Tina Memo No. 2008-002
Internal Memo

Problems with the Brainweb MRI Simulator in the Evaluation of Medical Image Segmentation Algorithms, and an Alternative Methodology

P. A. Bromiley

Last updated
20 / 12 / 2007

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Problems with the Brainweb MRI Simulator in the Evaluation of Medical Image Segmentation Algorithms, and an Alternative Methodology

P. A. Bromiley
Dept. of Medical Biophysics
Imaging Science and Biomedical Engineering Division
Medical School, University of Manchester
Manchester, M13 9PT, UK
paul.bromiley@man.ac.uk

Abstract

We demonstrate that simulated MR images obtained from Brainweb do not model the partial volume effect in a realistic fashion, and therefore cannot be used to evaluate medical image segmentation algorithms that rely on models of intensity distributions and incorporate partial volume effects. However, we make two observations; first, evaluation of segmentation algorithms on simulated data can only prove consistency between the assumptions incorporated into the simulation and segmentation algorithms, and second, given this constraint, a method for producing approximately noise-free MR images is all that is required to draw any conclusions that could be drawn through the use of Brainweb simulated images. We use these observations to motivate an alternative method for evaluating medical image segmentation algorithms, based on the use of the multi-dimensional segmentation algorithm provided by TINA. This method uses segmentations of multi-dimensional data to reconstruct noise-free MR images; these can then be used in Monte-Carlo experiments to measure the parameter stability of the segmentation, and also to assess the presence of most forms of potential bias.

1 Introduction

Simulated MR image volumes play an important role in the development of some types medical image analysis algorithms, notably segmentation algorithms, as they allow evaluation of absolute segmentation accuracy by comparison with the phantoms or atlases used to produce them. Performing the corresponding experiment with clinical[1] data would require the production of gold-standard segmentation results by an expert observer, which would be both time consuming and expensive. However, if the evaluation is to be meaningful, the simulation must model all of the major effects seen in real MR images in a realistic fashion. These include, in order of priority:

- spatial locations of the relevant tissues;
- tissue intensities, calculated through application of the Bloch equations to the T1 time, T2 time and proton density (PD) of the tissues and the parameters of the scan (e.g. field strength, repetition time, echo time etc.);
- the partial volume effect at tissue boundaries;
- additive image noise;
- the point spread function of the MR scanner
- inhomogeneity effect.

These requirements must be interpreted in the context of the known statistical properties of MR images and capabilities of available pre-processing algorithms: the simulation need not incorporate features that can be dealt with effectively by existing algorithms, and approximations may be used where their effects are known to be minimal. For example, efficient skull-stripping algorithms are available, and the intra-cranial volume consists predominantly of grey matter (GM), white matter (WM), and cerebro-spinal fluid (CSF); therefore, simulations incorporating

[1] We adopt the phrase "clinical" MR to denoted MR image volumes acquired from subjects, as opposed to simulated images.

only these three tissues may still be useful. Effective inhomogeneity correction algorithms are available (see [7] and references therein), and so simulations with 0% inhomogeneity effect can be useful. The noise in MR magnitude images is known to be Rician [6], the result of addition in quadrature of the Gaussian noise on the real and imaginary images; however, the Gaussian and Rician distributions are identical to a good approximation at if the signal-to-noise ratio is greater than 3, and so Gaussian noise may be used in the simulation in cases where only the brain tissues are considered. These observations limit the required features of the simulation considerably; it need only incorporate realistic tissue locations (in order to allow the use of spatial information in the segmentation), incorporate realistic modelling of the tissue intensities, together with additive Gaussian noise (in order to allow the use of intensity information in the segmentation) and incorporate the partial volume effect in order to provide a realistic challenge for segmentation.

Including the partial volume effect in the above list may prove contentious since many popular medical image segmentation algorithms, e.g. those provided by the SPM [1] and FSL [17] software packages, do not incorporate it. This may be valid in certain situations. Partial voluming is the result of the combination of several tissue within a single voxel, when both tissue contribute to the resulting intensity. Therefore, it is largely relevant only to MR images (and not, for example, to CT where the intensity of the voxels is dictated primarily by a single tissue), and becomes increasingly significant with increasing voxel dimensions; in thick slice acquisitions it may account for up to 30% of the voxels [13]. It is only observed as a separate contribution to the intensity histogram when the means of the two corresponding pure tissues are widely separated compared to their standard deviations. Under these circumstances, the partial volume voxels will be observed as an approximately uniform distribution between the peaks produced by the two pure tissues. The uniform nature of the distributions is the result of two effects. First, in general the locations of tissue boundaries is not correlated with the locations of voxel boundaries. Therefore, all possible fractional combinations of tissue are equally probable i.e. the distribution of fractional combinations is uniform. Second, the Bloch equations that govern the image formation process in MR are linear, and so the intensity of a voxel containing a mixture of pure tissues is given by a linear combination of the intensities of the pure tissues it contains, weighted by their fractional contributions. Therefore, all possible partial volume intensities are equally probable i.e. their distribution is uniform [2].

The importance of partial voluming in MR image segmentation under the circumstances outlined above has led to the development of segmentation algorithms that incorporate it. Initial developments focused on models of the intensity histogram incorporating uniform distributions for partial volume voxels [15, 16, 11, 12], although non-uniform distributions have also been investigated [3, 8]. The TINA software also contains such an algorithm, which is also capable of analysing multi-dimensional data (i.e. where several MR images of the same anatomical region have been acquired using different pulse sequences) and incorporating intensity gradient into the intensity histogram (to aid in the disambiguation of pure tissue and partial volume contributions, since the latter occur at tissue boundaries i.e. locations of high intensity gradient); the development of the algorithm is described in [14, 13, 20, 2, 18]. The development of such algorithms requires validation, and if this is to be performed by measuring the absolute segmentation accuracy on simulated images, through comparison to the phantoms or atlases used to produce them, then the simulations must incorporate realistic models of partial volume effects.

The Brainweb MRI simulation package [4, 10, 9, 5] has become a de-facto standard in medical image analysis validation over recent years, due in part to its availability over the Internet (www.bic.mni.mcgill.ca/brainweb/). However, if the considerable body of literature based on Brainweb simulated images is to have any relevance, we must be sure that the simulations meet the requirements listed above. In this report, we demonstrate that the modelling of partial volume effects in Brainweb simulated images is not performed in a realistic fashion, and therefore segmentation algorithms that incorporate partial volume effects, and so would have increased accuracy on real MR data where such effects are significant, may show inconsistent or even reduced performance when evaluated on Brainweb images. This observation implies a clear requirement for an alternative method for evaluating segmentation algorithms when the partial volume effect is to be considered.

In order to suggest such an alternative, we must first clearly identify the evaluation methodology we wish to adopt. The general aim is to perform a segmentation on real MR data to obtain a set of tissue phantoms describing the locations of each pure tissue, use these to produce simulated images, add noise, segment the simulated images to produce maps of the tissue locations, and compare these maps to the tissue phantoms in order to measure the absolute accuracy of the segmentation. This can be incorporated into a Monte-Carlo experiment. Since the

[2] Note that this simple model is an approximation: e.g. in 2D where a tissue boundary passes through a voxel, it intersects with two of the voxel boundaries, and it is the position of both of these intersections that have uniform distributions. The area of each pure tissue is therefore dictated by the product of two uniform distributions i.e. an exponential distribution, and so we would expect the partial volume distribution to peak at either end, close to the pure tissues. However, since the intensity model used in the segmentation will also incorporate pure tissue distributions at these locations, these distributions will take up the excess contribution above uniformity from the partial volume voxels. The effect on accuracy will be minimal, as the voxels concerned consist predominantly of one pure tissue. The error introduced in this way will therefore be dictated by the ratio of the frequencies of occurrence of the two pure tissues, weighted by the ratio of frequencies of occurrence of pure tissue to partial volume voxels

segmentation inevitably involves fitting a model of some form to the data, we can interpret it as a maximum-likelihood process, and this makes clear the significant implicit limitations of such an evaluation. We can only evaluate two aspects of the system as a whole: the effects of image noise on stability, and the goodness-of-fit of the model used in segmentation to that used in simulation. The former can be evaluated through the standard deviation of the tissue volumes, overlap, or fitted parameters i.e. random errors: the latter through consistent under- or over-estimation of tissue volumes or fitted parameters i.e. systematic errors or bias. We can therefore draw one of only two conclusions. If there is no evidence of bias, we can conclude that the models used in segmentation and simulation are consistent and, to the (unknown) extent that these models are also consistent with the image formation process in real MR images, the random errors describe the dependence of segmentation accuracy on image noise, and can be used in comparative evaluations ("shoot-outs") of multiple segmentation algorithms. If there is evidence of bias, we can only conclude that the simulation and segmentation algorithms are inconsistent. In this circumstance, at least one of the models must also be inconsistent with the image formation process in real MR images, but we will not know which. In addition, the well-known properties of maximum-likelihood (e.g. asymptotic efficiency) are only guaranteed if the model is a good fit to the data; in this case, if the model used in the segmentation is a good fit to that used in the simulation. If this is not the case, as will be demonstrated by the presence of bias, then the results of maximum-likelihood are unknown, and so the random errors on the evaluation results tell us nothing about the performance on real data. We can even assume that when a researcher develops a new segmentation algorithm based on fitting a model of image intensities to the intensity histogram of the data, they will use the most accurate model available at the time, and furthermore that the segmentation algorithm will be developed subsequently to the simulation algorithm used to evaluate it. Any bias in the evaluation results is therefore more likely to indicate failures in the simulation package than failures in the segmentation algorithm.

The limitations in the standard approach to evaluation of segmentation algorithms using simulated data can suggest an alternative approach that is at least as powerful. We need to evaluate three features of any new segmentation algorithm:

- Is there any bias (i.e. consistent over- or under-estimation of the tissue volumes) on the segmented tissue maps or, equivalently, the fitted parameters? If the answer is yes, we can ignore the next two aspects, since the algorithm is fundamentally unusable in its current form, and requires more development work.
- What are the random errors on the segmented tissue maps? This information will feed through to any subsequent analysis.
- Are the spatial locations of the segmented voxels correct (i.e. we could in theory develop an algorithm that identified a volume of white matter, equal to the volume of CSF, as CSF: the measured volume might appear to have low random errors and no bias, but the result would be fundamentally wrong).

The primary perceived advantage of the standard approach to segmentation evaluation using synthetic data, namely segmenting real MR data to produce a set of tissue phantoms, using these to generate noise-free simulated MR images, adding noise (perhaps in a Monte-Carlo process), segmenting the simulated images with the algorithm we wish to test, and measuring the overlap of the resulting tissue maps with the phantoms, is that it provides a direct measurement of the third of these aspects. However, this is clearly not the case in practice: it only measures the degree to which the assumptions used to generate the algorithm under test match those used to generate the phantoms, and in comparison tells us little about the performance on real data. With this primary advantage of the existence of the phantoms removed, we need not have them at all, except to generate the simulated images.

2 Brainweb Simulated Images

We have identified several problems with the use of Brainweb simulated images in performance evaluation of medical image segmentation algorithms:

- confusion of partial voluming with image noise in the generation of the phantoms;
- generation of histogram artefacts in the simulation of the images;
- non-stationarity of the simulator;
- lack of relevance to performance on genuine MR data.

The following sections demonstrate each of these issues in turn.

2.1 The Construction of the Brainweb Simulator

The Brainweb simulator is described in [4], [10], [9], and [5]. It is based on a set of tissue phantoms generated from 27 high-resolution ($1mm^3$ isotropic voxels), low-noise, T1-weighted gradient-echo acquisitions of the same individual, which were registered into a common stereotaxic space where they were sub-sampled and intensity averaged [5]. Intensity non-uniformity was reduced by deconvolution with the non-uniformity blurring kernel derived from the intensity histogram of the image. The image was then segmented: a trained neuroanatomist identified points within pure tissues and a fuzzy minimum distance classifier was applied to identify the tissue contents of each voxel (see below). Manual editing and masking were then applied to produce the final set of tissue phantoms.

The simulation process includes up to four steps. NMR parameters for each tissue and the Bloch equations are used to generate a zero-dimensional simulation describing the intensity of each pure tissue. The tissue intensities are weighted by the tissue proportions in each voxel in the phantoms to produce synthetic images. The effects of sampling in the Fourier domain are addressed, followed by addition of Gaussian noise fields to the real and imaginary components of the images and magnitude image reconstruction. Finally, intensity inhomogeneity fields are applied to generate the final simulated image volumes.

2.2 Confusion of Partial Voluming with Noise

A significant error is contained within the minimum distance classifier used to generate the tissue phantoms used in the Brainweb classifier. According to [5], in order to deal with partial voluming, each voxel was assigned an n-dimensional tissue membership vector, where n was the number of tissue classes considered. Each component of the vector f_i represented the proportion of the ith tissue within the voxel, and the vectors were normalised such that $0 < f_i < 1$ and $\sum_{j=1}^{n} f_j = 1$. For voxels with intensity equal to the mean intensity m_i of any tissue class i, the ith tissue fraction f_i was set to 1 and all other components to 0. For all other voxels, the components of the vector were estimated to be inversely proportional to the distance between the voxels intensity g_i and the class mean m_i

$$f_i = \frac{1}{|g_i - m_i|}$$

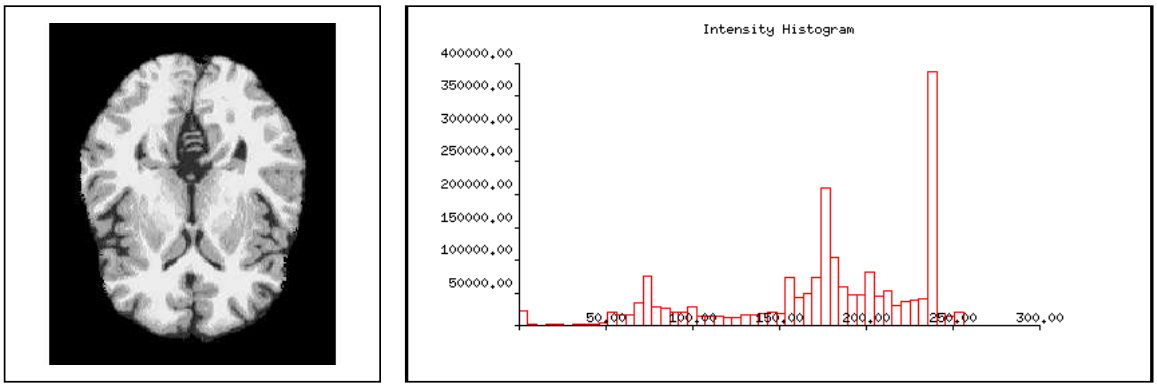
again normalised such that the components of the vector summed to unity. Manual intervention and prior anatomical knowledge were used to remove partial volume contributions that were known not to exist.

The problem with this approach is that it confuses partial voluming with image noise. Real MR images contain two types of intensity distribution, pure tissue distributions (which we may assume to be delta functions if all non-uniformity has been removed) and partial volume distributions, which take the form of extended distributions between the pure tissue distributions for any pairs of tissues that share a common boundary. When noise is added during the image acquisition process, both types of distribution are convolved with a noise distribution. Therefore, movement of the voxel intensity away from the mean intensity of a pure tissue class may be due either to the effects of convolution with the noise distribution or to partial voluming with another tissue. The classification process used in the Brainweb simulation ignores the first effect, effectively assuming that pure tissue classes are not affected by noise. This has two effects. First, the number of partial volume voxels is considerably overestimated. Second, the shape of the partial volume distributions is changed compared to that in the underlying images: the excess partial volume voxels are more likely to occur close to the pure tissue intensities, and so the partial volume distributions are enhanced at these locations. The result is that the phantoms used in the simulations have non-realistic partial volume distributions.

2.3 Histogram Artefacts in the Simulated Images

The Brainweb T1-weighted simulation with 1mm thick slices, zero noise and zero inhomogeneity was downloaded as gzip compressed raw short data, uncompressed using gunzip version 1.3.12, and read into TINA using the raw image reader. The fuzzy tissue phantoms were also obtained and loaded into TINA using the same procedure, then scaled to lie between 0 and 1 such that voxel intensity represented the volumetric contribution of the tissue to each voxel. The T1 simulated image volume was then multiplied by the summed phantoms for grey matter, white matter, CSF and glial matter to segment these tissues. Figure 1 shows an example slice (number 142) from the resulting image volume together with the intensity histogram of the volume.

Figure 1 should contain a maximum of five peaks in intensity (background, CSF, GM, WM and glial matter). The peaks for the background, CSF, GM and WM are clearly visible: there is too little glial mater to produce a distinct peak. However, the histogram also contains secondary peaks either side of the main peaks for the pure



(a) (b)

Figure 1: The CSF, GM, WM and glial matter from slice 142 of the Brainweb T1 simulated image with 0% noise and 0% inhomogeneity (a), and the intensity histogram of the whole volume (b).

tissues: these are particularly visible either side of the GM peak. Figure 2 shows the phantoms for the brain tissues, together with the intensity histograms for the whole phantom volumes, smoothed with one iteration of tangential smoothing in order to reveal the structure more clearly. Secondary peaks are not apparent in these histograms, implying that they are generated at some point in the reconstruction (this also eliminates the possibility that the secondary peaks were generated by a bug in the TINA raw image reader or histogramming routine).

The spatial locations of the voxels in the secondary peaks were crudely identified by thresholding the image. After masking out the non-brain tissues, the image was scaled lie between 0 and 255 grey levels, such that the CSF, GM and WM peaks occurred at approximately 75, 180 and 240 grey levels respectively. Figure 3 shows the results of thresholding between 90 and 105 grey levels, 145 and 165 grey levels, and 190 and 210 grey levels, to locate the peaks on the high intensity side of the CSF, the low intensity side of the GM, and the high intensity side of the GM respectively. This process reveals that the secondary peaks correspond to the locations of the tissue boundaries i.e. they are partial volume voxels.

The image simulation process used by Brainweb, as described in the literature [4, 10, 9, 5] involves only multiplication of the phantoms with mean tissue intensities derived from a simulation of the Bloch equations, resampling and noise addition, and production of a magnitude image from the Fourier transform of the complex data. Of these steps, only the resampling could preferentially alter the intensities of some of the edge voxels in order to produce the secondary peaks observed in the intensity histogram (i.e. side-lobes of the sinc function, generated by the truncation artefact). However, the lack of the tertiary lobes of the sinc function in the histogram, the lack of visible ringing in the simulated images, the spatial location of the voxels in the secondary peaks only at the tissue boundaries (rather than on either side of them) and the height of the secondary peaks all cast doubt on this interpretation.

As described in the introduction, various models for the intensity distributions of partial volume voxels have been suggested. Earlier work focused on uniform distributions [15, 16, 11, 12], based on the simple logical arguments given above. Some researchers have investigated non-uniform models [3, 8], in the case of Joshi and Brady derived through extensive simulation work. However, these non-uniform models always produce intensity distributions that peak at either end, close to the pure tissue distributions: there is no support in the literature for partial volume intensity distributions that feature peaks in the middle, away from the pure tissue distributions. In addition, the minor peaks appear to be non-symmetric, biased towards the nearest pure tissue. Figure 4 shows a sample slice from an IRTSE MR image of a young normal subject and, whilst an approximately uniform distribution is present as a long tail between the locations of the peaks for CSF (-1800 grey levels) and GM (-600 grey levels) there is no evidence for peaks contained within this partial volume distribution. We must therefore conclude that the Brainweb simulated data does not incorporate realistic modelling of partial volume effects, and therefore that its use in evaluating segmentation algorithms that incorporate modelling of these effects is likely to lead to meaningless results, following the arguments outlined in the introduction.

2.4 Other Issues

One of the main, perceived advantages of an MRI simulator such as Brainweb is that it provides a standard evaluation problem for medical image segmentation algorithms. Independently derived results from various groups and

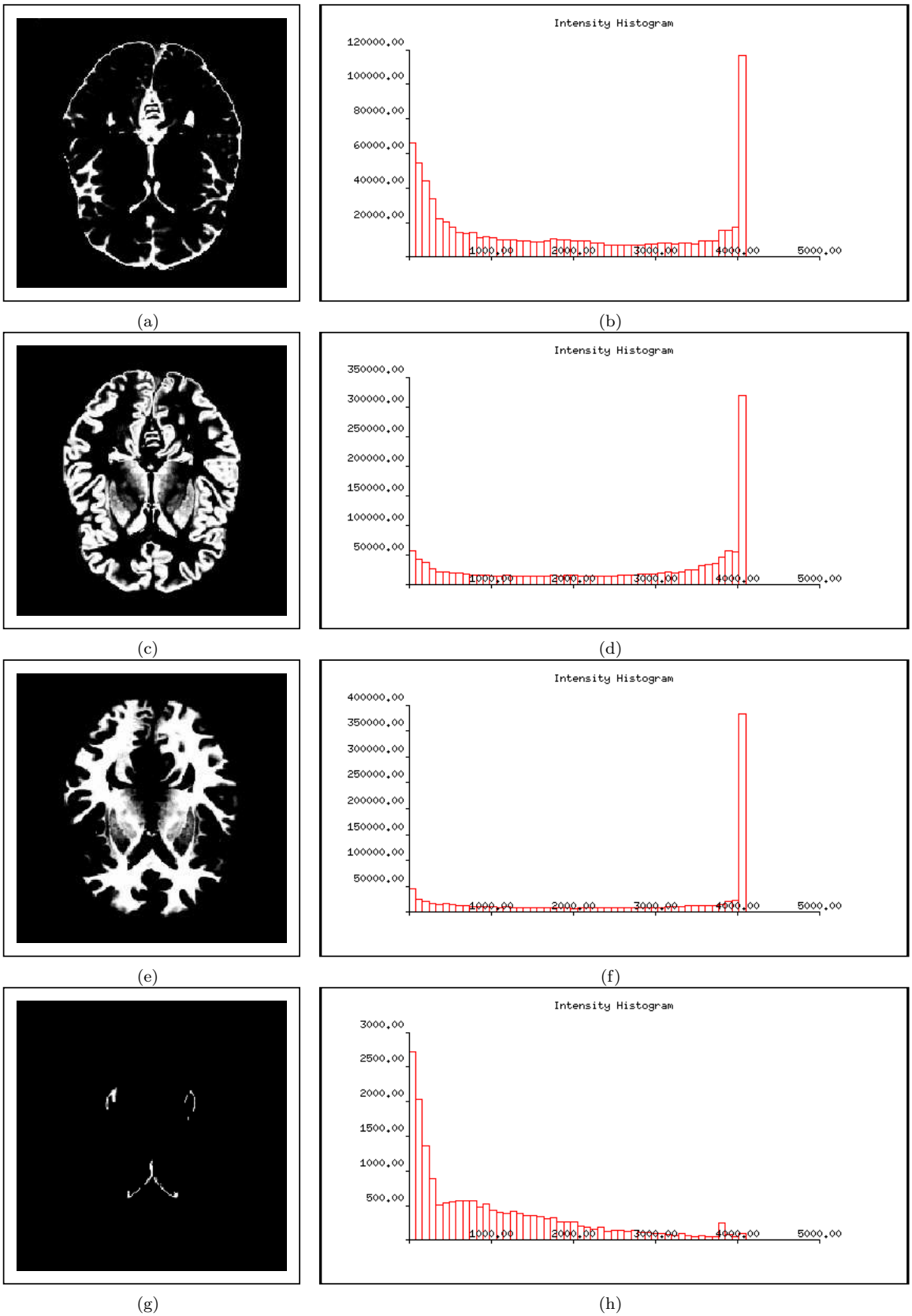


Figure 2: Slice 142 of the Brainweb brain tissue phantoms, and the intensity histograms of the whole volumes: CSF (a,b), GM (c,d), WM (e,f) and glial matter (g,h)

algorithms can in theory be compared directly. However, the Brainweb web-site (www.bic.mni.mcgill.ca/brainweb) states that "the SBD (Simulated Brain Database) is still considered 'under development' both in terms of the

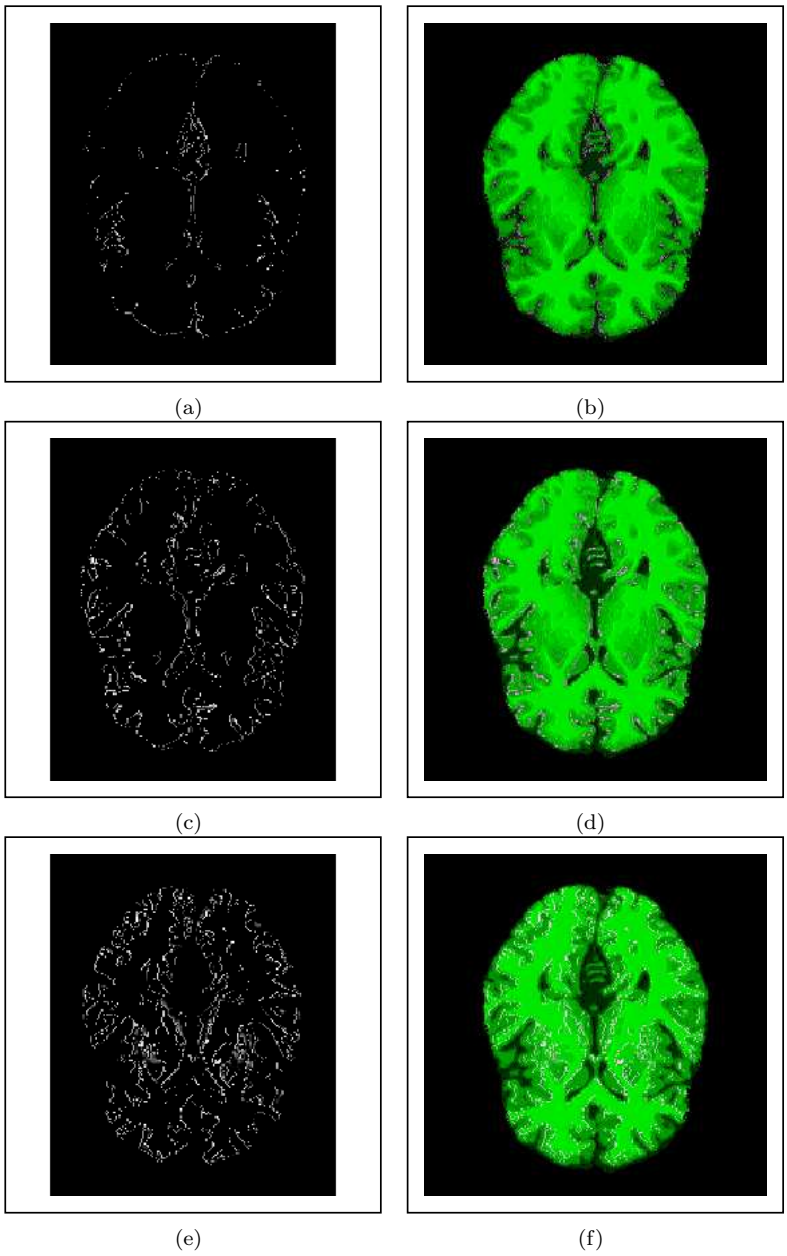
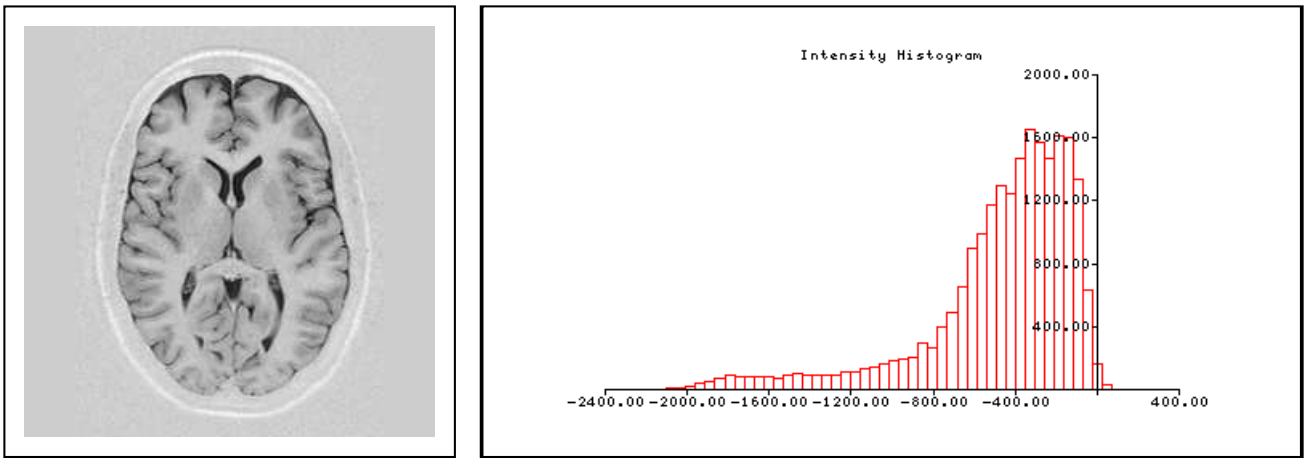


Figure 3: The spatial locations of the voxels in the secondary peaks on the high intensity side of the CSF (a,b), the low intensity side of the GM (c,d) and the high intensity side of the GM (e.f). The left column shows the identified voxels: the right column shows them overlaid on the original image.

anatomical model and the simulation itself. What you get today may not be the same as what you get tomorrow!". In addition, there is no system of version numbering or list of updates to the simulator. The perception that Brainweb provides a standard problem is therefore false: evaluation results produced using the simulated images cannot be directly compared without first confirming that exactly the same set of simulated images were used.

The problems introduced by this lack of stationarity cannot be assumed to be negligible. For example, the "normal brain database" simulates images with ten tissue classes: background, CSF, grey matter, white matter, fat, muscle/skin, skin, skull, glial matter and connective tissue. Twenty anatomical models derived from twenty normal brains are also available on the web-site, but use a different set of tissue classes: background, CSF, grey matter, white matter, fat, muscle, muscle/skin, skull, blood vessels, connective tissue, dura matter and bone marrow. The addition and removal of entire tissue classes demonstrates that significant changes may take place in the procedures used by the Brainweb group.

The glial matter class has been removed and additional classes for blood vessels, dura matter and bone marrow added. The glial matter class is of particular importance. In the Brainweb simulations, this class is identified as a



(a) (b)

Figure 4: A sample slice from an IRTSE MR image of a normal subject (a), and its intensity histogram (b).

thin surface along the inside of the ventricles. It therefore accounts for a significant amount of the partial voluming present within the images: either partial voluming of glial matter with both CSF and white matter, where the class is included, or partial voluming of white matter with CSF, if it is removed. Most segmentation algorithms, applied to clinical MR images, would not include this class and so would treat this tissue as white matter and white matter partial voluming with CSF. Many users of the Brainweb simulations also ignore the glial matter and focus only on the grey matter, white matter and CSF, using the phantoms to identify these tissues (e.g. [19]). However, this procedure introduces a significant difference between evaluation on Brainweb images and application to clinical MR: a significant amount of partial voluming between CSF and white matter has been removed from the former in comparison to the latter. Since partial voluming is one of the most significant aspects of the MR image formation process as regards segmentation, this casts doubt on the relevance of evaluations performed on Brainweb simulated images to the performance of segmentation algorithms on clinical MR.

Finally, it should be noted that the entire concept of testing segmentation algorithms on simulated images, which themselves were derived from phantoms generated from segmentations of clinical MR images, is deeply flawed. The segmentation algorithm being evaluated is attempting to re-generate the phantoms from the simulated images, for comparison to the original phantoms. Such evaluations therefore address two issues: sensitivity to the image noise added during the simulation process, and the agreement between the assumptions used in the segmentation algorithm being evaluated and those used in the segmentation algorithm used to generate the original phantoms. This may be particularly important if the segmentation algorithm being evaluated assumes an intensity model since, as demonstrated above, the intensity model assumed by the minimum-distance classifier used in the generation of the Brainweb phantoms confuses partial voluming with image noise, leading to unrealistic partial volume distributions in the phantoms. It is therefore possible that a segmentation algorithm that assumes a partial volume distribution that is realistic for clinical MR images will be penalised when evaluated on Brainweb images in comparison to one that assumes a partial volume distribution unrealistic for clinical MR, but which corresponds closely to that contained within the Brainweb phantoms.

3 An Alternative Methodology

The problems with the use of Brainweb simulated images for performance evaluation of medical image segmentation algorithms identified above produce a requirement for an alternative methodology. Such a methodology must retain the primary advantage of Brainweb, namely removal of any need to produce gold-standard data, and avoid the problems associated with the production of simulated images.

The TINA MR segmentation algorithm [14, 13, 20, 2, 18] fits an intensity model to the histogram of an MR image volume using EM optimisation, and produces estimates of the volumes of each tissue within each voxel. Therefore, the segmentation process can be inverted by multiplying the tissue volume fractions by the mean tissue intensities in the fitted model: the result is a noise-free version of the model that was fitted to the original images. Since this is available, we can treat it like any other model and evaluate it through two experiments: first, Monte-Carlo testing of parameter stability (in order to measure the effect of image noise on random errors: this will produce a result equivalent to applying error propagation to the algorithm) and second testing the goodness-of-fit of the model to

the data. The second can be accomplished by measuring the number of voxel values in the reconstructed model that vary from the voxel values in the original image by more than three standard deviations of the image noise. Since the number of voxels in this group can be calculated for a perfect model fit (by integrating the noise distribution from the three standard deviation point to infinity), any excess is a direct measure of the number of voxels for which the model does not fit the data and thus a direct measure of the systematic error on the segmentation. Therefore, through the application of this methodology both the systematic and random errors produced by the segmentation algorithm are evaluated on clinical data, without recourse to gold-standard segmentations. Tina Memos 2003-007 and 2005-013 demonstrate the application of this approach.

4 Conclusions

Brainweb simulated data exhibits anomalous peaks in the intensity histogram, the spatial locations of which correspond to tissue boundaries, implying that they represent partial volume voxels. However, whilst there is disagreement in the literature over whether the intensity distribution of partial volume pixels is uniform, or approximately uniform in the middle with peaks at either end, there is no support for partial volume distributions that peak in the middle. Real MR data also shows no minor peaks in the middle of the partial volume intensity distributions. We must therefore conclude that Brainweb simulated MR images incorporate an unrealistic representation of the partial volume process, and that any intensity-based segmentation algorithm that incorporates modelling of the partial volume effect will give anomalous results when evaluated on Brainweb data. We expect these anomalies to manifest themselves as biases on fitted parameters, possibly dependent on the level of noise added to the images.

Since Brainweb data cannot be used to evaluate intensity based segmentation algorithms that include modelling of the partial volume effect, we have identified an alternative evaluation methodology. This involves a direct test of the goodness-of-fit of the model used in the segmentation process to the original image data, and therfore has advantages in terms of statistical interpretation. It retains the primary advantage of the use of Brainweb simulated data, namely the removal of any need to generate gold standard segmentations, whilst also removing the need for an image simulation. Since the evaluation is performed directly on clinical MR, it provides a realistic estimate of the performance that will be obtained in clinical usage.

References

[1] J Ashburner and K Friston. Multimodal image coregistration and partitioning-a unified framework. *NeuroImage*, 6:209–217, 1997.

[2] P A Bromiley, N A Thacker, M L J Scott, M Pokrić, A J Lacey, and T F Cootes. Bayesian and non-Bayesian probabilistic models for medical image analysis. *Image and Vision Comput.*, 21(10):851–864, 2003.

[3] J P Chiverton. *Probabilistic Partial Volume Modelling of Biomedical Tomographic Image Data*. PhD thesis, Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey GU2 7HX, U.K., Aug 2006.

[4] C A Cocosco, V Kollokian, R K-S Kwan, and A C Evans. Brainweb: Online interface to a 3D MRI simulated brain database. *NeuroImage*, 5(4):S425, 1997.

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

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

[7] Z Hou. A review on MR image intensity inhomogeneity correction. *International Journal of Biomedical Imaging*, 2006:1–11, 2006.

[8] N Joshi and J M Brady. A non-parametric model for partial volume segmentation of MR images. In *Proceedings BMVC'05*, pages 919–928, 2005.

[9] R K-S Kwan, A C Evans, and G B Pike. An extensible MRI simulator for post-processing evaluation. *Visualization in Biomedical Computing (VBC'96). Lecture Notes in Computer Science*, 1131:135–140, 1996.

[10] R K-S Kwan, A C Evans, and G B Pike. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging*, 18(11):1085–1097, 1999.

[11] D H Laidlaw, K W Fleischer, and A H Barr. Partial-volume Bayesian classification of material mixtures in MR volume data using voxel histograms. *IEEE Trans. Med. Imag.*, 17(1):74–86, 1998.

[12] K Van Leemput, F Mayes, D Vandermuelen, and P Suetens. A unifying framework for partial volume segmentation of brain MR images. *IEEE Trans. Med. Imag.*, 22(1):105–119, 2003.

[13] M Pokrić, N A Thacker, and A Jackson. The importance of partial voluming in multi-dimensional medical image segmentation. In *Proc. MICCAI*, pages 1293–1294, 2001.

[14] M Pokrić, N A Thacker, M L J Scott, and A Jackson. Multi-dimensional medical image segmentation with partial voluming. In *Proc. MIUA*, pages 77–80, 2001.

[15] P Santago and H D Gage. Quantification of MR brain images by mixture density and partial volume modelling. *IEEE Trans. Med. Imag.*, 12:566–574, 1993.

[16] P Santago and H D Gage. Statistical models of partial volume effect. *IEEE Trans. Med. Imag.*, 4:1531–1540, 1995.

[17] S M Smith, M Jenkinson, M W Woolrich, C F Beckmann, T E J Behrens, H Johansen-Berg, P R Bannister, M De Luca, I Drobnjak, D E Flitney, R Niazy, J Saunders, J Vickers, Y Zhang, N De Stefano, J M Brady, and P M Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):208–219, 2004.

[18] N A Thacker, M Pokrić, and D C Williamson. Noise filtering and testing illustrated using a multi-dimensional partial volume model of MR data. In *Proc. BMVC*, pages 909–919, Kingston, London, 2004.

[19] J Tohka, E Krestyannikov, I D Dinov, A MacKenzie Graham, D W Shattuck, U Ruotsaainen, and A W Toga. Genetic algorithms for finite mixture model based voxel classification in neuroimaging. *IEEE Trans Med Imag*, 26:696–711, 2007.

[20] D C Williamson, N A Thacker, S R Williams, and M Pokrić. Partial volume tissue segmentation using grey-level gradient. In *Proc. MIUA*, pages 17–20, 2002.