentation of evolutions. Manually or semi-automatically delimiting lesions has been shown to be unreliable (important total lesion load inter-rater variability has, for instance, been reported in Mitchell et al. (1996)). Manually delimiting lesion evolutions is even more imprecise.

The evaluation protocol consists in 3 steps, and involves two experts:

(1) Expert-1 marks evolving lesions between pairs of images on the whole database.
(2) The automatic change detection system marks evolving lesions between pairs of images on the whole database.
(3) Expert-2 evaluates the detections of expert-1 and of the automatic system. Expert-2 will decide which detections are valid.

The first step, the labeling by expert-1, is very time-consuming. Expert-1 must consider each pair of images, and find evolutions by visually inspecting each slice. This is the procedure normally used by clinicians. Here, expert-
was assisted by rigid registration, linear intensity normalization, and used multi-planar volume visualization software. Note that expert-1 did not use the preprocessing used by the automated detection system (deformable matching, joint histogram normalization etc.), as these are integral parts of the automated system, and it is our purpose to evaluate them here. The last step, the validation done by expert-2 is much faster, as he only has to evaluate previously designated evolutions.

Matching between evolutions labeled by expert-1 and clusters detected by the automatic system was done manually. In some very rare cases, a single evolution labeled by expert-1 was divided into several detected clusters. In those cases, extra clusters were labeled as false-positives.

After evaluation by expert-2, automatic detected evolutions may be labeled as:

**Valid lesion evolution** confirmed by expert-2.

**Valid non-lesion evolution** Confirmed by expert-2 as being a real evolution, but is not a lesion evolution (example: ventricular expansion).

**False detection** : rejected by expert-2.

Figure 9 shows the graphical interface used for the evaluation. Cycling animations of compared images are overlaid with expert-1’s detections and automatic detections. Expert-2 can efficiently cycle through detections to validate or reject them.

The threshold $\lambda$ of the automatic detection system may be tuned, giving a trade-off between detection sensitivity and error probability. This trade-off is tuned for each pair $(a, b)$ of compared images. The detection probability $p_{\text{det}}(\lambda, (a, b))$ is defined for each pair of images as the ratio between valid detections for threshold $\lambda$ and the total number of valid detections (including expert-1’s detections). Let $n_{\text{err}}(\lambda, (a, b))$ be the number of errors for threshold $\lambda$. The $n_{\text{err}}$ versus $p_{\text{det}}(n_{\text{err}}, (a, b))$ plot is comparable to a ROC curve, and is characteristic of the detector’s performance. The average detection probability $p_{\text{det}}(n_{\text{err}})$ on all image pairs is shown on figure 10. The final result of the automated detection system is a sorted list of detected clusters for each image pair (as described in section 5).

4.2.2 Experimental Results

The processing chain was applied to over 200 source images of different modalities with no user intervention. The robustness and reliability of processing in such an automated system are very important.

A total of 93 validated lesion evolutions, from both expert-1 and the au-
Fig. 9. Graphical interface used for evaluating the performance of the detection system. The three multi-planar views display cycling animations of the images at the two time points that are being compared. On the top left view (RARE) are also shown expert-1’s detections. On the top right view are shown automatic detections. The lower right view is a 3D display that helps keeping track of spatial positioning. This software was written using ImLib3D and Medimax (Bosc et al., 2002)

tomated system, were found in the database. Expert-1 found a total of 35 validated lesion evolutions. Figure 11 gives a description of the sizes of the automatically detected evolutions. The stated volumes are a rough estimate (see sections 4.2.1 and 5) computed by using a common, manually selected threshold for the whole data base. Figure 11 also shows that evolutions correctly detected by expert-1 as well as those missed by expert-1 had size characteristics comparable to other evolutions. Evolutions missed by expert-1 were therefore not particularly smaller or larger than other evolutions. Of the 5%
Fig. 10. Performance of the automated detection system. For comparison, the expert’s detection rate was evaluated at 42%. Multi-modal detection eliminates many false detections and therefore allows us to attain higher detection rates.

Fig. 11. Sizes of automatically detected evolutions. Voxels are 2mm x 2mm x 2mm. (see section 4.2.2).

of evolutions missed by the automatic detection system (e.g. not in the first 30 automated detections), most were very small (total volume of 2 to 4 voxels). These lesions are of a size which is equivalent to the Gaussian profiled simulated lesions with $0.4 < \sigma_r < 0.65$. The remaining evolutions, missed by the automatic detection system had intensities that varied only slightly, and it was not clear for expert-2 whether these were actual evolutions. Evolutions identified by expert-1 but rejected by expert-2 are difficult to characterize, as they have diverse origins (see section 5).

Figure 10 shows the performance of the automatic detection system. It plots
the average detection probability against the number of falsely detected clusters. Expert-1 detected evolutions with a probability of 0.42. This means that more than one out of two lesion evolutions were missed by expert-1. 37% of expert-1’s detected evolutions were considered incorrect by expert-2. The automatic system therefore performed slightly better than expert-1 for an equivalent false detection rate. However, the automatic system provides much higher detection rates of 95% for higher false alarm rates.

5 Discussion

As shown by the experimental results, manual change detection misses many (58%) lesion evolutions. Manually scanning the full volume of two images for changes is indeed very time consuming and experts tend to skip areas that may contain valid evolutions. Furthermore, the first expert is also misled by acquisition artifacts, such as contrast changes, whereas these are corrected in our automated system. For example, if a lesion has a Gaussian profile, a contrast change may make the lesion seem to grow or shrink. Without non-linear intensity normalization, expert-1 might be misled, and label it as a valid evolution.

The system presented here provides a tool that may be used for assisting a neurologist in finding lesion evolutions. The first 30 automatically detected evolutions, sorted in descending order of importance, are presented to the neurologist. He can scan these evolutions and reject incorrect detections. A large majority of residual false detections occur in CSF, specifically in the central parts of the sulci, which are difficult to reliably segment by automated methods. In practice, these are instantly rejected by the neurologist. Therefore, scanning the 30 detected evolutions of an image pair typically takes about one or two minutes, whereas manually comparing two images volumes may take over 20 minutes. It is planned that false detections in CSF regions will be removed by automatic segmentation in future work.

A major problem we faced while designing this change detection system was identifying which artifacts were creating the detection problems. Since human vision, and expert knowledge, naturally compensate for many artifacts, it is very difficult to determine the exact sources of many detection errors. Sub-voxel position differences are particularly hard to observe. Detailed inspection using several visualization tools was necessary. We found that animations of the successively registered images proved to be a very efficient visualization tool (animations are available at http://ipb.u-strasbg.fr/ipb/gitim/research/research_a.html)

The lesions we observed had many different sizes and shapes. The lesion evolu-
tions were not restricted to radially expanding or contracting spherical shapes. Our change detection system makes no assumption on this. We also observed a large amount of ventricular expansion and contraction, as well as changes inside the ventricles.

Change detection is constantly faced with the trade-off between eliminating artifactual changes and the risk of eliminating real, legitimate changes. Care was taken at each step to evaluate its impact on legitimate changes. The multi-scale deformable registration that we use, allows us to chose a maximum scale. At scale $i$ the deformation field is divided into $2^i$ cells in each direction. Choosing $i = 3$ gives us a low resolution field that is slow-varying so that the resulting deformation has very little influence on local features such as lesions (as experimentally verified in section 4.1). Choosing a lower resolution would provide less correction of undesired deformations (section 2.1). Choosing a higher resolution will reduce the maximal allowable size of evolutions before distortion by deformable registration appears. The intensity normalization is done on an image-wide basis. Therefore local image intensity changes have no measurable impact on the image-wide histograms. The low-pass filter that fixes the RF inhomogeneities is very large, and does not affect local image changes. The robustness of these decisions was evaluated using simulated lesions (section 4.1).

The simultaneous use of images of several modalities at each time-point greatly increased detection performance. The detection probability increases from 79% to 95% when using multi-modal data (figure 10). Indeed, multi-modality provides redundant information while noise and artifacts are uncorrelated between modalities. The multi-slice RARE images have residual artifacts that are distributed in a very different manner than the real 3D GE3D images. This redundant information eliminates many false alarms in multi-modal detection and thus increases the detection rate for a given false-alarm rate (Figure 10). Even in the case of manual change labeling, the multi-modal information is very useful for confirming or rejecting a change.

The system described here is oriented toward determining if and where there have been subtle changes. Our experience, as well as that reported by other authors (see section 4.2.1) is that quantifying *evolutions* is very inaccurate. Noise and residual artifacts make it difficult, even for a trained expert, to validate small evolutions. Quantifying the amount variation is even more difficult. Furthermore, the exact intensity profile of a lesion may change in complex ways, making it difficult to give a satisfactory, reliable, definition of what is understood by the *volume* of an evolution. Nevertheless, if one accepts these inaccuracies, our method may be very simply used for quantitative analysis (see Figure 11 and section 4.2.2). However, this might not, in certain cases, match with what a neurologist subjectively considers as the volume of the variation.
Another, more subjective, semi-automatic approach has been considered within our change detection framework. The change detection system presented in this paper presents the neurologist with a list of evolutions sorted in decreasing likelihood. For each validated evolution, the neurologist may vary a threshold on the likelihood level, and thus semi-automatically segment the evolution. Fully automatic segmentation techniques have also been considered (see for instance (Rey et al., 2002)).

The system we have described here is a general change detection system. It was applied to MS-lesion evolution, but no prior knowledge on size, shape, intensities, or locations of such evolutions was made. An important path to reducing false detection rates for a specific application would be to introduce such prior knowledge. Positional information such as knowing whether an evolution is in GM, WM or CSF regions would eliminate many false alarms. Information on intensity and size would help rejecting other non lesion evolutions.

Another path for improving performance would be to add a more sophisticated model for noise. Spatially variant noise maps, or noise distributions learned using PCA are possible directions.

6 Conclusions

We have presented a comprehensive framework for automatic change detection in serial MRI. The original contributions of this paper are the use of the following techniques in the specific case of change detection: the use of multi-modal information for change detection; an original non-linear joint histogram normalization technique; the use of a specific probabilistic detection scheme. The use of multi-resolution deformable image matching in this context is also new, as it has not been previously used as a realignment step for intensity comparison. We propose a systematical analysis of each major artifact and describe a reliable method for correcting it. The reliability was confirmed by the automatic processing of a large database of images of different types.

This automatic change detection has important applications for MS followup, and provides an objective marker of disease evolution.

References

Bosc, M., Vik, T., Armspach, J.-P., 2002. ImLib3D,Medimax: 3D image pro-


Prentice Hall.
Kikinis, R., Guttmann, C., Metcalf, D., Wells, W., Ettinger, G., Weiner, H.,
Jolesz, F., 1999. Quantitative follow-up of patients with multiple sclerosis
fluid volumes with MR imaging. I. methods, reliability and validation. Radio-
diogy 178, 115–122.
Lehmann, T., Gönner, C., Spitzer, K., 1999. Survey: interpolation methods in
medical image processing. IEEE Transactions on Medical Imaging 18 (11),
1049–1075.
Lemieux, L., Wieshmann, U., Moran, N., Fish, D., 1998. The detection and
significance of subtle changes in mixed-signal brain lesions by serial MRI
scan matching and spatial normalization. Medical Image Analysis 2 (3),
227–242.
Lim, K., Pfefferbaum, A., 1989. Segmentation of MR brain images into cere-
brospinal fluid spaces, white and gray matter. Journal of Computed Assisted
Tomography 13, 558–593.
The variability of manual and computer assisted quantification of multiple
Mitiche, A., Bouthemy, P., 1996. Computation and analysis of image motion:
A synopsis of current problems and methods. International Journal of Com-
image matching using constrained hierarchical parametric models. IEEE
Transactions on Image Processing 10 (7), 1081–1093.
images over multiscale nested subspaces. application to atlas-based MRI
segmentation., accepted submission to Pattern Recognition.
T1 hypointense lesion load in secondary progressive multiple sclerosis: a
comparison of pre versus post contrast loads and of manual versus semi
automated threshold techniques for lesion segmentation. Multiple Sclerosis
4, 408–412.
Rey, D., Subsol, G., Delingette, H., Ayache, N., 2002. Automatic Detection
and Segmentation of Evolving Processes in 3D Medical Images: Application
to Multiple Sclerosis. Medical Image Analysis 6 (2), 163–179.
of brain change over time. Tech. Rep. TR00SMS1, FMRIB John Radcliffe
Hospital, Headington, Oxford OX3 9DU, UK.
Transactions on Medical Imaging 19 (7), 739–758.