Medical Image Analysis 15 (2011) 267–282

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Evaluating intensity normalization on MRIs of human brain with multiple sclerosis

Mohak Shah [a,c,*], Yiming Xiao [a], Nagesh Subbanna [a], Simon Francis [b,c], Douglas L. Arnold [b,c], D. Louis Collins [b], Tal Arbel [a]

[a] Centre for Intelligent Machines, McGill University, Montreal, Canada
[b] Montreal Neurological Institute, McGill University, Montreal, Canada
[c] NeuroRx Research, Montreal, Canada

ARTICLE INFO

Article history:
Received 13 November 2009
Received in revised form 3 December 2010
Accepted 13 December 2010
Available online 25 December 2010

Keywords:
Intensity normalization
Heterogenous MRI data
Evaluation
Multiple sclerosis

ABSTRACT

Intensity normalization is an important pre-processing step in the study and analysis of Magnetic Resonance Images (MRI) of human brains. As most parametric supervised automatic image segmentation and classification methods base their assumptions regarding the intensity distributions on a standardized intensity range, intensity normalization takes on a very significant role. One of the fast and accurate approaches proposed for intensity normalization is that of Nyul and colleagues. In this work, we present, for the first time, an extensive validation of this approach in real clinical domain where even after intensity inhomogeneity correction that accounts for scanner-specific artifacts, the MRI volumes can be affected from variations such as data heterogeneity resulting from multi-site multi-scanner acquisitions, the presence of multiple sclerosis (MS) lesions and the stage of disease progression in the brain. Using the distributional divergence criteria, we evaluate the effectiveness of the normalization in rendering, under the distributional assumptions of segmentation approaches, intensities that are more homogenous for the same tissue type while simultaneously resulting in better tissue type separation. We also demonstrate the advantage of the decile based piece-wise linear approach on the task of MS lesion segmentation against a linear normalization approach over three image segmentation algorithms: a standard Bayesian classifier, an outlier detection based approach and a Bayesian classifier with Markov Random Field (MRF) based post-processing. Finally, to demonstrate the independence of the effectiveness of normalization from the complexity of segmentation algorithm, we evaluate the Nyul method against the linear normalization on Bayesian algorithms of increasing complexity including a standard Bayesian classifier with Maximum Likelihood parameter estimation and a Bayesian classifier with integrated data priors, in addition to the above Bayesian classifier with MRF based post-processing to smooth the posteriors. In all relevant cases, the observed results are verified for statistical relevance using significance tests.

1. Introduction

Multiple sclerosis (MS) is the most common neurological disorder affecting young adults in North America. MS is a demyelineating disease of the central nervous system (CNS) where myelin, the insulation around nerve fibre (axon) is attacked resulting in focal MS lesions (McAlpine, 1973). Magnetic Resonance Imaging (MRI) enables the visualization of MS lesions with very high contrast sensitivity and has proven to be critical in their identification. Most of the supervised parametric lesion identification and tissue type segmentation approaches (both automatic and semi-automatic) applied to brain MRI volumes rely, explicitly or implicitly, on strong assumptions regarding the shape of the underlying distribution of various tissue type intensities. These assumptions require images to have standardized intensity ranges (see, for instance, van Leemput et al., 2001; Dugas-Phocion et al., 2004). In fact, intensity standardization can be deemed necessary even for some unsupervised learning approaches that aim to make pooled estimates over multi-modal MRIs for lesion identification. The parameters of various automatic methods are learned from data that are assumed to have been sampled independently and identically distributed (iid) from a fixed, although arbitrary, distribution. However, large variations in intensity ranges can violate these assumptions significantly and adversely affect their success. These variations stem not only from the differences in the protocols of various MRI scan acquisitions, the different manufacturers and scanner-models, but also due the fact that scans from various sites can correspond to patients at varying stages of the disease. Data acquired from various sources, as well as from the same source

* Corresponding author at: Centre for Intelligent Machines, McGill University, 3480 University Street, Montréal, QC, Canada H3A 2A7. Tel.: +1 514 398 8702; fax: +1 514 398 7348.

E-mail addresses: mohak@cim.mcgill.ca (M. Shah), yiming.xiao@mail.mcgill.ca (Y. Xiao), nagesh@cim.mcgill.ca (N. Subbanna), simon@mrs.mni.mcgill.ca (S. Francis), doug@mrs.mni.mcgill.ca (D.L. Arnold), louis.collins@mcgill.ca (D.L. Collins), arbel@cim.mcgill.ca (T. Arbel).

1361-8415/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2010.12.003

but at different time points, generally do not have similar intensity ranges. Furthermore, it is well-known that the presence of pathology can impact the tissue intensity behaviors significantly. As a result, a lack of standard intensity scale makes it difficult to generalize the relative behavior of various tissue types across different volumes in the presence of disease.

To deal with such intensity variations, the raw MRI data undergoes some pre-processing before any meaningful learning for tissue classification can be done. There are two main steps in the pre-processing of MRIs: (1) the intensity inhomogeneity correction (Belaroussia et al., 2006) and (2) the intensity normalization.[1] The former is generally aimed at addressing scanner specific variations in individual MRIs. In this paper, we focus on the latter, that of, standardizing the intensity scales in MRIs.

Various intensity normalization methods have been proposed (see Section 2 for a review) and have met with varying degrees of success and also have their respective limitations. For this study, we choose the two-stage method of Nyul et al. (2000) for evaluation owing mainly to its ease of computation, customizability and speed while maintaining high accuracy. The method consists of a training stage to find the parameters of the standard scale and a transformation stage that maps the histograms of candidate volumes to the standard histogram scale (also known as Intensity of Interest or IOI) in a piece-wise linear manner. The scale has been shown to be relatively stable across different variations of the piece-wise intervals. In this work, we study these variations both qualitatively and quantitatively. We aim at assessing the effectiveness of this normalization on multi-site, multi-scanner MRIs of human brains with MS to examine the efficacy of the intensity normalization in the presence of pathology. We also investigate the role of normalization on the behavior of resulting tissue intensities with regard to the assumptions made by various automatic approaches (e.g., tissue classification) regarding their distributions. We further refine this analysis of the normalization procedure in the presence of pathology based on the MS lesion loads in the MRIs. The lesion load-dependent analysis enables us to evaluate the effect of varying degrees of pathology on the intensity scale normalization and its potential impact on the intensity behavior of various tissue types. Finally, we examine the efficacy of the normalization in obtaining a better separation between different tissue types as well as in MS lesion identification in multi-modal scenario. With image normalization becoming commonplace, simpler and effective methods such as that of Nyul et al. (2000) having an obvious appeal, for pooled analysis over the images and such a validation seems both critical and timely. Moreover, since in most cases, the application domain involves MRIs with pathology, such a validation, and more importantly, justification is rendered even more critical.

In order to validate the method, we have chosen a dataset consisting of 21 multi-spectral brain MRI volumes from 13 different scanners from three different scanner manufacturers: GE, Siemens and Philips. Our analysis is focused on single modality images since the normalization is applicable on individual modalities. With T1-weighted (T1w), T2-weighted (T2w) and Protein Density (PD) weighted images, the dataset contains a total of 63 volumes.[2] We mainly utilize statistical divergence criteria to assess whether the typical assumption of Gaussian distribution over individual tissue types in single modalities hold. In order to verify this, we compare the empirical data distribution against a Normal distribution generated according to sample statistics. Next, we perform a similar analysis for each modality after intensity normalization. The analysis after normalization, in addition to testing for the Gaussian compliance, also evaluates the role of intensity normalization in making the data more amenable to applications under this assumption. Performing evaluations on multiple modalities further allows us to test if the advantages rendered by the intensity normalization are indeed consistently observed across various image acquisition modes. Next, we break the above analysis to examine effects on scanners from different manufacturers to identify scanner-specific dependencies. We, then, also refine the analysis to take into account the effect of lesion load on intensity normalization. We follow this by a preliminary analysis on the effect of normalization in increasing separation between intensitites of different tissue types while at the same time making intensities of the same tissue more homogenous using both statistical divergence and applying a *k*-means clustering algorithm. Finally, we extend this analysis to multi-spectral case since most MRI segmentations algorithms take multi-modal intensities into account. In addition to the qualitative data analysis with scatterplot visualization, we compare the performances of three algorithms on identification of MS lesions from multi-spectral MRI volumes normalized according to the normalization method of Nyul et al. (2000) to account for biases introduced by any one approach: a *k*-mean clustering approach, an outlier detection based approach (van Leemput et al., 2001) and a Bayesian approach in conjunction with Markov Random Field (MRF) based smoothing (Harmouche, 2006). Further, to study the advantage of the method of Nyul et al. (2000) as well as the use of decile based piece-wise normalization against a simpler method, we also perform the respective comparisons on MRI intensities normalized according to a linear normalization method. In order to further study the dependence of these comparisons on the complexity of the algorithm used, we perform a similar analysis using Bayesian learning algorithms of varying complexity: a standard Bayesian classifier with maximum likelihood estimation, a Bayesian classifier with prior information, and a Bayesian classifier with MRF based post-processing. For all the relevant cases, we also perform a statistical significance test of the results obtain to verify whether the observed differences are indeed significant.

The rest of the paper is organized as follows: Section 3 describes the decile formulation of the intensity normalization of Nyul et al. (2000) utilized in this work. Section 4 gives details of the data acquisition and the image non-uniformity and inhomogeneity correction pre-processing. Section 6 presents the evaluation framework followed by qualitative and quantitative evaluation results in Sections 6.1 and 6.2 respectively. Results are discussed in Section 7. Finally, we conclude in Section 8.

2. Approaches to intensity normalization

Various intensity normalization methods have been proposed to deal with intensity scale inhomogeneity of various tissue types and have met with varying degrees of success. One of the first approaches to intensity normalization was the Dynamic Histogram Warping approach of Cox et al. (1995) aimed at finding an optimal alignment between two data sequences with different lengths. The optimization criterion used to measure the quality of the alignment was generally a matrix of monotonic and separable cost functions between intensity occurences. Cox et al. (1995) applied this approach to a pair of stereo images. Even though this approach performs acceptably in the case of stereo images, it is not easily customizable to the case of MRI intensities. Moreover, the results in the case of MRI intensity normalization are not uniform across sequences (Bergeest and Jäger, 2008). Next, we discuss intensity normalization methods developed specifically for MRI images.

Among the approaches aimed at MRI intensity normalization, some of the main include those based on utilizing even-order histogram derivatives (Christensen, 2003), template histogram matching using multiplicative correction field (Weisenfeld and Warfield,

[1] Intensity normalization is also referred to as intensity standardization. However, we use the former term in this paper.

[2] The dataset was obtained courtesy of NeuroRx Research, Montreal.

2004), region-specific normalization (Hellier, 2003), and mapping the normalization problem to non-linear registration problem (Jäger et al., 2006). Other intensity normalization approaches have also been proposed typically in the context of image registration and spatial alignment (see, e.g., Friston et al., 1995; Nestares and Heeger, 2000; Guimond et al., 2001; Bosc et al., 2003). These approaches have been shown to perform well in their respective contexts and sometimes over the specific image pairs under consideration. These later methods have not be validated extensively in other domains. Below, we discuss methods specifically proposed with regard to intensity normalization *independent* of any other image pre-processing such as image registration.

Christensen (2003) proposed a method utilizing even-order histogram derivatives to calculate characteristic values of white matter in various image modalities. This characteristic value is then utilized to normalize the image intensity values of various tissue types. The method depends on the even order derivatives of the image histogram. However, these derivates were chosen by optimizing an error function computed from simulated image histograms. Moreover, the methods seems to work better only in conjunction with a threshold-sensitive brain segmentation algorithm.

The approach of Weisenfeld and Warfield (2004) relied on matching a template histogram to a reference model density using multiplicative correction field. The approach aimed at altering global statistics of MRI of the template histogram to the model density while preserving local feature contrast. This was done by choosing a parameter field that minimized the Kullback–Leibler divergence measure between the two densities. However, performing a pixel-wise correction estimation makes this method slower than others described here.

Hellier (2003) proposed a head-region-specific intensity normalization of MRI brain volumes. The source and target histograms are approximated using a Gaussian Mixture model (GMM) computed through an Expectation–Maximization algorithm. This is followed by polynomial correction for intensity smoothing. The method was described in the context of the head images (including fat and muscle) and not on the images of brain extracted from the full head MRIs. Even though it can conceivably be applied to skull-stripped images by disregarding the fat and muscle modeling components of the GMM, the difficulty involved in such application and the effectiveness in such scenario are not known. The specificity of this method to head-region was also noted by Bergeest and Jäger (2008).

Jäger et al. (2006) proposed an intensity normalization method by mapping the problem to an image registration problem. They consider the probability densities of tissue types as images and apply a distance measure based non-rigid registration to the joint histograms resulting in a non-linear correction function for MRI intensities. The approach applies to multi-modal image sequences and requires the histogram to come from at least two modalities. The method gives good results. However, performing non-rigid registration operations over histograms and further utilizing all image sequences makes this method slower than approaches that perform scaling independent to other image pre-processing and over individual modalities.

Finally, Wells et al. (1996) provides a simultaneous estimate of tissue classification and estimation of a smoothly varying intensity inhomogeneity artifact. This method requires the specification of the class conditional probability density functions, which are difficult and time-consuming to acquire accurately. The densities are derived interactively from pre-identified representative training voxels. The method then uses an iterative expectation maximization manner to refine the parameter estimates for both classification and intensity normalization. Hence, if a tissue class conditional probability density for a test MRI scan is closer to the wrong class in the statistical model derived from the identified pixels, the initial segmentation model can be poor. One of the consequences of this observation appears also in terms of the requirement induced by the method to perform separate learning over multiple sites.

The two-stage method of Nyul et al. (2000) has some significant advantages over the methods described above. It is both easier to customize to various anatomical regions (and not just head, or skull-stripped images) and is fast in practice. Moreover, the method does not sacrifice accuracy for gains in speed. Some of these methods were compared by Bergeest and Jäger (2008). Their results further confirms these conclusions. For instance, the method of Jäger et al. (2006) was found to be about 30 times slower than that of Nyul et al. (2000) (taking approximately 1 min against 2 s per image volume for the latter) while the method of Weisenfeld and Warfield (2004) was found to be significantly slower (approximately 30 min for every MRI).[3] As a result the methods of Nyul et al. (2000) has been used in the intensity normalization step in many automatic MRI analysis and lesion detection approaches (see, for instance, Moonis et al., 2002; Anbeek et al., 2004; Datta et al., 2006; Harmouche, 2006; He et al., 2008; Bergeest and Jäger, 2008; Khan et al., 2008; Karimaghaloo et al., 2010; Elliott et al., 2010; Scully et al., 2010; Gronenschild et al., 2010). Another major advantage of this approach is that it does not rely heavily on specific statistical properties of tissue classes and has a lower computational complexity. Therefore, the method of Nyul et al. (2000) was chosen for evaluation in this work.

Earlier versions of their approach appeared in (Nyul and Udupa, 1999a; Nyul and Udupa, 1999b) where they present a quantitative and qualitative evaluation of this intensity normalization method with regard to intra- and inter-patient variations. However, the original formulation suffers from some limitations in the presence of pathology in the MRIs, in particular when the goal is tissue classification in the presence of this pathology. These limitations are detailed in Nyul and Udupa (1999c) and Nyul et al. (2000). Immediately relevant among these is how, in the case of application to MRIs of brains with MS, the method displays a switching behavior. That is, in the presence of pathology, the histogram foreground landmarks (the peak of the highest intensity or the mode) may not refer to the same tissue type in different images thereby rendering incorrect mapping of tissue types. This issue of switching behavior with the original formulation of the approach, when applied to MRIs of human brains with pathology, comes essentially from treating the whole intensity range over a single linear scale and performing the corresponding mapping. Nyul et al. (2000) suggested a variation of the original formulation with deciles piece-wise linear mapping to address these limitations. However, there is no evaluation available, to the best of our knowledge, of the effectiveness of this normalization method across scanners from different brands or manufacturers as well as variations occurring due to different machines of the same brand. Furthermore, it is not conclusively confirmed if the intensity normalization is equally effective in the presence of pathology in the brain and reflected in the resulting MRIs.

3. Decile normalization method of Nyul et al. (2000)

We consider an image $\mathscr{I}$ consisting of a set of voxels E and an intensity mapping function $f_{ei} : E \rightarrow \mathbb{N}_0$ that maps each voxel

[3] The evaluation was performed in significantly low resolution T1- and T2/FLAIR images, all from Siemens Avanto 1.5 T scanner. The T1-weighted images had a resolution of $208 \times 256 \times 19$ with 0.86 mm^2 and 7.2 mm slice thickness and TE = 14 and TR = 510. The T2 weighted FLAIR images had a resolution of $408 \times 512 \times 19$, pixel size of 0.43 mm^2 and 7.2 mm slice thickness and TE = 143 and TR = 9000. The simulations were performed on a computer with Intel Core2 CPU T5500 with 1.66 GHz and 2 gigabyte RAM. The reference histograms were obtained from lesion-removed images.

element $e \in E$ to an intensity value $i \in \mathbb{N}_0$, where $\mathbb{N}_0$ is the set $\{0, 1, 2, \ldots\}$ of natural numbers. Hence, $f_{ei}(e) \geqslant 0$, $\forall e \in E$ with the equality also satisfied if there is no measured data for element e. Each volume image can be described jointly by the set of voxels E and the associated intensity mapping function f_{ei}, $\mathscr{I} = (E, f_{ei})$. Hence, $\mathscr{I}$ represents, in our case, 3D image acquired through MRI. However, f_{ei} is generally not known since we have access only to the intensity values for each e.

We denote by $\mathscr{A} = \{A_1, A_2, \ldots, A_k\}$ the set of acquisition modalities of MRIs such that $|\mathscr{A}| = k$. In our case, we focus on mainly three image modalities: the PD, T1 and T2 weighted images giving $k = 3$.

Let $\mathscr{I}_{\mathscr{A}}$ denote the superset of all images $\mathscr{I}$ that can be acquired according to the protocols in $\mathscr{A}$. The histogram of an image $\mathscr{I}$ is denoted by $\hbar = (I, f_{ic})$ such that $I \subset \mathbb{N}_0$ is the set of all possible intensity values in $\mathscr{I}$ and $f_{ic} : I \to \mathbb{N}_0$ is intensity to count mapping function such that $f_{ic}(i) = -\{f_{ei}(e) = i\}-$. That is, the function $f_{ic}(i)$ outputs the number of voxels that have an associated intensity value of i. Finally, let i_{min} and i_{max} denote the minimum and maximum intensity values respectively in an image $\mathscr{I}$. That is, $i_{min} = \min\{f_{ei}(e) - e \in E, f_{ei}(e) \neq 0\}$ and $i_{max} = \max\{f_{ei}(e) - e \in E, f_{ei}(e) \neq 0\}$

3.1. Choosing histogram landmarks

The tails of the image histograms are generally pruned so as to make the algorithm robust against artifacts and outliers that may result in inter-patient and scanner variations. Although not the focus of this study, this also helps in minimizing intra-patient variations. This pruning results in an intensity range generally called intensity of interest (IOI). Let p_{low} and p_{high} denote the minimum and maximum percentile values of the overall intensity range that defines the boundaries of this IOI. That is, if we consider $p_{low} = 1$ and $p_{high} = 99$, then we effectively prune the lower and upper *1* percentile of the intensity values. Now, let $i^{\hbar}_{min}$ and $i^{\hbar}_{max}$ be the minimum and maximum intensity values in the histogram $\hbar$corresponding to p_{low} and p_{high} respectively. We consider the more common bimodal histograms for the MR images. The peak (first mode) in the histogram distribution corresponds to the background while the second mode corresponds to the foreground intensities. In order to identify the landmarks, the foreground was identified using a thresholding approach. The overall mean intensity of the image was chosen as the threshold to separate the background from the foreground. The histogram landmarks are the histogram-specific parameters used for intensity normalization and include p_{low} and p_{high} described above in addition to the second modes at each decile as detailed below.

3.2. The normalization

In the decile formulation of the normalization method, using the histogram for the foreground of the image, we have the intensity-landmark configuration C_L as:

$$C_L = [p_{low}, m_{10}, m_{20}, m_{30}, m_{40}, m_{50}, m_{60}, m_{70}, m_{80}, m_{90}, p_{high}]$$

where each $m_i, i \in \{10, 20, \ldots, 90\}$ denotes the ith percentile of the histogram associated with the foreground part of the image with mode m. Due to the intervals used here, this variation is also referred to as the decile version. Finally, let i^s_{min} and i^s_{max} be the minimum and maximum intensity values in the standard scale respectively (the superscripts denoting the standard scale). Given this setting, the intensity normalization algorithm is as shown in Tables 1 and 2. A graphical illustration is provided in Fig. 1.

The two-stage intensity normalization algorithm can be summarized as follows. For each image in the training set of images, we determine the image histogram. From this image intensity histogram, we determine the range of intensity values of interest. That is, we determine the intensity values $i^{\hbar}_{min}$ and $i^{\hbar}_{max}$ corresponding to p_{low} and p_{high}, the lower and upper percentile intensity bounds respectively. The values falling outside these bounds are discarded as outliers. Now, we identify the intensity values $m^i_{10}, \ldots, m^i_{90}$ that correspond to each decile of this intensity range of interest. The minimum and maximum intensity values of the image $i^{\hbar}_{min}$ and $i^{\hbar}_{max}$ are then mapped to the corresponding minimum and maximum intensity values i^s_{min} and $imax^s$ of the standard intensity scale respectively. Note that i^s_{min} and i^s_{max} are the user selected parameters of the learning algorithm. The choice generally depends on the trade-off between the desired resolution of the resulting intensity normalization scale and the desired efficiency in the intensity mapping in each segment of this intensity range under the decile formulation. Once this mapping is done, new landmarks corresponding to each decile of the mapped image $m'^i_{10}, \ldots, m'^i_{90}$ are calculated. The landmarks for the standard scale are then calculated from the rounded means of each of the landmarks from the mapped image from the training set of images.

The standard scale landmarks obtained at the end of the training phase are then used to perform the transformation of a new MRI image to the standard scale as follows: Given a new MRI image $\mathscr{I}_j$, we compute the histogram $\hbar_j$ of the new image and determine the minimum and maximum intensity values corresponding to p_{low} and p_{high} percentiles. We also determine the intensity values corresponding to each decile of this intensity histogram from the intensity values corresponding to p_{low} and p_{high} respectively. The end points of each decile is an intensity landmark for the new image. Hence, these intensity values corresponding to the decile landmarks enables segmenting the image in ten segments. Each of

Table 1
Intensity normalization algorithm: Training. Adapted from Nyul and Udupa (1999b) and Nyul et al., 2000.

Algorithm: Training
Input: The Image set $\mathscr{I}_j$, $(j = 1, 2, \ldots, n)$ such that $\mathscr{I}_j \subset \mathscr{I}_{\mathscr{A}}$, p_{low} and p_{high}, i^s_{min}, i^s_{max}, and C_L
Output: $\{m_k - k = 10, 20, \ldots, 90\}$
begin
for $j = 1$ to n do
compute the histogram $\hbar_j$ of $\mathscr{I}_j$;
determine intensity values $i^{\hbar_j}_{min}$ and $i^{\hbar_j}_{max}$ corresponding to p_{low} and p_{high}, and the landmark locations $m^j_{10}, m^j_{20}, \ldots, m^j_{90}$ on $\mathscr{I}_j$;
map $[i^{\hbar_j}_{min}, i^{\hbar_j}_{max}]$ of $\mathscr{I}_j$ onto $[i^s_{min}, i^s_{max}]$ linearly;
find the new mapped landmark locations $m'^j_{10}, m'^j_{20}, \ldots, m'^j_{90}$;
end for;
calculate the rounded means $m^s_{10}, m^s_{20}, \ldots, m^s_{90}$ of $m'^j_{10}, m'^j_{20}, \ldots, m'^j_{90}$ respectively, over $j = 1, 2, \ldots, n$;
end

Table 2
Intensity normalization algorithm: Transformation. Adapted from Nyul and Udupa, 1999b and Nyul et al., 2000.

Algorithm: Transformation
Input: An Image $\mathscr{I}_j \in \mathscr{I}_{\mathscr{A}}, p_{low}, p_{high}, i^s_{min}, i^s_{max}, m^s_{10}, m^s_{20}, \ldots, m^s_{90}$.
Output: Transformed image $\mathscr{I}^s_j$
begin
compute the histogram $\hbar_j = \left(I_j, f^j_{ic}\right)$ of $\mathscr{I}_j$;
determine $i^{\hbar_j}_{min}$ and $i^{\hbar_j}_{max}$ corresponding to p_{low} and p_{high}, and the landmark locations $m^j_{10}, m^j_{20}, \ldots, m^j_{90}$ on $\mathscr{I}_j$;
map sections of the scale of $\mathscr{I}_j$ linearly to the standard scale I_s of the standard histogram $\hbar_s = (I^s, f^s_{ic})$;
map the intensity value of every voxel $e \in \mathscr{I}_j$ so as to obtain $\mathscr{I}^s_j$;
end

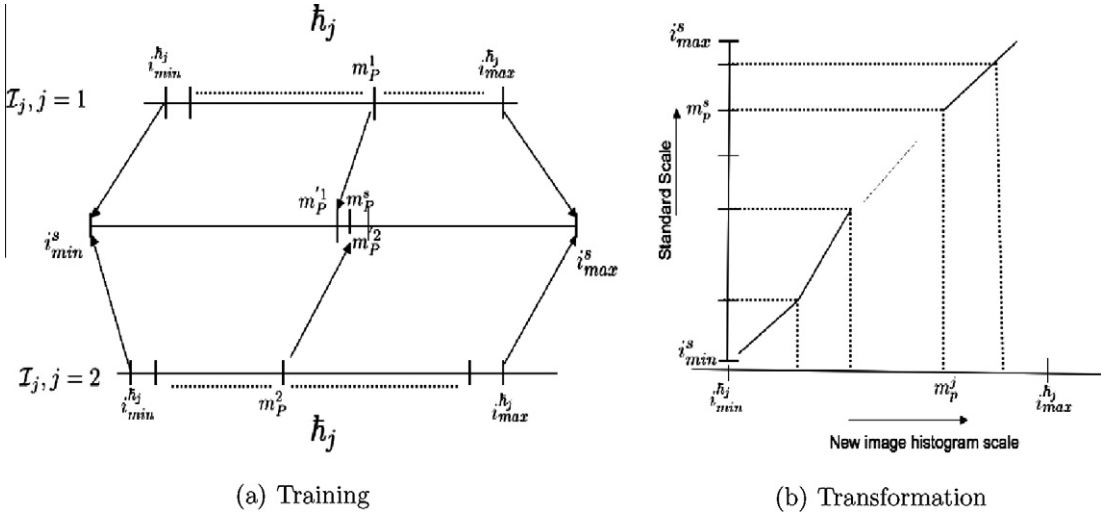


Fig. 1. Illustration of Nyul Normalization adapted from (Nyul et al., 2000). Please refer to the text for notations. *Left:* The figure illustrates the training stage with the intensity landmarks from the input images (2 shown here, one top and the other bottom) mapped to the standard scale (middle). The standard scale intensity landmarks are then obtained by taking the means of the mapped intensity landmarks. *Right:* The figure illustrates the transformation stage where a new input image histogram (on the horizontal axis) is mapped to the standard scale in a piece-wise linear fashion.

the segment is then linearly mapped to the corresponding segment of the standard image intensity scale determined in the training stage of the algorithm. This outputs the standardized image $\mathscr{I}_j^s$ of the input image $\mathscr{I}_j$.

4. Data acquisition and pre-processing

A group of subjects MRI scans (T1w, T2w, and PDw modalities) were obtained to test the performance of the MRI intensity normalization technique of Section 3 across MRI scanners made by different manufacturers, different brands of MRI scanners from the same manufacturer, and MRI scans of subjects that have different loadings of MS lesion volumes. The MRIs were selected based on a combination of diverse criteria. The selected group includes MRI scans of a total of 21 subjects, with seven subjects acquired from MRI machines made by each of the three major MRI machine manufactures (GE, Phillips, and Siemens) making a total of 63 image volumes to test on. Table 3 provides details on the scanners used in the study along with the associated number of patients. Within each sub-group of seven cases, the sampled patients also possess varying sizes of ventricles, and varying loads of MS lesions. Each MRI volume has a resolution of $1 \times 1 \times 3$ mm for voxels with the images acquired axially. Each axial slice is 3mm thick with each volume containing about 50 slices to cover the brain from vertex to foramen magnum.

Different MRI modalities acquired may not correlate spatially upon acquisition due to various reasons such as patient movement and scanner behavior over time. An alignment is required as a result so as to obtain a deterministic mapping between voxels of the MRI image of the same patient in different modalities. Therefore, prior to sampling the images the different image modalities are aligned to a common stereotaxic space so as to achieve this spatial correlation between brain volumes (Collins et al., 1994).

This brings up the question as to whether to perform inhomogeneity correction before, or after, the intensity normalization since the order of performing these two processing steps can have potential impact on the resulting MR Images. Indeed, the combined effects of these two techniques can potentially generate different final results depending on the order in which these two processes are performed and the number of iterations performed for them. Madabhushi and Udupa (2005), have investigated the image quality after interplay between the intensity normalization and inhomogeneity correction procedures. The results of their analysis of nearly 4000 cases suggest that improved normalization can be achieved by preceding it with the inhomogeneity correction process. The data also show that longer sequences of repeated correction and normalization operations do not considerably improve image quality. Therefore, all MRI scans from the database were first processed by a single iteration of image inhomogeneity correction followed by an iteration of the decile formulation of the image intensity normalization algorithm of Section 3. The N3 method of Sled et al. (1998) was used to perform the intensity inhomogeneity correction.

Another important factor contributing to the variations in the tissue intensities is the limited resolution of MRIs which results in more than one tissue represented by each voxel. As a result, the boundaries separating the tissues of one type from another can have a mixture of intensities from both these tissue types resulting in partial volume effects (PVEs). The PVE makes it difficult to categorize the voxels to any single tissue type since the intensity behavior of voxels depicting such behavior is indeed not representative of any single tissue type. Further, the effects of normalization on voxels with PVE are not clear. Hence, we limit ourselves to studying the intensity behavior of various tissue types from voxels that are free from such partial volume effects. We refer to such voxels as "pure" tissue voxels. Thus, on average, approximately 1000 "pure" samples of each tissue type were manually identified on the data from each subject. Subsequently, the effects of normalization are studied and evaluation performed on these "pure"

Table 3
Details of Various machines used to obtain MRI volumes for this study; "NA" denotes the unavailability of the information on the corresponding coil type.

Manufacturer	Model	Coil	Field (T)	No. of patients
GE	Genesis Signa	NVPA	1.5	1
	Genesis Signa	Head	1.5	2
	Signa Excite	Head	1.5	3
	Signa	Head	1.5	1
Philips	Intera	Head	1.5	5
	Infinion1.5T	NA	1.5	2
Siemens	SymphonyVision	NA	1.5	2
	Avanto	NA	1.5	1
	Symphony	NA	1.5	2
	Allegra	NA	3	2

tissue samples for each tissue category. Voxels corresponding to the tissues belonging to four tissue classes, i.e., white matter (WM), cortical gray matter (CGM), deep gray matter (DGM) and Cerebrospinal fluid (CSF), were manually sampled uniformly from various anatomical locations. The axial view of the brain volume is primarily used for sampling owing to higher in-plane resolution. The two complementary views (Coronal and Sagittal) were used to avoid voxels with partial volume effect and other variations from being sampled (for instance, voxels from sulci or fissures). Further, sampling is performed in alternate slices so as to obtain representative samples from the whole volume averaging about 20 sampled slices per volume.

Due to the fact that MS lesions are more difficult to be identified correctly, lesions of the selected clinical cases are first identified automatically using an automated method utilizing an expert based clinically established lesion classification protocol (Francis, 2004) and are then further validated by five expert radiologists. In the particular dataset that we had access to, the protocol for establishing "ground truth lesions" was as follows: "ground truth lesions" were selected based on a consensus among experts with a voxel labeled as lesion when three or more experts agree. For example, a voxel can be labeled as a lesion when 3 out of 5 experts agree. The data selected for this study also contain brain volumes with varying clinical lesion loads (measured in terms of the volume of brain tissues labeled as lesions) from low (very sparse lesions, less than 5 cc lesion load) to very high (more than 25 cc of lesion load). We also perform a lesion load-dependent analysis of the normalization behavior of tissue intensities in MRI volumes in Section 6.4.

5. Evaluation methodology and experiments

The standard scale histogram parameters were obtained from a training set of 100 MRI volumes (T1w, T2w and PDw) of human brains with MS from scanners from various manufacturers at multiple sites with subjects at varying stages of MS pathology. The parameters then defined the standard scale used for normalization of subsequent images. The same parameters were used for all image modalities. Note that the only dependence of the method in terms of landmark identification comes from the training data. As a result, we have tried to incorporate as wide a range as possible to reflect variations in scanners, sites as well as the extent of pathology present in the MRIs utilized for the purpose. This requires a higher number of training MRIs so that the resulting landmark configuration is robust to such variations.

The evaluation of the normalization algorithm was performed keeping in perspective the potential advantages and use of the resulting normalized images. The main factors that were explored include:

(1) The effects of MRIs coming from heterogenous sources including scanners from different manufacturers as well as different scanner-models from the same manufacturer.

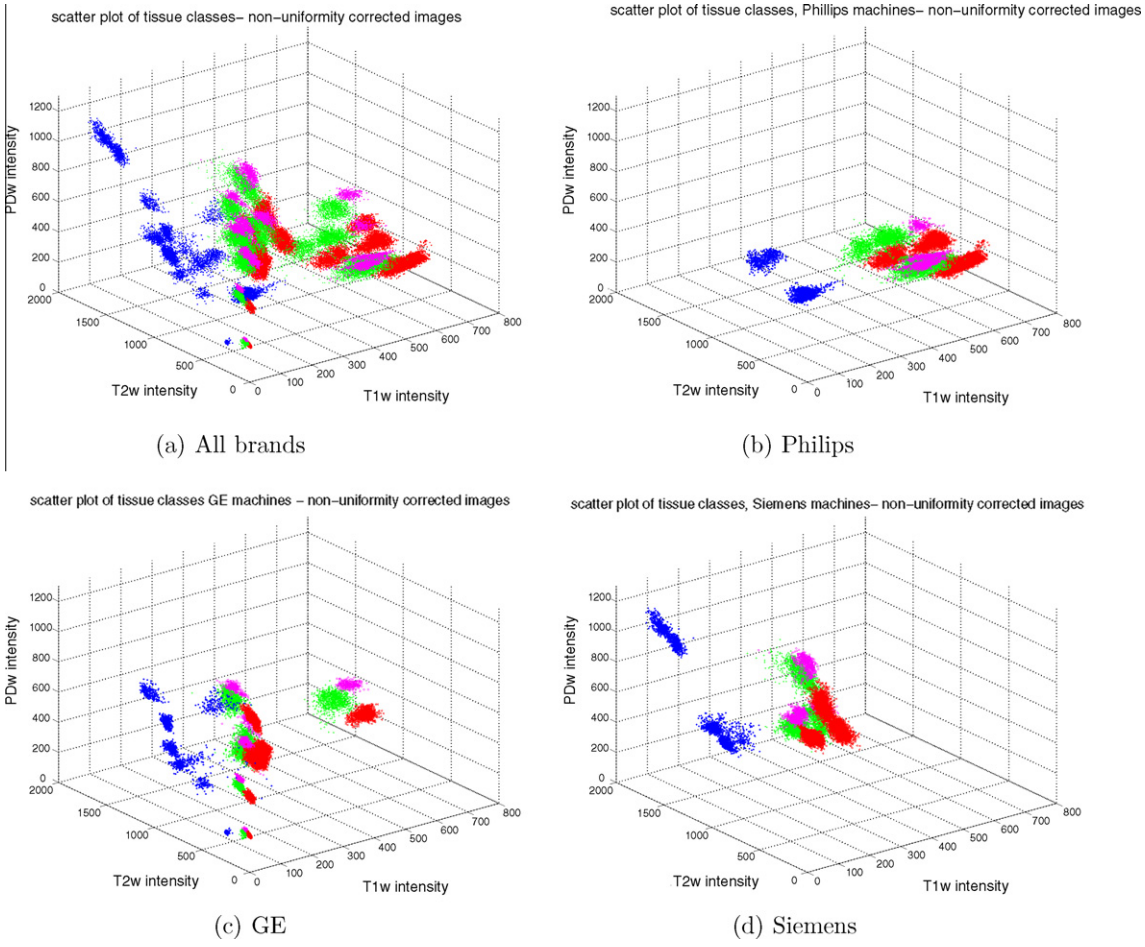


Fig. 2. Scatterplots of various brand machines before normalization. The intensities in red denote white matter, in green denote cortical gray matter, in magenta denote deep gray matter, and in blue denote CSF. Lesions are not included in the un-normalized image intensities since they were generally sparse and did not exhibit a characteristic behavior in the intensity space. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(2) Testing whether the tissue intensities behave in accordance with the assumptions made by various automatic lesion segmentation and analysis techniques (e.g., whether they are Gaussians).
(3) Whether normalization helps in increasing the efficacy of tissue classification.
(4) The effect of the amount of lesion load on the accuracy of tissue intensity normalization.
(5) The advantage of the decile based normalization against a simple linear normalization of data in automatic segmentation of tissues in multi-modal brain MRIs of patients with MS.

The dataset chosen for testing has been selected so as to take all these factors into account. In particular, to complement the qualitative multi-modal analysis, we perform the following quantitative evaluation experiments:

(a) We study divergence between the empirical distribution and a Normal distribution defined according to sample statistics. Doing this before normalization and then after normalization gives an indication of whether the data is more (or less) representative of a Gaussian over sample statistics. Further, we study these effects across various modalities. The aim to this exercise is to evaluate if normalization has any effect on this data distribution assumption and if so, whether this effect makes the data more amenable to the application of automatic segmentation approaches.
(b) We break the above analysis over scanner brands to study if the above effect is indeed visible across different settings and if this is the case uniformly.
(c) We further study the effect of varying lesion load on the above analysis of (a).
(d) We perform validation experiments to ascertain whether the normalization helps in homogenizing same tissue type intensities while simultaneously increasing inter-tissue distances using a simple *k*-means clustering on the sampled (pure) tissue voxels.
(e) The analysis of (d) is extended to full scale test data to study the empirical benefits of normalization on the problem of MS lesion identification. The use of deciles scale increases the number of histogram landmarks yielding a piece-wise linear normalization model for intensities. It has also been argued that the use of a higher number of histogram landmarks result in a better defined standard histogram against which the tissue intensities are normalized (Nyul et al., 2000). We employ *k*-means, an outlier based and a Bayesian segmentation algorithms over multi-modal MRIs to evaluate the advantage of normalization, especially, in comparison to a simpler linear scaling based method to further confirm this argument.
(f) Finally, we study the dependence of this effect on the degree of complexity of learning algorithm, over multi-modal MRIs to verify if the effect observed in (e) can indeed be attributed to normalization (and not the level of sophistication of the learning algorithm). To assess this dependency, we utilized three Bayesian classifiers with increasing complexity: the first is a simple Bayesian classifier using a Maximum Likelihood parameter estimation, the second incorporates an integration of a data prior while the third performs a Markov Random Field based smoothing as a post-processing step.

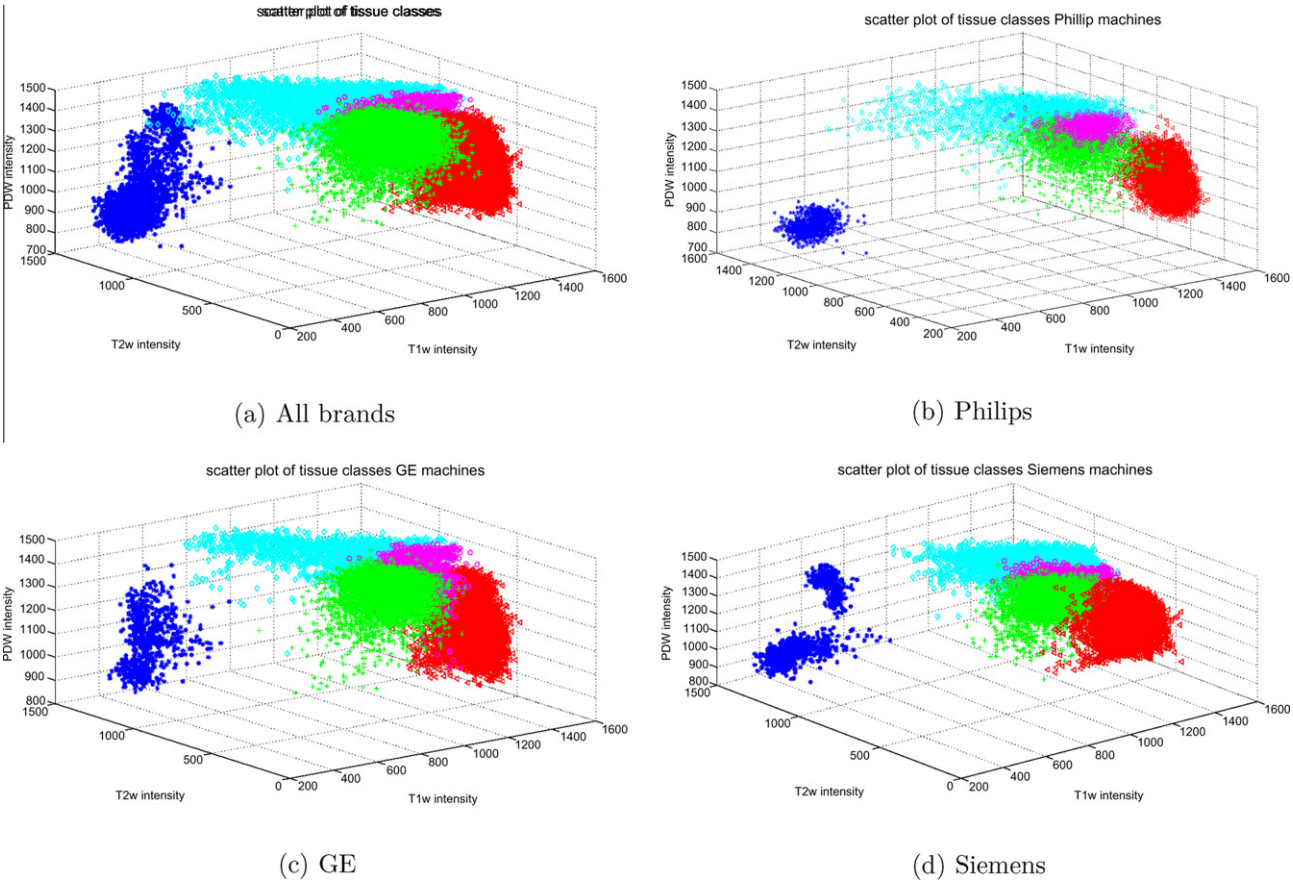


Fig. 3. Scatterplots of the intensity resulting from images acquired from various brand machines after normalization. The intensities in red denote white matter, in green denote cortical gray matter, in magenta denote deep gray matter, in cyan denote lesions and in blue denote CSF. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Finally, we also study the statistical significance of the results obtained in relevant cases using a matched pair *t*-test over the sampling distribution with a significance level of 0.05 (i.e., 95% confidence).

6. Results

6.1. Qualitative evaluation

We first present qualitative effect of the intensity normalization on the multi-spectral MRI volumes. We study both the overall effect of normalization on MRI volumes and the effect of normalization on MRIs obtained from scanners from various manufacturers. Fig. 2 presents the scatter plots of the multi-spectral volumes grouped in tissue types for all MRI volumes as well as for the MRI volumes grouped by scanner manufacturer. Each point in the scatterplot represents the three dimensional intensity vector for the corresponding voxel with each element in the vector denoting the intensity value in one of the acquired image modalities (T1w, T2w, PDw). The same data after intensity normalization is shown in Fig. 3.

From Fig. 2 it can be seen that intensities of each set of tissue classes (WM, Deep GM, Cortical GM, and CSF) resulting from images acquired from machines of a particular manufacturer have a characteristic spatial distribution. When all subjects are plotted together, the variations in imaging acquisition from different scanner brands and their sub-series are visible.

As shown in Fig. 3, the variations in the position of tissue intensity distributions relative to each other are largely reduced as a result of the intensity normalization. The same effect can also be seen in the case of scanners from different manufacturers. A standardized characterization of intensity distributions for various tissue types on a standard scale has significant implications on the performance of both the automatic lesion identification as well as tissue segmentation algorithms. The only concern seem to be in the plots for the Siemens machine after normalization (Fig. 3d) where CSF seem to form two different clusters. However, it was found upon examination of the data that an additional cluster was formed as a consequence of a slight mismatch in the normalization of two volumes whereby a brain mask was applied at a later stage. This effect is hence not attributable to any anomaly in the image normalization technique. Some qualitative results of the effect of intensity normalization over MRI volumes with different lesion loads are later shown in Fig. 8.

6.2. Quantitative evaluation

As a first step toward quantitatively evaluating the efficacy of the intensity normalization on MRIs of human brains with

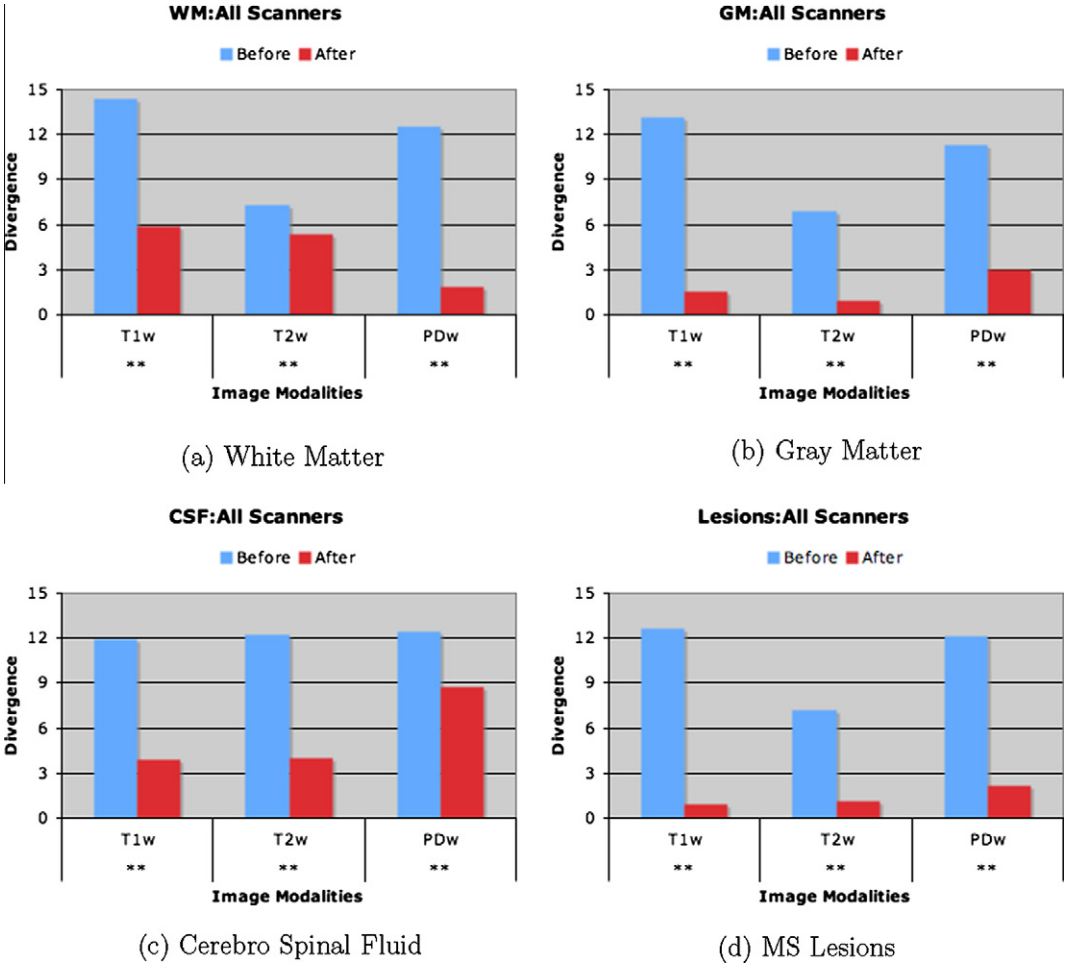


Fig. 4. Divergence measures between distribution fits before and after normalization over all MRI volumes taken together. The blue bars in each graph denote, for a given modality and tissue type, the Jeffreys divergence measure between the Gaussian distribution centered at the sample mean with sample variance, and the histogram model generated from tissue intensities on the un-normalized volumes. The red bars correspond to the same measure over volumes after normalization. An absence of bar indicates that the divergence was too small to plot. A "**" under each comparison denotes the result to be statistically significant at 95% confidence level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

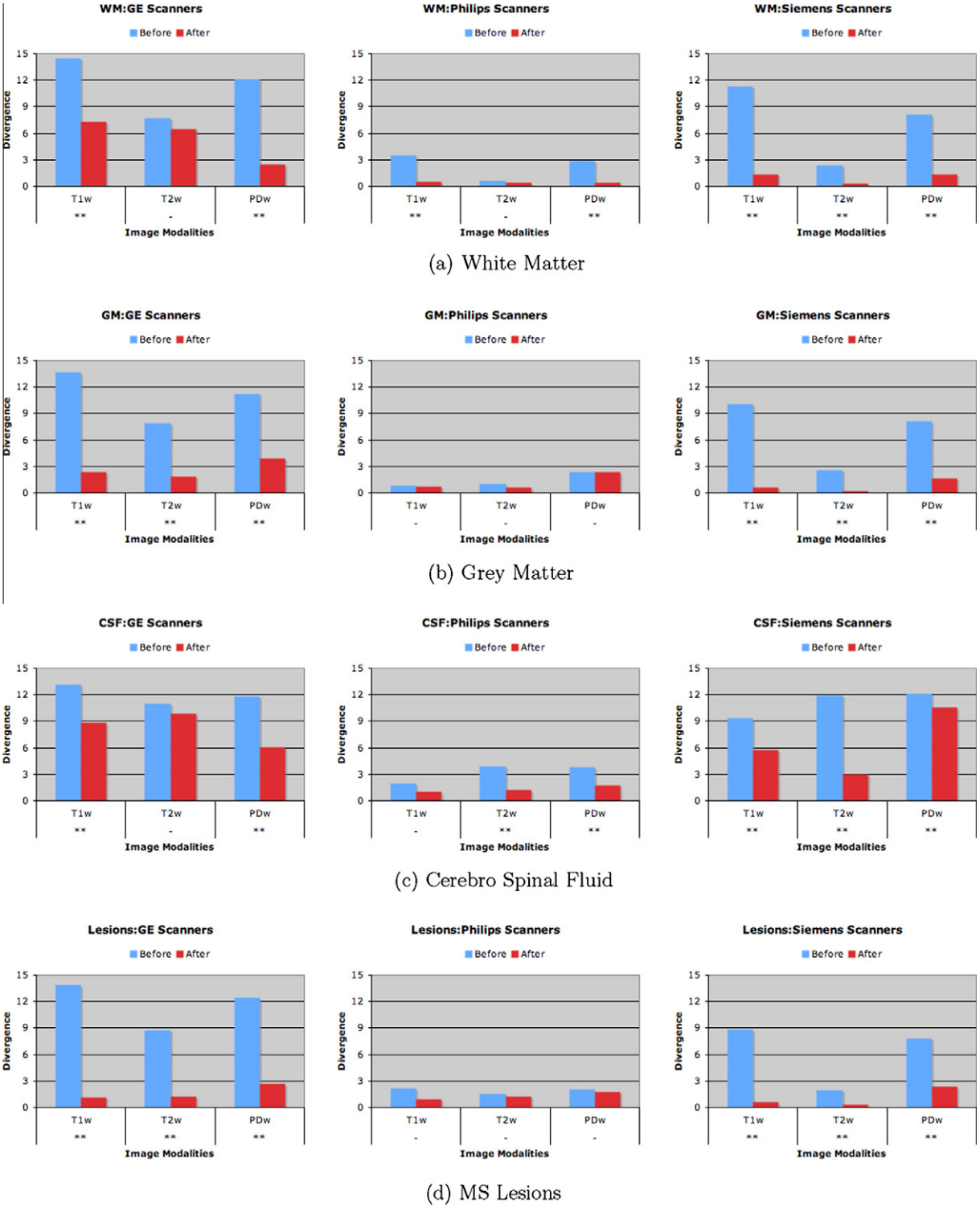


Fig. 5. Breaking the analysis of Fig. 4 according to scanner brands. Each row corresponds to measures for a given tissue type. Each column refers to divergence measures for MRI volumes from different machines: GE, Philips, and Siemens. The conventions are the same as given in the legend of Fig. 4. A "-" indicates that the corresponding results were not found to be statistically significant at 95% confidence level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pathology, we wish to investigate how representative the voxel intensities from various tissue types are of a Gaussian model built around the necessary statistics computed from the data. The null hypothesis assumes that the voxel intensities come from an underlying Gaussian distribution centered at the sample mean with sample covariance. Hence, if this hypothesis holds, we should not find any significant advantage towards data modeling as a result of intensity normalization. We should be able to quantify this effect in terms of the proximity of voxel intensities to a Gaussian distribution as a result of the intensity normalization. Let us now verify if the null hypothesis holds.

Since the intensity normalization is applied on a single modality basis, we restrict all our analysis for distributions of tissue intensities to single modalities as well. Naturally, in case the tissue

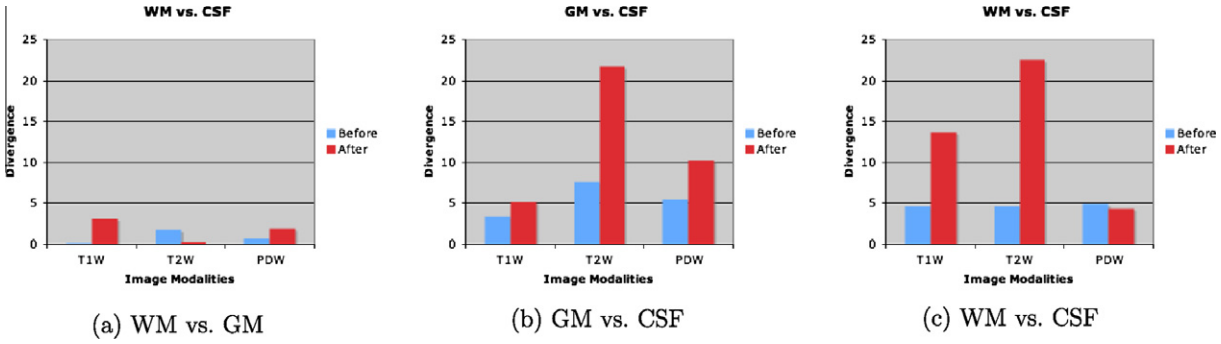


Fig. 6. Jeffreys Divergences between tissue types before (blue bar) and after (red bar) normalization. The other notations are the same as previous figures. Note that the divergence increases after normalization, indicating improved tissue separation for all cases (except WM:GM in T2w and WM:CSF in PDw). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

intensities follow a normal distribution in individual modalities then they can be depicted at least a mixture of Gaussians across the spectrum of modalities, a basis utilized by the approaches motivated by the Gaussian Mixture Model (GMM) based tissue analysis and segmentation (e.g., Schroeter et al., 1998; van Leemput et al., 2001; Dugas-Phocion et al., 2004).

We first compute, for the un-normalized MRI volumes, the necessary statistics of a Gaussian distribution from the data. We compute the mean vector μ_t^a and the variance $(\sigma_t^a)^2$ over the MRI intensity values for tissue type t where $t \in \mathscr{T}$ such that $\mathscr{T}$ is the set of all tissue types and $a \in \mathscr{A}$ such that $\mathscr{A}$ denotes the set of acquisition modalities. The set $\mathscr{T}$ considered in our case includes *CorticalGM*,*WM*, *DeepGM*,*CSF*, and *Lesion*. We can then generate samples of Gaussian distribution centered at μ_t^a and with a variance $(\sigma_t^a)^2$, i.e., $\mathscr{N}(\mu_t^a, (\sigma_t^a)^2),\ \forall t \in \mathscr{T},\ \forall a \in \mathscr{A}$.

Next, for each tissue type $t \in \mathscr{T}$, we generate histograms from the data over three modalities taking into account 98 data percentiles (upper and lower 1 percentile data left out as noise and outliers). This gives an account of the actual model of the data for the given tissue type.

Recall that the null hypothesis assumes that the data distribution obtained from the data histogram does not differ significantly from the Gaussian distribution over the sample mean and variance that is assumed to generate the samples. Hence, if we were to calculate a metric based upon how much do the distributions generated in these manner differ, we would not find any significant difference.

In order to verify this claim, we employ Jeffreys divergence measure to calculate the distance between these two distributions, both before and after normalization. Jeffreys divergence is a symmetric measure of similarity between two distributions. Low values of divergence represent less difference between the two distributions and hence would partially validate the null hypothesis claim.

In the above case of un-normalized data, if the Normal distribution $\mathscr{N}\left(\mu_t^a, (\sigma_t^a)^2\right)$ built using the sample statistics be denoted by $p(\cdot)$ and the data histogram be denoted by $q(\cdot)$ on intensities i of a given tissue type in any given image modality, then, the Jeffreys Divergence (Jeffreys, 1946) (JD) between $p(i)$ and $q(i)$ is defined as:

$$D(p,q) = \int_{-\infty}^{\infty} p(i) \log \frac{p(i)}{q(i)} di + \int_{-\infty}^{\infty} q(i) \log \frac{q(i)}{p(i)} di \quad (1)$$

Difference between any two distributions can thus be determined using the JD criterion. For instance, when we perform modeling in the similar manner as above on the normalized data, we can obtain a measure of difference between the actual histogram of normalized intensities (say $q(\cdot)$) against a Normal distribution obtained from (normalized) sample statistics (say $p(\cdot)$). We use the discretized approximation of $D(p,q)$ by replacing the integrals with the summations over a fixed number of bins. Please note that

Table 4
Effect of normalization on tissue separation and automatic clustering: *k*-means Classification error rates (in percentages, %) for various tissue types before and after normalization.

Tissue type	% Error before normalization	% Error after normalization
CSF	29.7	0.003
WM	57.4	1.15
GM	58.9	6.07
Lesion	44.8	13.18

in addition to considering various studies that claim that the tissue distribution can be approximated as Gaussian (van Leemput et al., 2001; Harmouche, 2006), we also performed goodness of fit tests with other well-known distributions such as Rician, Inverse-Gaussian, and Weibull. The results from these tests too suggested Gaussian distribution to be able to approximate the underlying data generating model most closely (even for intensities from images before normalization) for all tissue types with the exception of lesions whose distribution shape has not yet been established. We have nevertheless adopted the commonly employed Normality assumption on tissue types and attempt to better understand the effect of normalization of this distribution shape.

It should be noted here that Nyul and Udupa (1999c) used a metric called the Normalized Mean Squared Differences (NMSD) to assess the quality of normalization. The NMSD metric measures the mean squared intensity difference between the image before normalization and after normalization. However, the measurement does not take lesions into account as it is restricted to the original lesion-removed image. This metric is not sufficient in our case since we aim to test factors different from the ones generally analyzed when assessing the intensity normalization on any given image. Moreover, unlike Nyul and Udupa (1999c), we do not work with lesion-removed images since we are interested in understanding the normalization behavior in the presence of MS pathology.

The data statistics and the Jeffreys divergence values over all the volumes before and after intensity normalization are shown in Fig. 4. If the sample indeed comes from a Gaussian distribution then this difference would be small. Also, if intensity normalization does not play any role then the divergence value after normalization should not differ too much from the one before normalization. This analysis is then broken by scanner brands and the results are shown in Fig. 5.

6.3. Tissue divergences

We perform two analyses to show how normalization increases between-tissue discrimination. For the first part, we employ the Jeffreys divergence measure between the distributions of pairs of

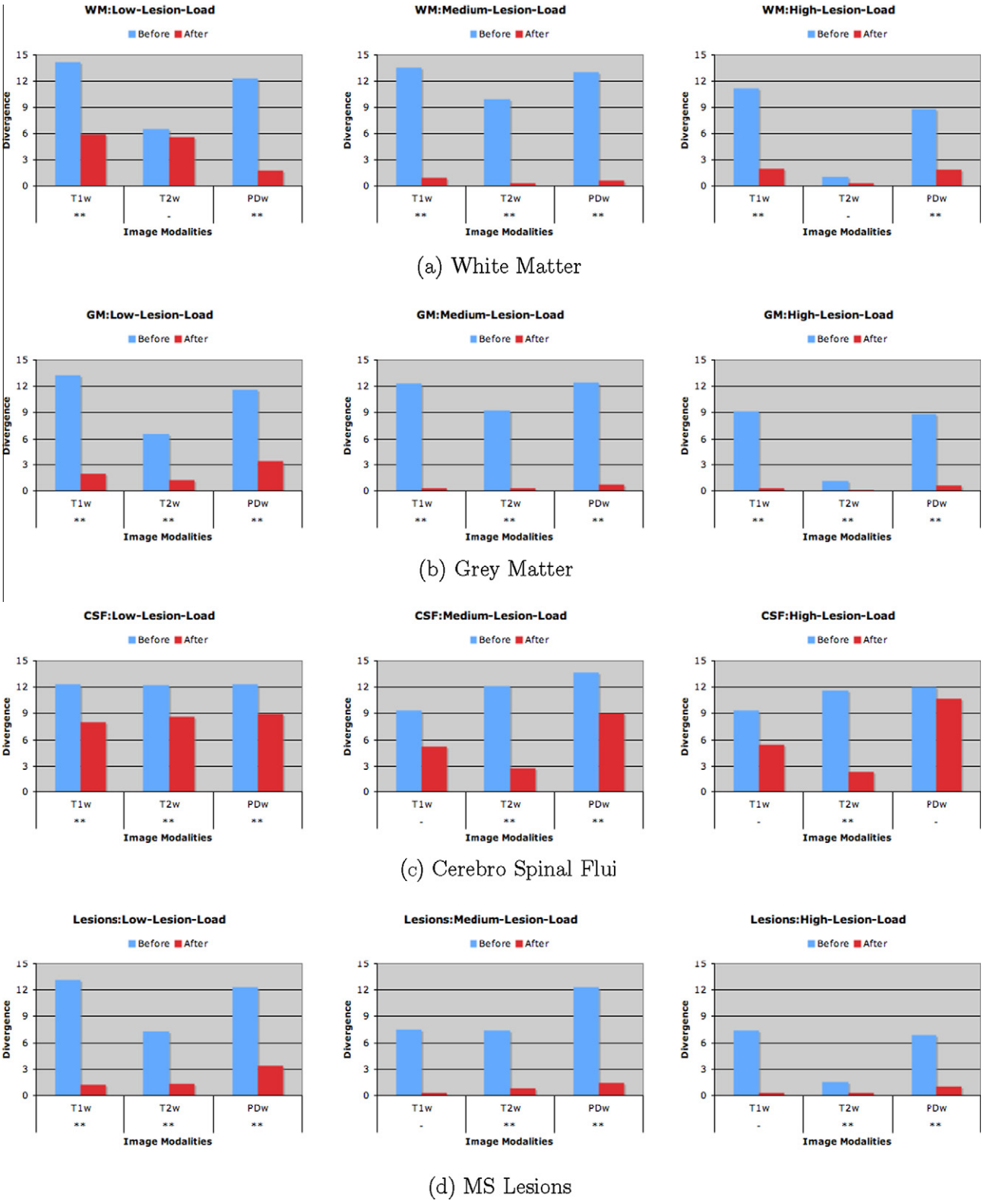


Fig. 7. Divergence measures between distribution fits before and after normalization. The blue bars in each graph denote, for a given modality and tissue type, the Jeffreys divergence measure between the Gaussian distribution centered at the sample mean with sample variance, and the histogram model generated from tissue intensities on the un-normalized volumes. The red bars correspond to the same measure over volumes after normalization. An absence of bar indicates that the divergence was too small to plot. Each row corresponds to measures for a given tissue type. Each column refers to divergence measures for MRI volumes with different lesion loads: low, moderate, and high. As before, a "$**$" beneath each comparison denote that it was found to be statistically significant at 95% level while "-" indicates otherwise. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tissue types before and after normalization within each modality. In order to do this, we obtain the histogram models over the voxel intensities for the pairs of tissue to be compared. For instance, for WM and CSF tissues, we obtain their histogram models first from the MRI volumes before normalization. We calculate the distance between these two distributions using the Jeffreys divergence

measure as described above. Next, we do the same for normalized images. If the intensity normalization improves the tissue constrast then we should see an increase in these divergence measures between the two distributions on the normalized images. For this analysis, we concern ourselves with the characteristics of primary tissue types, that is, the GM (combined Deep GM and Cortical GM), WM and CSF. Since lesions have a non-uniform presence across different volumes resulting in occasional sparse samples, we do not include the lesions in this analysis. Fig. 6 demonstrates that intensity normalization contributes to a better relative scale mapping of various tissue types resulting in improved tissue type separation in intensity space. Note that since the analysis was done on pure samples in a pooled manner (unlike above), the statistical significance tests applied above cannot be employed meaningfully in this case. We hence relegate this discussion to the next part where we directly evaluate the performance of supervised MS identification approaches.

The second part of our analysis with regard to the effect of intensity normalization focuses on the tissue intensity behaviors in multi-spectral intensity space. By examining the behavior, we perform unsupervised learning of the tissue clusters in different MRI brain volumes with the number of clusters set to 5 to reflect the number of different tissue type samples obtained in our sampling experiment described in Section 4. However, the Deep GM and Cortical GM are combined into one category. Effective normalization should give clusters that are more accurate (in terms of correctly mapping tissues to the corresponding tissue types based on their multi-spectral intensities) on brain volumes after normalization as compared to the un-normalized volumes.

An Expectation–Maximization based soft clustering algorithm was utilized to group the voxels corresponding to the same tissue type in k clusters. This method is referred to here as k-means approach. We present aggregate classification errors for different tissue types after the clustering algorithm has converged. Note again that this preliminary analysis was done on pure samples obtained manually. Hence, we do not present κ-measures for tissue classification accuracies here. Rather, our main aim is to demonstrate how intensity normalization brings tissues of one type closer on intensity scales and grouping them using k-means technique results in more uniform clusters, and hence, low classification error. These results are reported in Table 4. We will study the effectiveness of MS identification approaches soon.

6.4. Lesion load-dependent analysis

The 3D intensities for MS lesions are different from the rest of the tissue classes and the development of pathology may largely affect the shape of the histograms used to normalize the targeted MRIs, in terms of a decrease in white matter volumes and/or an increase of CSF volumes. Most of the current intensity normalization methods can cope well with subjects with healthy brains or low MS lesion volume loads, but the performance effect on subjects with severe pathology (in this case high MS lesion loads) is not well understood. In order to better study the effect of the lesion loads on the resulting intensity behaviors, we perform a lesion load-dependent analysis on the impact of the intensity normalization procedure.

With the help of the labels corrected by the experts, the MS lesion volumes with varying clinically defined lesion loads (0 cc to more than 25 cc) are evaluated and sorted for the 21 subjects. From the minimum to the maximum volumes, 33 and 66 percentiles of the MS lesion volume range are used as thresholds. These thresholds roughly corresponded to dividing the MRI volumes in groups with lesion load less than 5 cc, between 5 and 12 cc and more than 12 cc respectively. The 21 subjects are accordingly categorized into three groups, low, moderate, and high lesion load. The group with a high lesion load is sparse. Due to this limitation on the number of subjects with high MS lesion volumes, we limit our analysis to a total of nine subjects (three subjects for each group).

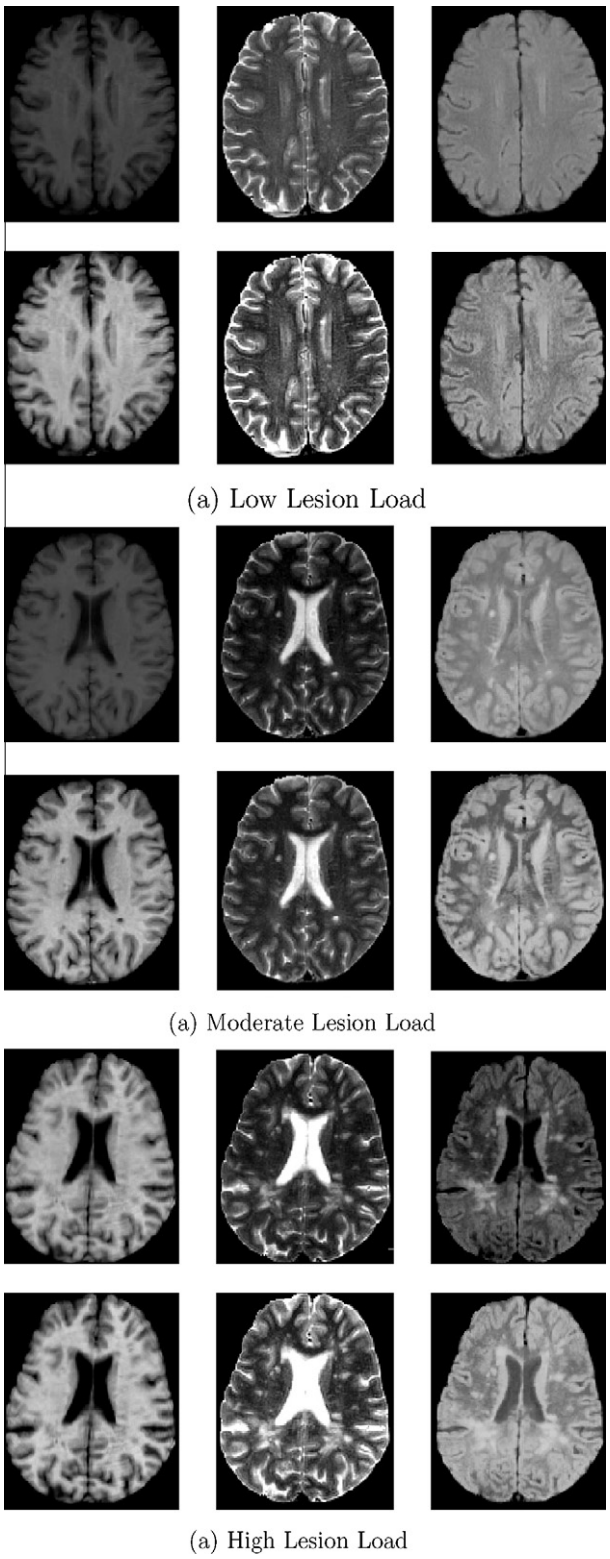
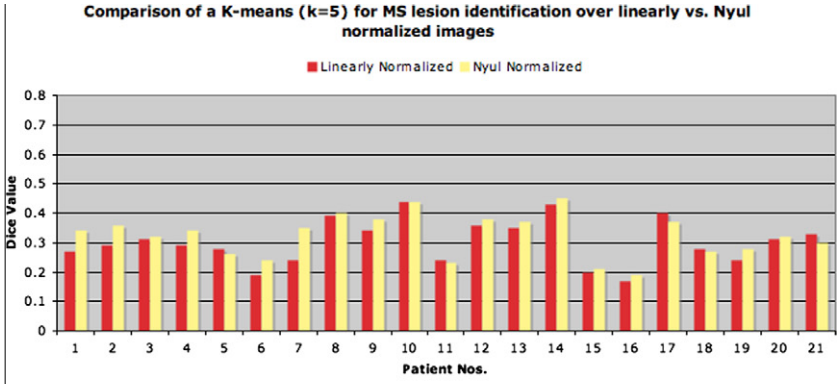
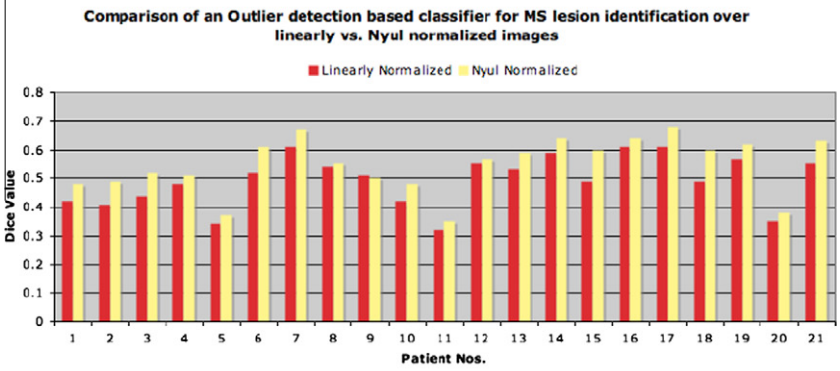


Fig. 8. Lesion load and intensity normalization. The first, third and fifth row give T1, T2 and Pd images before normalization for subjects with low, medium and high lesion loads respectively. The second, fourth and sixth rows give corresponding normalized images.

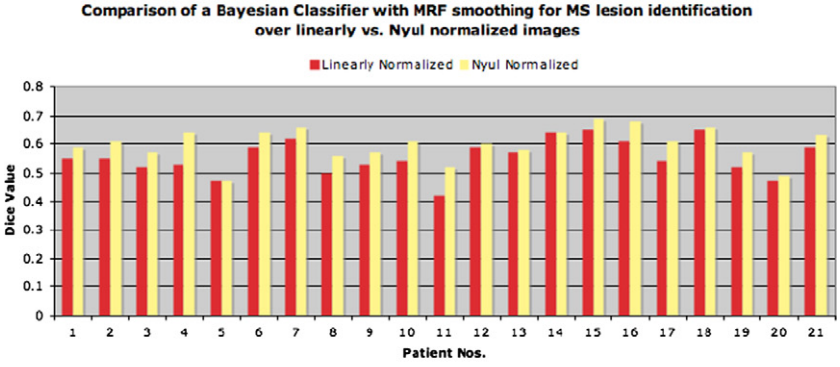
We perform a Jeffreys divergence based analysis on various tissue types for these groups with varying lesion loads. These results are presented in Fig. 7.



(a) K-means Classifier (p-value: 0.011)



(b) Outlier Detection Based Classifier of van Leemput et al. (2001) (p-value: < 0.0001)

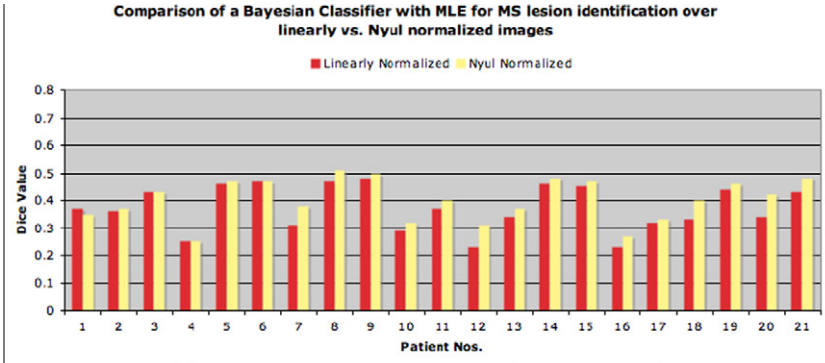


(c) Bayesian Classifier of (b) in conjunction with MRF Smoothing (p-value: < 0.0001)

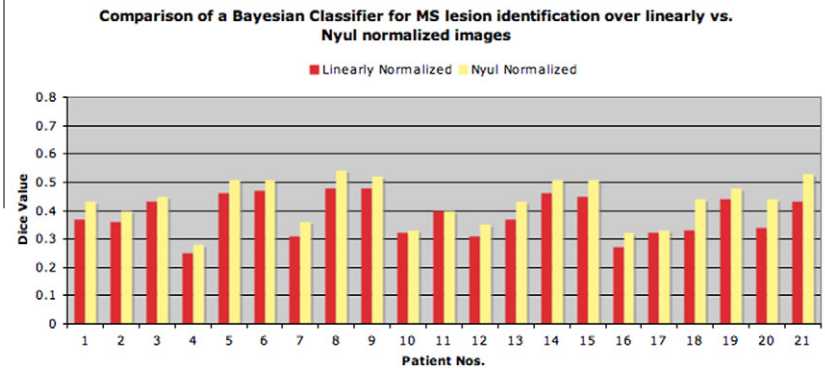
Fig. 9. Results comparing the effects of linear normalization (red bars) to the decile normalization (yellow bars) of Section 3 on various MS lesion identification approaches. Each bar shows the obtained value of the Dice coefficient against a consensus labeling of 5 experts (with a consensus threshold of 3). The *p*-value in the parentheses is associated with the statistical significance testing using a matched pair *t*-test over the mean Dice coefficient values. As can be seen, the results in the case of all the classifiers are found to be significant at or above 0.01 (99%) confidence level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.5. Effect on MS lesion identification

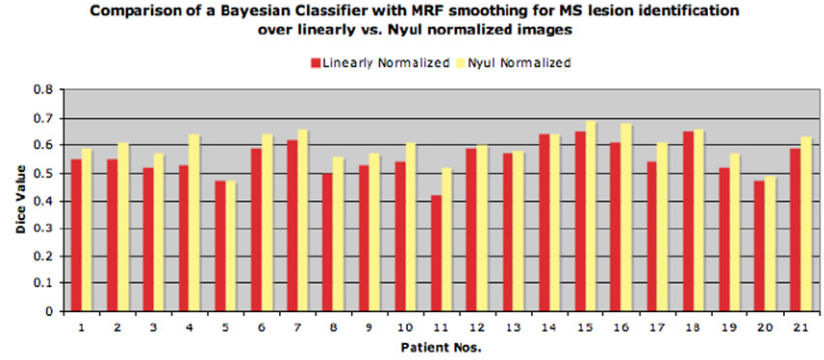
The ultimate test of the benefits of the intensity normalization on the supervised approaches to MS lesion identification is of course assessed in terms of the affect on their accuracy in identifying MS lesions from the multi-modal MRIs. Hence, our final set of experiments focuses on analysing this effect. We mainly analyze three approaches to MS lesion identification over multi-modal MRIs (T1w, T2w and PDw modalities). A Bayesian approach with Markov Random Field smoothing proposed by Harmouche (2006), an outlier detection based approach proposed by van Leemput et al. (2001) and also compare these with a *k*-means approach with EM optimization with k set to the number of classes of interest (in our case $k = 5$ for GM, WM, CSF, Lesions and Background). The results of Dice coefficient (Dice, 1945) of agreement over the identified MS lesions against a consensus labeling obtained from 5 expert raters on each image. A consensus threshold of 3 out of 5 raters for accepting a lesion label, as used in the experiments, are shown in Fig. 9. In each case, we compare the results of the classifier applied to images normalized according to the Nyul normalization described above against the images normalized according to a simple linear mapping in the same range as chosen by the Nyul method to study the effect of both the Nyul normalization as well as the deciles. We also performed statistical



(a) A Maximum Likelihood Estimator (p-value: 0.00011)



(b) Bayesian Classifier with Prior information included from an healthy atlas (p-value: < 0.0001)



(c) Bayesian Classifier of (b) in conjunction with MRF Smoothing (p-value: < 0.0001)

Fig. 10. Results comparing the effects of linear normalization (red bars) to the decile normalization (yellow bars) of Section 3 on various Bayesian tissue identification approaches with increasing sophistication. Each bar shows the obtained value of the Dice coefficient against a consensus labeling of 5 experts (with a consensus threshold of 3). The *p*-value in the parentheses is associated with the statistical significance testing using a matched pair *t*-test over the mean Dice coefficient values. As can be seen, the classifiers applied to images with the Nyul normalization of Section 3 outperforms the corresponding classifiers over linear normalization. These results, in the case of all the classifiers, are found to be significant at or above 0.01 (99%) confidence level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

significance tests using a paired *t*-test the *p*-values corresponding to which are also indicated in the parentheses in the legend of each figure.

Further, as seen in Fig. 10 the Bayesian classifier with MRF smoothing outperforms the other two classifiers. In order to study the effect of the level of sophistication of classifier against the normalization, we also studied a plain Maximum Likelihood estimator over the images with no prior information as well as a simple Bayesian classifier with prior information but no post-processing and compared these against the Bayesian classifier of Harmouche (2006) that uses an MRF to smooth the results of the Bayesian classifier. Note that the learning algorithm employed by Harmouche (2006) is a Bayesian classifier that obtains the prior information using an anatomical atlas of healthy brain tissues. It then performs Bayesian inference using a Maximum a posteriori estimate and subsequently utilizes an MRF to smoothen the output of the Bayesian classifier. In this respect, the three classifiers utilized in Fig. 10 reflect degrees of increasing sophistication. Similar to above, the *p*-values of a paired *t*-test for statistical significance are provided in the parentheses of each figure's legend.

7. Discussion

Our results demonstrate that intensity normalization plays a significant role in enabling both efficient tissue modeling and rel-

ative scale mapping in the context of heterogenous MRIs of human brains with pathology. As shown in Figs. 2 and 3, intensity normalization results in more homogenous intensity values for voxels of the same tissue type. The qualitative effects can be seen more clearly at the image level in the results of Fig. 8 where clearly the normalization results in images with improved contrasts in various modalities.

Over individual modalities, it results in data that are more representative of the underlying distribution under the Gaussian assumption. This is reflected in the reduced Jeffreys divergence measure over the distribution differences after normalization (see Fig. 4) over all data. It can also be seen in Fig. 5 that, overall, the same observations and conclusions hold when breaking this analysis down to groups of MRIs from scanners of particular manufacturers. These results were found to be statistically significant for most cases. However, there are a few isolated where the results were not found to be statistically significant. Consider, for instance, the cases of T2w modality in WM modeling in GE scanners, or all modalities over lesions in Philips scanners. The examination of the data suggests different reasons for this behavior. With regard to lesions, as mentioned above, the distribution behavior of tissue intensities are not well understood and hence such results are expected. Indeed, more analysis with more volumes is required to gain an in-depth understanding of behaviors of lesion intensities. The only other relevant instance where such difference is not found to be statistically significant happends to be over the WM and CSF modeling in T2w images of GE scanners. In all other cases, as can be seen, the JD values before normalization are indeed already relatively low. Hence, discerning a statistically significant effect over rescaling of data is extremely difficult. In the case of GE scanners, this behavior can in part be attributable to the input image intensity behavior even before normalization while in part to greater variability in the scanners used over GE. As a result, the data-clusters before normalization, even over individual modalities, are relatively widely dispersed as compared to other scanner brands. Another effect of the dispersed data-clusters in GE machines, as seen in Fig. 2c can then be seen the relatively high JD values in Fig. 5 over GE machines as compared to the other two brands. The effect of this behavior in GE is also reflected in the elevated values for the combined analysis over all scanners of Fig. 4. However, note that, in no case, does the normalization affect the intensities adversely with respect to the distribution assumptions made over them.

We also studied the effect of intensity normalization on tissue type separation both by comparing the tissue contrasts in each modality (see Fig. 6) and by studying the performance of a *k*-means clustering algorithm over the multi-spectral intensity space (Table 4). As can be seen clearly in Table 4, intensity normalization is instrumental in obtaining more accurate mapping of voxels to their respective tissue types based on their multi-spectral intensity profiles resulting in reduced clustering error on the normalized data. We would like to note the only discrepancy in the tissue contrast analysis between WM:GM in T2w scans as shown in Fig. 6. One possible reason can be the effect of the global 98% wide dynamic range mapping (from 1% to 99%) resulting in a range compression for WM and GM. While tissue separation is indeed of interest here, our main goal is the effectiveness of normalization in the context of MS lesion identification. Hence, instead of basing our conclusions on results over smaller "pure" samples, we do so over a more elaborate analysis of MS identification algorithms applied to the test volumes.

Before we discuss the segmentation results, however, we would like to note the effectiveness of the normalization across varying degrees of pathology as confirmed by results of Fig. 7. Note in particular that since the number of volumes considered in each category was limited as a result of sparseness of lesions in the high lesion load segment, the results over T1w images in the case of MS lesions (Fig. 7d) are prohibited from being taken to be statistically significant. However, what is relevant here is the general trend toward a reduced divergence making the data more suited to the application of various automated segmentation methods with regard to their assumptions over tissue intensities. Moreover, when this is not necessarily the case, it does not have any adverse effect on intensity distributions. At the qualitative level, a sense of the improved contrast in the normalized images is conveyed by Fig. 8.

Finally, we focus on the ultimate goal of the normalization exercise: more effective MS lesion identification. While for the methods that deal with pooled estimates to generalize the tissue intensity behaviors and subsequently using these in MS lesion identification, meaningful results are not obtained over un-normalized images for obvious reasons. Hence, we do not present results of applying these approaches on un-normalized data. In the respective studies as well as algorithm framework, it is amply clear that these approaches necessitate some type of rescaling of the data. In order to validate if indeed the method of Nyul et al. (2000) provides an advantage as compared to a simple linear rescaling of the intensities, we compare the results of the MS lesion identification approaches over these two normalization techniques. This, in effect, also gives an insight into the importance of piece-wise linear mapping approach adopted over deciles.

In the first set of experiments over MS identification, we compare the following approaches: a *k*-means learning algorithm with EM optimization over its parameters, the outlier based approach of van Leemput et al. (2001), and the Bayesian approach with Markov Random Field based smoothing of Harmouche (2006). We compare the obtained Dice agreement estimates against the ones obtained when the approaches are applied over linearly normalized image intensities. The linear normalization is a simple rescaling of the tissue intensities in the same range that is used for the decile normalization of Section 3. The results of Fig. 9 show that indeed the advantage afforded by the decile method is both apparent and statistically significant (see the associated *p*-values of a matched paired *t*-test). Next, we also wish to verify if the observed effects are indeed attributable to normalization and not to the difference in the level of sophistication of approaches. To do this, we perform another study over increasing sophistication of a Bayesian classifier while controlling for other factors. The results of Fig. 10 show the performance of MS identification approach based on a simple Maximum likelihood parameter estimation (with no prior information) over a Bayesian classifier (see Fig. 10a), a classical Bayesian classifier with prior information taken into account (Fig. 10b), and finally the classical Bayesian classifier followed by a MRF based smoothing of bayesian posteriors over class memberships of tissues (Fig. 10c). Again, as before, in each case, we compare the results against a simple linear normalization. The results again confirm the added advantage offered by the decile approach in both absolute (increased Dice values) as well as statistically significant sense (see the associated *p*-values) uniformly with increasing level of sophistication of the algorithm.

8. Conclusion

Tissue intensity behaviors in brain MRI volumes can vary significantly due to the variations in acquisition protocols, scanner differences, heterogeneity of source, and possibly due to intensity inhomogeneity correction applied to obtain uniform images. The presence of pathology in the brain can further aggravate this problem. As a result, intensity normalization plays a very important role in standardizing the tissue intensities for various tissue types across different brain MRI volumes. Supervised automated tissue

classification methods rely heavily on the intensity normalization. Various intensity normalization procedures have been proposed to address the issue of standardizing the tissue intensity range across MRIs. The method of Nyul and Udupa (1999b) and its variants have obtained a wide acceptance among these mainly as a result of the ease of application and speed of execution without the loss of accuracy (see, for instance, Moonis et al., 2002; Anbeek et al., 2004; Xue et al., 2004; Datta et al., 2006; Harmouche, 2006; Bergeest and Jäger, 2008; Shi et al., 2008; Khan et al., 2008; Karimaghaloo et al., 2010; Elliott et al., 2010). However, the effectiveness of this method on the MRIs with variations such as above and in the presence of MS had not been aptly explored.

In this work, we demonstrated the effectiveness of the decile formulation of this approach across a multi-site multi-scanner MRI data in the presence of varying MS lesion loads. We examined the effect of intensity normalization on the tissue intensity behavior in different modalities (T1, T2 and PD weighted). Utilizing the Jeffreys divergence criteria for measuring the difference in the distributions, we also verified whether the assumptions of an underlying Gaussian distribution on different tissue types made by various automatic tissue segmentation and visualization approaches hold. We also verified that the findings hold across varying lesions loads. Finally, we also studied the effect of the procedure on tissue separation among different tissue types over various MS lesion identification approaches and also verified that the advantages offered are indeed attributable to normalization by assessing them over a range of algorithms of varying sophistication while controlling for other factors. We believe that this extensive evaluation justifies utilization of the intensity normalization procedure of Nyul and Udupa (1999b) as a pre-processing step for image normalization for various segmentation and visualization techniques. An important observation that the study makes is that a relatively simple extension of the linear scaling in the form of decile based piece-wise linear normalization can prove to be a reasonable trade-off of minor computational complexity in favor of better discriminating ability of learning algorithms.

Acknowledgement

This work was supported by a Canadian National Science and Engineering Research Council Strategic Grant (STPGP 350547-07).

References

Anbeek, P., Vincken, K.L., van Osch, M.J.P., Bisschops, R.H.C., van der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage 21, 10371044.

Belaroussia, B., Millesb, J., Carmec, S., Zhua, Y.M., Benoit-Cattina, H., 2006. Intensity non-uniformity correction in MRI: existing methods and their validation. Medical Image Analysis 10 (2), 234–246.

Bergeest, J., Jäger, F., 2008. A comparison of five methods for signal intensity standardization in MRI. In: Tolxdorff, Braun, Deserno, Handels, Horsch, Meinzer, (Eds.), Bildverarbeitung für die Medizin (Bildverarbeitung für die Medizin 2008, Algorithmen, Systeme, Anwendungen, Proceedings des Workshops vom 6. bis 8., 2008). Springer, Berlin, pp. 36–40.

Bosc, M., Heitz, F., Armspach, J., Namer, I., Gounot, D., Rumbach, L., 2003. Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. NeuroImage 20, 643–656.

Christensen, J.D., 2003. Normalization of brain magnetic resonance images using histogram even-order derivative analysis. Magnetic Resonance Imaging 21 (7), 817–820.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3d intersubject registration of MR volumetric data in standardized talairch space. Journal of Computer Assisted Tomography 18, 192–205.

Cox, I.J., Roy, S., Hingorani, S.L., 1995. Dynamic histogram warping of image pairs for constant image brightness. In: Proceedings of International Conference on Image Processing, ICIP, pp. 2366–2369.

Datta, S., Sajja, B.R., He, R., Wolinsky, J.S., Gupta, R.K., Narayana, P.A., 2006. Segmentation and quantification of black holes in multiple sclerosis. NeuroImage 29, 467474.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dugas-Phocion, G., Gonzalez, M.A., Lebrun, C., Chanalet, S., Bensa, C., Malandain, G., Ayache, N., 2004. Hierarchical segmentation of multiple sclerosis lesions in multi-sequence MRI. In: ISBI04, vol. 1, pp. 157–160.

Elliott, C., Francis, S., Arnold, D., Collins, D.L., Arbel, T., 2010. Bayesian classification of multiple sclerosis lesions in longitudinal MRI using subtraction images. In: MICCAI-2010, Beijing, China, pp. 290–297.

Karimaghaloo, Z., Shah, M., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T., 2010. Detection of gad-enhancing lesions in multiple sclerosis using conditional random fields. In: MICCAI-2010, Beijing, China, pp. 41–48.

Khan, R., Wang, L., Beg, M.F., 2008. FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping. NeuroImage 41, 735–746.

Francis, S., 2004. Automatic Lesion Identification in MRI of Multiple Sclerosis Patients. MS thesis. Department of Neurology and Neurosurgery McGill University, Montréal.

Friston, K.J., Ashburner, J., Frith, C.D., Poline, J.B., Heather, J.D., Frackowiak, R.S.J., 1995. Spatial registration and normalization of images. Human Brain Mapping 2, 965–1189.

Gronenschild, Ed.H.B.M., Saartje, B., Floortje, S., Vuurman, E.F.P.M., Uylings, H.B.M., Jolles, J., 2010. A time-saving and facilitating approach for segmentation of anatomically defined cortical regions: MRI volumetry. Psychiatry Research: Neuroimaging 181 (3), 211–218.

Guimond, A., Roche, A., Ayache, N., Meunier, J., 2001. Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections. IEEE Transactions on Medical Imaging 20 (1), 58–69.

Harmouche, R., 2006. Bayesian MS Lesion Classification Modelling using Regional and Local Spatial Information. Master's thesis. McGill University.

He, R., Datta, S., Sajja, B.R., Narayana, P.A., 2008. Generalized fuzzy clustering for segmentation of multi-spectral magnetic resonance images. Computerized Medical Imaging and Graphics 32 (5), 353–366.

Hellier, P., 2003. Consistent intensity correction of MR images. In: Proceedings of International Conference on Image Processing, ICIP, vol. 1, pp. 1109–1112.

Jäger, F., Deuerling-Zheng, Y., Frericks, B., Wacker, F., Hornegger, J., 2006. A new method for MRI intensity standardization with application to lesion detection in the brain. In: Kobbelt, L., Kuhlen, T., Aach, T., Westermann, R. (Eds.), Vision Modeling and Visualization 2006. Aka GmbH, Aachen Berlin, pp. 269–276.

Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A 186, 453–461.

Madabhushi, A., Udupa, J.K., 2005. Interplay between intensity standardization and inhomogeneity correction in MR image processing. IEEE Transactions on Medical Imaging 24 (5), 561–576.

McAlpine, D., 1973. Multiple sclerosis: a review. British Medical Journal 2 (5861), 292–295.

Moonis, G., Liu, J., Udupa, J.K., Hackney, D.B., 2002. Estimation of tumor volume with fuzzy-connectedness segmentation of MR images. American Journal of Neuroradiology 23, 352–363.

Nyul, L.G., Udupa, J.K., 1999a. An approach to standardizing the MR intensity scale. In: Mun, S.K., Kim, Y. (Eds.), Medical Imaging 1999: Image Display, SPIE Proceedings, vol. 3658, pp. 595–603.

Nestares, O., Heeger, D.J., 2000. Robust multiresolution alignment of MRI brain volumes. Magnetic Resonance in Medicine 43, 705–715.

Nyul, L.G., Udupa, J.K., 1999b. On standardizing the MR image intensity scale. Magnetic Resonance in Medicine 42, 1072–1081.

Nyul, L.G., Udupa, J.K., 1999c. New variants of a method of MRI scale normalization. In: Kuba, A., Smal, M., Todd-Pokropek, A. (Eds.), Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI), Lecture Notes in Computer Science, vol. 1613. Springer Verlag, Heidelberg, pp. 490–495.

Nyul, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging 19 (2), 143–150.

Schroeter, P., Vesin, J.M., Langenberger, T., Meuli, R., 1998. Robust parameter estimation of intensity distributions for brain magnetic resonance images. IEEE Transactions on Medical Imaging 17 (2), 172–186.

Scully, M., Anderson, B., Lane, T., Gasparovic, C., Magnotta, V., Sibbitt, W., Roldan, C., Kikinis, R., Bockholt, H.J., 2010. An automated method for segmenting white matter lesions through multi-level morphometric feature classification with application to lupus. Frontiers in Human Neuroscience 4 (Article 27).

Shi, Y., Qi, F., Xue, Z., Chen, L., Ito, K., Matsuo, H., Shen, D., 2008. Segmenting lung fields in serial chest radiographs using both population-based and patient-specific shape statistics. IEEE Transactions on Medical Imaging 27 (4), 481–494.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A non-parametric method for automatic correction of intensity non-uniformity in MRI data. IEEE Transactions on Medical Imaging 17, 87–97.

van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A.C.F., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Transactions on Medical Imaging 20 (8), 677–688.

Weisenfeld, N.I., Warfield, S.K., 2004. Normalization of joint image-intensity statistics in MRI using the Kullback–Leibler divergence. In: IEEE Interational Symposium on Biomedical Imaging, vol. 1, pp. 101–104.

Wells III, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. IEEE Transactions on Medical Imaging 15, 429–442.

Xue, Z., Shen, D., Davatzikos, C., 2004. Determining correspondence in 3D MR brain images using attribute vectors as morphological signatures of voxels. IEEE Transactions on Medical Imaging 23 (10), 1276–1291.