SCI INSTITUTE
TECHNICAL REPORT

Adaptive, Nonparametric Markov Modeling for Unsupervised, MRI Brain-Tissue Classification

Suyash P. Awate, Tolga Tasdizen, Norman Foster, Ross T. Whitaker

UUSCI-2006-008

Scientific Computing and Imaging Institute
University of Utah
Salt Lake City, UT 84112 USA

January 23, 2006

Abstract:

This paper presents a novel method for brain-tissue classification in magnetic resonance (MR) images that relies on a very general, adaptive statistical model of image neighborhoods. The method models MR-tissue intensities as derived from stationary random fields. It models the associated higher-order statistics nonparametrically via a data-driven strategy. This paper describes the essential theoretical aspects underpinning adaptive, nonparametric Markov modeling and the theory behind the consistency of such a model. This general formulation enables the method to easily adapt to various kinds of MR images and the associated acquisition artifacts. It implicitly accounts for the intensity nonuniformity and performs equally well on T1-weighted MR data with or without nonuniformity correction. The method minimizes an information-theoretic metric on the probability density functions associated with image neighborhoods to produce an optimal classification. It automatically tunes its important internal parameters based on the information content of the data. Combined with an atlas-based initialization, it is completely automatic. Experiments on real, simulated, and multimodal data demonstrate the advantages of the method over the current state-of-the-art.

THE
UNIVERSITY
OF UTAH

Adaptive, Nonparametric Markov Modeling for Unsupervised, MRI Brain-Tissue Classification

Suyash P. Awate [a,*] Tolga Tasdizen [a] Norman Foster [b]
Ross T. Whitaker [a]

[a] *School of Computing, Scientific Computing and Imaging Institute, University of Utah, 50 S Central Campus Dr., Salt Lake City, UT 84112, USA*

[b] *School of Medicine, Center for Alzheimer's Care, Imaging and Research, University of Utah, 30 N 1900 E, Salt Lake City, UT 84132, USA*

Abstract

This paper presents a novel method for brain-tissue classification in magnetic resonance (MR) images that relies on a very general, adaptive statistical model of image neighborhoods. The method models MR-tissue intensities as derived from stationary random fields. It models the associated higher-order statistics nonparametrically via a data-driven strategy. This paper describes the essential theoretical aspects underpinning adaptive, nonparametric Markov modeling and the theory behind the consistency of such a model. This general formulation enables the method to easily adapt to various kinds of MR images and the associated acquisition artifacts. It implicitly accounts for the intensity nonuniformity and performs equally well on T1-weighted MR data with or without nonuniformity correction. The method minimizes an information-theoretic metric on the probability density functions associated with image neighborhoods to produce an optimal classification. It automatically tunes its important internal parameters based on the information content of the data. Combined with an atlas-based initialization, it is completely automatic. Experiments on real, simulated, and multimodal data demonstrate the advantages of the method over the current state-of-the-art.

Key words: Adaptive image modeling,
PACS: 87.57.Nk Data-driven brain-MRI classification,
1991 MSC: 62G07 Nonparametric density estimation,
1991 MSC: 62M40 Markov random fields,
1991 MSC: 94A15 Information theory.

1 Introduction

Tissue classification in magnetic resonance (MR) images of human brains is an important problem in biomedicine. The fundamental task in tissue classification is to classify the voxels in the volumetric (3-dimensional) MR data into gray-matter, white-matter, and cerebrospinal-fluid tissue types. This has numerous applications related to diagnosis, surgical planning, image-guided interventions, monitoring therapy, and clinical drug trials. Such applications include the study of neuro-degenerative disorders such as Alzheimer's disease, generation of patient-specific conductivity maps for EEG source localization, determination of cortical thickness and substructure volumes in Schizophrenia, and partial-volume correction for low-resolution image modalities such as positron emission tomography.

Manual classification of high-resolution 3D images is a tedious task, making it impractical for large amounts of data. Because of the complexity of this task, such classifications can be very error prone and exhibit nontrivial inter-expert and intra-expert variability [1]. *Fully automatic* or *unsupervised* methods, on the other hand, virtually eliminate the need for manual interaction. Therefore, such methods for brain-tissue classification have received significant attention in the biomedical image processing domain.

Recent developments in automatic brain-tissue classification have led to state-of-the-art systems that typically incorporate the following strategies: (a) parametric statistical models, e.g. Gaussian, of voxel grayscale intensity for each tissue class, (b) Markov-random-field models to enforce spatial smoothness on the classification, (c) methods to correct for the nonuniformity that is inherent to MRI, and (d) using probabilistic-brain-atlas information in the classification method. However, several biomedical factors continue to pose significant challenges to the state of the art, such as:

(1) The intensities and contrast in MR images varies significantly with the pulse sequence, and several other variable scanner parameters. The quality of MR data also shows a certain amount of variation when produced at multiple sites with different MR scanners.
(2) Magnetic resonance imaging (MRI) acquisition artifacts, intensity nonuniformity (bias field), the Rician nature of the noise in magnitude-MR data [2], and partial voluming effects [3] can cause the data to significantly deviate from the Gaussian models, thereby compromising the quality of the

[*] Corresponding Author.
Email addresses: suyash@cs.utah.edu (Suyash P. Awate), tolga@cs.utah.edu (Tolga Tasdizen), Norman.Foster@hsc.utah.edu (Norman Foster), whitaker@cs.utah.edu (Ross T. Whitaker).
URL: http://www.cs.utah.edu/∼suyash (Suyash P. Awate).

classification.

(3) Most methods treat the nonuniformity as multiplicative noise and explicitly correct the MR intensities to reduce its effect. However, for certain kinds of data, e.g. neonatal brain MRI, the nonuniformity correction can pose a serious challenge because of higher intensity variability for each tissue class as well as lower intensity contrast [4].

To address these issues in an effective way, unsupervised classification approaches need to *adapt* to the data. One adaptation strategy is to automatically *learn* the underlying image statistics from the data and construct a classification strategy based on that model.

This paper presents a novel method for MRI brain-tissue classification that incorporates an adaptive nonparametric model of neighborhood statistics. It is a more complete version of some of our previous work [5]. The strategy is to learn the image-neighborhood/Markov statistics from the input data itself using nonparametric density estimation. We show that this approach enables the method to perform well *without* any explicit nonuniformity correction. Incorporating the information content in the neighborhoods in the classification process virtually eliminates the need for explicit smoothness constraints on the classification, and provides optimal regularization. The method produces an optimal classification by minimizing an entropy-based metric defined on the *higher-order* Markov probability density functions (PDFs). It adjusts all its important internal parameters automatically using a data-driven approach and information-theoretic metrics. Combined with an atlas-based initialization, it is fully automatic. It incorporates the atlas information in the classification process in a straightforward manner. Experiments on real, simulated, and multimodal data demonstrate the significant advantages of the method over the current state-of-the-art.

The rest of the paper is organized as follows. Section 2 discusses works in MR-image classification and nonparametric Markov modeling and their relationship to the proposed method. Section 3 presents the mathematical basis of the proposed method, which relies on an adaptive Markov-random-field image model. Section 4 formulates the classification as an optimal-segmentation problem associated with an information-theoretic goodness measure on higher-order image statistics. Section 5 focuses on the application of the proposed method to brain-tissue classification. It explains why the method does not require explicit nonuniformity correction, and describes the usage of the atlases during initialization and classification. Section 6 gives the validation results and analysis on numerous real and simulated images. Section 7 summarizes the contributions of the paper and presents ideas for further exploration.

2 Related Work

This section discusses works in MR-image classification and nonparametric Markov modeling along with their relationships to the proposed method. It compares and contrasts the proposed strategy, in brief, with the key ideas around which the classification strategies have evolved such as: (a) decision based on voxel grayscale intensity, (b) use of regularization schemes based on local interactions among voxels in the classification, and (c) incorporating spatial priors based on probabilistic atlases.

Early approaches for tissue classification typically used segmentation-based methods that did not explicitly account for the effect of noise in the data. Such approaches used image-denoising methods in a pre-processing step, e.g. Gerig et al. [6] use a non-linear diffusion technique. Current strategies incorporate more effective schemes that perform classification, without such pre-processing, while dealing with the noise and nonuniformity in the data.

Wells et al. [7] present a method that couples tissue classification with nonuniformity correction based on maximum-likelihood parameter estimation. They use the expectation-maximization (EM) algorithm of Dempster et al. [8] to simultaneously estimate the unknown bias field and the classification. Leemput et al. [9,3] extend this approach by posing the problem in the context of mixture density estimation to estimate the grayscale intensity PDFs for each tissue type. They apply the EM algorithm to estimate these PDFs as well as the bias and, in turn, the classification. Their approach assumes that the tissue-intensity distribution conforms to a parametric Gaussian PDF whose parameters are obtained via the EM algorithm.

The EM-classification algorithm does not impose any smoothness constraint on the classification and, therefore, it is susceptible to outliers in the tissue intensities. Several authors [10,11,9,3,12,13] have extended the EM-classification algorithm to incorporate spatial smoothness via Gibbs/Markov priors on the label image. For instance, Kapur et al. [10] use spatially-stationary Gibbs priors to model local interactions between neighboring labels. Typically, these methods modify single-voxel tissue-probabilities based on energies defined on local configurations of classification labels. They assign lower energies to spatially-smooth segmentations and, therefore, make them more likely. However, such strong Markov models can over regularize the fine structured interfaces, e.g. the one between gray matter and white matter. Hence, it is often necessary to impose additional heuristic constraints [11,9,3]. Ruf et al. [14] extend the EM approach to perform spatial regularization by incorporating the spatial coordinates of the voxels, in addition to their grayscale intensities, in the feature vector. They model each tissue class *spatially* by a *constrained* Gaussian-mixture model on the coupled feature space. Identical to previous

EM-based approaches, for each class, they constrain every Gaussian to have exactly the same mean in the intensity subspace.

Researchers have also used active contour models [15,16] to impose smoothness constraints for segmentation. These methods typically attempt to minimize the area of the segmentation boundary, an approach that can over regularize interfaces. Furthermore, the results obtained using these models are typically quite sensitive to some internal contour parameters. Hence, these methods typically entail careful manual parameter-tuning.

Classification techniques based on anatomical atlases have been widely studied in the literature [17–19]. These techniques convert the classification problem into a deformable-registration problem between the MR-image and the anatomical brain atlas. Once the registration is done, the method uses the resulting transformation to map the anatomical structure from the atlas onto the data to produce a segmentation. However, such methods rely heavily on the availability and accuracy of the anatomical atlas as well as the quality of the deformable registration.

The proposed method, in contrast to typical EM-based strategies, does not impose any parametric model on the tissue intensities. Instead, it automatically adapts to a model well-suited to the data via a data-driven nonparametric density estimation scheme. It exploits Markovity for regularization, but it applies the Markov model to the intensity data without any explicit smoothness constraints on the classification. Learning the Markov model from the input data itself, it provides optimal regularization for the segmentation. Moreover, this approach enables the method to perform well without any explicit bias correction. It does not use anatomical atlases but rather incorporates prior information via probabilistic atlases for the initialization and during the optimization.

Learning higher-order Markov statistics nonparametrically entails estimation of PDFs in high-dimensional spaces. For instance, for a first-order local neighborhood having 6 voxels, i.e. 2 neighbors along each cardinal axis, we need to estimate PDFs on a 7-dimensional space (center voxel along with its neighbors). Researchers analyzing the statistics of natural images in terms of local neighborhoods draw conclusions that are consistent with Markov image models. Lee et al. [20] as well as Silva and Carlsson [21] analyze the statistics of 3-pixel × 3-pixel neighborhoods in images, in the corresponding high-dimensional spaces, and find the data to be concentrated in clusters and low-dimensional manifolds exhibiting nontrivial topologies.

The literature presents some examples of algorithms that learn the statistics of image neighborhoods. Popat and Picard [22] were among the first to use nonparametric Markov sampling in images. Their approach models the higher-order nonlinear image statistics via cluster-based nonparametric density estimation. They apply it to image restoration, image compression, and texture classification. However, their approach relies on a training sample, which limits its practical use—the proposed method learns the Markov statistics of the signal directly from the input data.

The method in this paper and [5] builds on our previous work in [23,24], which lays down the building blocks for unsupervised learning of higher-order image statistics and proposes entropy reduction on higher-order statistics for *restoring* generic gray scale images. This paper describes the essential theoretical aspects underpinning adaptive, nonparametric Markov modeling and the theory behind the consistency of such a model. It also provides a different perspective towards the optimal choice of parameters in the associated nonparametric density estimation.

The literature dealing with texture synthesis also sheds some light on the principles underlying the proposed method. Texture-synthesis algorithms rely on image statistics from an input image to construct novel images that bear a qualitative resemblance to the input [25–27]. Although this is a different application, and these algorithms do not rely on information-theoretic formulations, they demonstrate the power of neighborhood statistics in capturing essential aspects of image structure. For some of these applications, Levina [28] proves that the empirically-learned Markov statistics converge asymptotically to the true texture statistics. This proof of convergence is also applicable towards the nonparametric learning of the Markov statistics in the proposed method.

3 Adaptive Image Modeling via Nonparametric Markov Random Fields

This section presents the statistical foundation of the proposed segmentation-based classification method, which relies on an adaptive Markov-random-field image model.

A random field [29] is a family of random variables $X(\Omega; T)$, for some index set T, where, for each fixed $T = t$, the random variable $X(\Omega; t)$ is defined on the sample-space Ω. If we let T be a set of points defined on a discrete Cartesian grid and fix $\Omega = \omega$, we have a realization of the random field called the *digital image*, $X(\omega, T)$. In this case T is the set of voxels in the image. For two-dimensional images t is a two-vector. We denote a specific realization $X(\omega; t)$ (the intensity at voxel t), as a deterministic function $x(t)$.

For the formulation in this paper, we assume X to be a Markov random field. This implies that the conditional PDF of a random variable $X(t)$ at voxel t given all other voxel intensities is exactly the same as the conditional PDF conditioned on only the voxel intensities in the *neighborhood* or spatial proximity of voxel t. Essentially, this enforces the concept of local statistical dependence of voxel intensities during image formation. A formal definition of Markovity relies on the notion of a neighborhood, which we define next.

If we associate with T a family of voxel neighborhoods $N = \{N_t\}_{t \in T}$ such that $N_t \subset T$, $t \notin N_t$, and $u \in N_t$ if and only if $t \in N_u$, then N is called a neighborhood system for the set T. Voxels in N_t are called neighbors of voxel t. We define a random vector $Y(t) = \{X(t)\}_{t \in N_t}$ corresponding to the set of intensities at the neighbors of voxel t. We also define a random vector $Z(t) = (X(t), Y(t))$ to denote image regions, i.e. voxels coupled with their neighborhoods. For notational simplicity we use the short hand for random variables $X(t)$ as X and their realizations $x(t)$ as x, dropping the index t. Based on this general notion of a neighborhood, Markovity implies that

$$P\left(X(t) | \{X(s) = x(s)\}_{s \in T \setminus \{t\}}\right) = P\left(X(t) | Y(t) = y(t)\right). \tag{1}$$

3.1 Unsupervised Learning of Higher-Order Markov Statistics

The proposed method exploits the Markovity property in images. However, we know neither the functional forms nor the parameter values for the joint or conditional Markov PDFs. MR images obtained with varying MRI-parameter values, e.g. T1, T2, and PD, or varying noise/bias fields belong to different Markov models. For a segmentation method to be easily applicable in all such cases we need an adaptive Markov model that we learn from the input data itself. To achieve this goal, we model images using a piecewise-stationary (or piecewise-homogenous) Markov model, and then employ a data-driven nonparametric-density-estimation technique to estimate the unknown Markov PDFs.

A stationary region is one where the higher-order Markov PDFs are exactly the same for all voxels in that region [29]. This is an instance of the shift-invariance property. For brain MR images, the Markov statistics in individual parts of the brain, such as white matter or gray matter, are similar for all voxels and, hence, the piecewise-stationary model holds. Indeed, the successful high-quality classifications produced by the proposed method corroborate this claim. Exploiting stationarity, where we have many observed voxel/neighborhood intensities derived from a single higher-order PDF, we can employ a nonparametric Parzen-window density estimation scheme to estimate the higher-order PDF.

This approach has significant advantages. The proposed method automatically *learns* the Markov model from the data and constructs a segmentation strategy based on that model, making it adaptive. As discussed later in Section 6, the method performs well for low noise levels where some state-of-the-art techniques using parametric intensity models break down due to partial-voluming effects. That section also shows that incorporating neighborhoods in the segmentation strategy inherently provides optimal regularization, thereby eliminating the need for explicit regularization. Moreover, this approach enables the method to perform well without any explicit bias correction.

3.2 Nonparametric Parzen-Window Density estimation

The estimation of higher-order Markov PDFs introduces the challenge of high-dimensional, scattered-data interpolation, even for modest-sized image neighborhoods (7 dimensional space in this paper). High-dimensional spaces are notoriously challenging for data analysis (regarded as *the curse of dimensionality* [30,31]) because they are so sparsely populated. Despite theoretical arguments suggesting that density estimation beyond a few dimensions is impractical, the empirical evidence from the literature is more optimistic [31,22]. The results in this paper confirm that observation. Furthermore, stationarity implies that the random vector $Z = (X, Y)$ has identical marginal PDFs, thus lending itself to more accurate density estimates [31,30]. The proposed method also relies on the neighborhoods in natural images having a lower-dimensional topology in the multi-dimensional feature space [20,21]. Therefore, in the feature space, locally, the PDFs of images are lower dimensional entities that lend themselves to better density estimation.

We use the Parzen-window nonparametric density estimation technique [32,33] with an n-dimensional Gaussian kernel $G_n(z, \Psi_n)$, where Ψ_n is the covariance matrix. Having no *a priori* information on the structure of the PDFs, we choose an isotropic Gaussian, i.e. $\Psi_n{=}\sigma I_n$, where I_n is the $n \times n$ identity matrix. For a stationary process, the Parzen-window estimate is

$$P(Z = z(t)) \approx \frac{1}{|A_t|} \sum_{u \in A_t} G_n(z(t) - z(u), \Psi_n), \tag{2}$$

where the set A_t is a small subset of T chosen at random for each voxel t from voxels in the spatial proximity of t. We refer to this sampling strategy as *local sampling* (more details later in Section 5.5.1). The local sampling strategy enables the proposed method to learn the higher-order Markov PDF in the stationary region. In Section 6.1, we show that this sampling strategy enables the proposed method to perform equally well on both biased and unbiased MR-data without explicit bias correction. The random selection results in a stochastic approximation for the PDFs that alleviates the effects of spurious local maxima introduced in the finite-sample Parzen-window density estimate [34].

The Parzen-window estimation technique possesses good convergence properties. For an infinite sample size, the Parzen-window density estimate converges to the true PDF irrespective of the value of the kernel parameter σ [32,33] . This suggests that with increasing data we get progressively better density estimates. However, with a finite sample size, which is precisely the case in real life, the density estimate varies with the kernel parameter value. In this case, using optimal values of the Parzen-window parameters is critical for success, and manually tuning the value can be difficult in high-dimensional spaces. Section 3.3.1 describes a data-driven technique to estimate an optimal kernel parameter value.

3.3 Markov Consistency for the Adaptive Model

The convenience of the Markovity property on the random field comes with additional constraints. Besag was among the pioneers who analyzed the stochastic models for systems of spatially-interacting random variables. Via his proof of the Hammersely-Clifford theorem [35], also known as the Markov-Gibbs equivalence theorem, Besag showed that the conditional Markov PDFs $P(X(t)|Y(t))$ must be restricted to a specific form in order to give a *consistent* structure to the entire system. A consistent system is one where we can obtain each conditional PDF, using rules of probabilistic inference, from the joint PDF $P(\{X(t)\}_{t\in T})$ of all the random variables in the system.

The higher-order Markov PDFs that the proposed method learns empirically from the data do, indeed, yield a consistent system asymptotically i.e. as the amount of data tends to infinity. The previous section explained that the Parzen-window density estimate gives the true PDF estimate as the sample size becomes infinite. This convergence holds when the *observations* in the sample are independently generated from a single underlying PDF. The observations, in our case, are the neighborhood-intensity vectors. The stationarity of the Markov random field implies that all observations are derived from a single PDF. Using a subset U of the entire voxel-set T, such that no two voxels in the subset have overlapping neighborhoods ($\forall a, b \in U : N_a \cap N_b = \phi$), produces independent observations ($\{z(u)\}_{u\in U}$) and such a scheme achieves consistency asymptotically. However, the constraint of non-overlapping neighborhoods leads to a wastage of a large amount of data ($\{z(t)\}_{t\in T\setminus U}$). It turns out that we can indeed use the data at all the available (overlapping) neighborhoods ($\{z(t)\}_{t\in T}$) and yet converge to the true PDF asymptotically, as proved by Levina [28].

3.3.1 Optimal Parzen-Window Kernel Parameter

Besag analyzed the problem for estimating parameters of PDFs for a *parametric* Markov random field [36] that also concerned the use of overlapping neighborhoods described in the previous section. A maximum-likelihood estimate of the parameter requires independent observations, thereby implying non-overlapping neighborhoods. However, Besag suggested a maximum pseudo-likelihood scheme that used overlapping neighborhoods giving better estimates [36]. Geman and Graffigne [37] later proved that the maximum-pseudo-likelihood parameter estimate did indeed converge to the true maximum-likelihood estimate asymptotically. We use a similar strategy to estimate the optimal Parzen-window kernel parameter, i.e. we choose the maximum pseudo-likelihood value of the Gaussian standard-deviation, σ, kernel parameter. We observe that this maximum-likelihood choice is equivalent to the choice that minimizes the entropy of the higher-order Markov statistics for the entire image. The optimal σ is

$$\underset{\sigma}{\operatorname{argmax}} \prod_{t \in T} P(z(t)) = \underset{\sigma}{\operatorname{argmax}} \sum_{t \in T} \log P(z(t)) = \underset{\sigma}{\operatorname{argmin}} \left(- \sum_{t \in T} \log P(z(t)) \right) \quad (3)$$

With this maximum-likelihood strategy, treating the entire Markov random field as a single stationary and ergodic random field, the optimal σ is an approximation to

$$\underset{\sigma}{\operatorname{argmin}} \sum_{z \sim P(Z)} (-\log P(z)) = \underset{\sigma}{\operatorname{argmin}}\, E_{P(Z)} \left[-\log P(Z) \right] = \underset{\sigma}{\operatorname{argmin}}\, h(Z), \quad (4)$$

where $z \sim P(Z)$ means that z is randomly generated from the PDF $P(Z)$, and $h(Z)$ is the entropy of the random variable Z. Indeed, the relationship between log-likelihood and entropy has been well known and has been utilized in some other works, such as [34].

4 Optimal Segmentation via Mutual-Information Maximization on Higher-Order Statistics

This section formulates the classification problem as an optimal-segmentation problem associated with an information-theoretic goodness measure. It begins by forming a connection between information-theoretic measures, such as mutual information and entropy, and optimal segmentation.

Mutual information between two random variables measures the mutual dependence between them [38]. Independent random variables convey no information about each other, and so their mutual information is zero (minimal). Functionally dependent random variables, on the other hand, contain all the information about each other. That is, knowing the value of one random variable tells us exactly the value of the other random variable. These ideas apply well to the problem of image segmentation [39]. For a good segmentation, knowing the voxel neighborhood uniquely tells us the voxel class. Also, knowing the voxel class gives us a good indication of what the voxel neighborhood is.

Consider a random variable $L(t)$ associated with each voxel $t \in T$. Then $L(t)$ gives the class that voxel t belongs to. We define the optimal segmentation as the one that maximizes the mutual information between L and Z, i.e.

$$I(L, Z) = h(Z) - h(Z|L) = h(Z) - \sum_{k=1}^{K} P(L = k) h(Z|L = k). \tag{5}$$

The entropy of the higher-order PDF associated with the entire image, $h(Z)$,, is a constant for an image and we can ignore it during the optimization. Let $\{T_k\}_{k=1}^{K}$ denote a mutually-exclusive and exhaustive decomposition of the image domain T into K regions such that $T_k = \{t \in T : L(t) = k\}$. Then, for a region T_k, (5) treats the entropy of the higher-order statistics in that region to quantify its goodness measure.

Entropy is a measure of randomness or uncertainty associated with a PDF [38]. Regions having low entropies for higher-order PDFs possess a strong homogeneity in their higher-order statistics. Such homogeneity is characteristic of regions that comprise voxels of a single tissue type, e.g. white matter or gray matter. On the other hand, regions comprising voxels of multiple tissue types exhibit decreased regularity leading to increased randomness and, in turn, higher entropy. Letting $P_k(Z(t) = z(t))$ be the probability of observing the voxel neighborhood $z(t)$ given that the voxel t belongs to the region k, i.e.

$$P_k(Z(t) = z(t)) \approx \frac{1}{|A_t|} \sum_{u \in A_t, A_t \subset T_k} G_n(z(t) - z(u), \Psi_n), \tag{6}$$

where the set A_t is a small subset of T_k chosen at random for each voxel t from voxels in the spatial proximity of t. The entropy, in turn, is

$$h(Z|L = k) = -\int_{\Re^d} P_k(Z(t_k) = z(t_k)) \log P_k(Z(t_k) = z(t_k)) dz, \tag{7}$$

where $d = |N_t| + 1$ is the neighborhood size, and t_k is any voxel in region k—$\forall t_k \in T_k$, the PDF $P_k(\cdot)$ remains the same due to the stationarity assumption.

Equation (5) implies that the optimal segmentation is the one that minimizes a weighted average of entropies associated with the set of K PDFs. With the present mutual-information-based energy, reducing entropies of larger regions in the image is given more importance or weight in direct proportion to their size—the weights are the probability of occurrence of the classes $P(L = k)$ in the given image. Equations (5) and (7) give the optimal segmentation as

$$\{T_k^*\}_{k=1}^K = \underset{\{T_k\}_{k=1}^K}{\operatorname{argmin}} \left(\sum_{k=1}^K P(L = k) h(Z|L = k) \right) \tag{8}$$

$$= \underset{\{T_k\}_{k=1}^K}{\operatorname{argmin}} \left(-\sum_{k=1}^K P(L = k) \int_{\Re^d} P_k(Z(t_k) = z(t_k)) \log P_k(Z(t_k) = z(t_k)) dz \right). \tag{9}$$

Treating entropy as the expectation of negative log-probability and approximating the expectation, in turn, by the sample mean [38], we get

$$\{T_k^*\}_{k=1}^K = \underset{\{T_k\}_{k=1}^K}{\operatorname{argmin}} \left(-\sum_{k=1}^K P(L = k) E_{P_k(Z)} \left[\log P_k(Z(t_k)) \right] \right) \tag{10}$$

$$= \underset{\{T_k\}_{k=1}^K}{\operatorname{argmin}} \left(-\sum_{k=1}^K P(L = k) \frac{1}{|S_k|} \sum_{z \sim P_k(Z(t_k)), z \in S_k} \log P_k(z) \right), \tag{11}$$

where $|S_k|$ equals the size of the k-th sample S_k, and $z \sim P_k(Z(t_k))$ means that z is randomly generated from the PDF $P_k(Z(t_k))$.

Assuming ergodicity [29], in addition to stationarity, enables us to approximate ensemble averages with spatial averages. Hence we have

$$\{T_k^*\}_{k=1}^K \approx \underset{\{T_k\}_{k=1}^K}{\operatorname{argmin}} \left(-\sum_{k=1}^K P(L = k) \frac{1}{|T_k|} \sum_{t \in T_k} \log P_k(z(t_k)) \right), \tag{12}$$

where $|\cdot|$ is an operator giving the cardinality of sets. Taking $P(L = k) = |T_k|/|T|$ gives

$$\{T_k^*\}_{k=1}^K \approx \underset{\{T_k\}_{k=1}^K}{\operatorname{argmin}} \left(\frac{-1}{|T|} \sum_{k=1}^K \sum_{t \in T_k} \log P_k(z(t_k)) \right). \tag{13}$$

To minimize the energy we manipulate the regions T_k using a gradient-descent optimization strategy. Section 5.1 gives the details.

5 MR-Image Brain Tissue Classification

We now focus on classifying brain-MR images. Our goal is to segment the image into 4 regions corresponding to the (a) white matter, (b) gray matter, (c) cerebrospinal fluid, and (d) every other tissue type. This section starts by giving a high-level version of the proposed classification algorithm. It then explains why the method performs well without an explicit bias correction mechanism. It also describes a few ways of incorporating the information in the probabilistic atlases into the proposed method.

5.1 High-Level Classification Algorithm

Based on the PDFs estimated using an initial classification, we need to decide which PDF/region each voxel should belong to. We can see that the energy in (13) can be reduced if each voxel t is assigned to the class k that maximizes the probability $P_k(z(t))$. This is an iterative process where the PDFs define a classification that, in turn, redefines the PDFs. Because the PDFs get implicitly redefined after every iteration, via the updated classification, the PDF estimates *lag*, so to speak, behind the classification. This is a widely used standard numerical optimization scheme. (Some recent work [40] focuses on avoiding the lag in the PDFs by introducing additional terms, but is not the focus of this paper.)

Given a classification $\{T_k^m = \{t \in T : L^m(t) = k\}\}_{k=1}^K$ at iteration m, the algorithm iterates as follows:

(1) $\forall k, t$: Estimate $P_k^m(z(t))$ nonparametrically, as described in Section 3.2.
(2) $\forall t : L^{m+1}(t) = \text{argmax}_k \, P_k^m(z(t))$.
(3) Stop upon convergence, i.e. when $|T_k^{m+1} - T_k^m|_2 < \delta, \forall k$, where δ is a small threshold.

5.2 Implicit Bias-Field Handling

Typical MR images are characterized by a slowly-varying multiplicative field that is known as the bias field. This bias field is inherent to MRI and is caused by equipment limitations and patient-induced electrodynamic interactions [9]. It can lead to significant performance degradation of intensity-based classification techniques. Most methods treat the bias field as multiplicative noise and explicitly correct the intensities, e.g. as a pre-processing step or iteratively during the segmentation [9], to reduce the effect of the bias field. However, for some applications, such as neonatal brain MRI, the bias correction can pose a serious challenge [4]. This is because such images, as compared to adult MR images, exhibit higher intensity variability for each tissue class as well as lower intensity contrast.

The proposed method, in contrast to typical classification strategies, does not rely on explicit bias correction. Along with learning the Markov statistics, it implicitly accounts for the bias field as well, treating it as a part of the Markov statistics, and then bases the classification on these statistics. The local sampling, as described previously in Section 3.2, plays a critical role in this process. It bases the classification-update decision on the local statistics which are corrupted almost identically with the bias field, thereby effectively eliminating the effect of the bias field on the classification. The results in Section 6 confirm this observation.

5.3 Using Probabilistic Atlases for Initialization

The classification algorithm described in Section 5.1 needs an initialization $\{T_k^0\}_{k=1}^K$. We use co-registered probabilistic atlases for the white matter, gray matter, and the cerebrospinal fluid for the purpose. We obtain these atlases from the ICBM repository [41], which also provides an average-T1 image registered with these atlases. These atlases give the *a priori* probability for a voxel belonging to one of these tissue types. We define the initialization as the maximum-a-priori estimate. To obtain this estimate the atlases need to be registered to the data. Hence, we first register the average-T1 image to the data using an affine transformation and then use the transformation to resample the three atlases. The algorithm is:

(1) Perform affine registration between the average-T1 image associated with the atlas and the data.
(2) Resample the white-matter, gray-matter, and cerebrospinal-fluid atlases based on the transformation obtained in the previous step.
Let $P_k^a(t), k = 1, 2, 3$ be the *a priori* probability, given by the atlas, for the t-th voxel belonging to the k-th tissue type.
(3) Compute the probabilities for the class (say class $k = 0$) comprising all the non-brain tissue types: $\forall t : P_0^a(t) = 1 - \sum_{k=1}^3 P_k^a(t)$.
(4) $\forall t : L^0(t) = \text{argmax}_k \, P_k^a(z(t))$.

5.4 Using Probabilistic Atlases as Priors During Classification

Probabilistic atlases have utility in the proposed classification algorithm as well. Instead of using data-driven probabilities alone for the classification updates, we can employ a Bayesian estimation strategy to compute the probabilities. Here, the likelihood term is the same as the data-driven probabilities that we have computed so far. The posterior is the likelihood multiplied by a prior that we derive from the probabilistic atlas. In contrast, we can also derive prior terms from global priors that provide no spatial probabilistic information. In this paper, however, the focus is on probabilistic atlases.

For the proposed method, empirical evidence suggests that using the atlas directly as a prior can strongly dominate over the likelihood. For instance, for regions where the prior probability is zero or near zero, the likelihood can have little effect. In such a case, the final segmentation would be very much like the initialization that we obtain via registration. Section 6.2 discusses the effect of different priors on the proposed method in more detail. To increase the effect of the data-driven likelihood term, we need to weaken the prior. We have investigated two ways of achieving the same and we discuss both of them next.

(1) One way of weakening the atlas prior is to use the atlas for discriminating only between two tissue types, namely brain and non-brain tissue. Here, we sum the atlas probabilities for the white matter, gray matter, and cerebrospinal fluid to create one composite atlas that only gives the spatial probability for any kind of brain tissue. That is,

$$\text{For } k = 1, 2, 3, \forall t : P_k^a(t) = 1 - P_0^a(t) \tag{14}$$

We call this as the *2-class* prior.

(2) Another way of reducing the strength of the prior is to voxel-wise rescale the atlas probabilities in such a way that the probabilities continue to add up to one but are less discriminating between the tissue types. We have used the following function for the desired effect.

$$\forall k, t : P_k^a(t) = (1 - v)/4 + vP_k^a(t), \tag{15}$$

where $v \in [0, 1]$ is a parameter. For the redefined prior probabilities $\forall t : \sum_{k=1}^{4} P_k^a(t) = 1$. A value of $v = 1$ makes no change to the atlas probabilities, whereas $v = 0$ makes every class equiprobable. In this paper we experiment with a moderate value of $v = 0.5$. We call this the *scaled-atlas* prior.

5.5 Implementation Issues

5.5.1 Local Sampling

The samples A_t used to estimate the PDFs $P_k(z(t))$ should consist of voxel neighborhoods that are spatially nearby voxel t. This enables the method to learn the Markov statistics for piecewise-stationary images along with the bias field (explained previously in Section 5.2). To achieve this, we choose a unique sample for each voxel, randomly, distributed according to an isotropic 3-dimensional Gaussian PDF on the image coordinates, centered at the voxel in question. Thus, the set A_t is biased and contains more voxels near the voxel being processed. Some of our previous work [23,24] explains this in more detail. In case of anisotropic MR data, we must weight the Gaussian variances along cardinal directions, making sampling more isotropic. To achieve this, we divide the σ along each axis by the grid spacing along that axis. This strategy gives consistently better results than uniform sampling. We have used a Gaussian PDF with a pre-weighted variance of 225 voxels-squared. However, we have found that it performs well for any choice of variance that encompasses more than several hundred voxels. The empirical results in Table 1 (shown in Section 6.1) confirm that the performance of the proposed method degrades gracefully for suboptimal values of this parameter.

5.5.2 Neighborhood Size and Shape

In this paper, while working with 3D MR data, we use a neighborhood comprising 6 voxels which correspond to the two voxel neighbors in each of the three cardinal directions. In case of anisotropic MR data we must weight the intensities, making neighborhoods more isotropic. We incorporate such fuzzy weights by using an anisotropic feature-space distance metric, $\|z\|_M = \sqrt{z^T M z}$, where M is a diagonal matrix with the diagonal elements being the appropriate weights on the influence of the neighbors on the center voxel. Some of our previous work [23,24] explains this in more detail. We select the weight for each neighbor to be reciprocal of the grid spacing along its associated axis.

5.5.3 Data-Driven Choice for the Sample Size

Section 3.3.1 described that we choose the maximum pseudo-likelihood (or, equivalently, minimal entropy) value of the Gaussian standard-deviation, σ, kernel parameter. We have found [23,24] that for sufficiently large $|A_t|$, the choice of σ is not sensitive to the value of $|A_t|$, thereby enabling us to automatically set $|A_t|$ to an appropriate value before the classification begins. We have implemented the Newton's method [42] to find the optimal parameter value. Thus, given the Markov neighborhood and the local-sampling Gaussian variance, the method chooses the critical Parzen-window kernel parameters σ and $|A_t|$ automatically in a data-driven fashion using information-theoretic metrics.

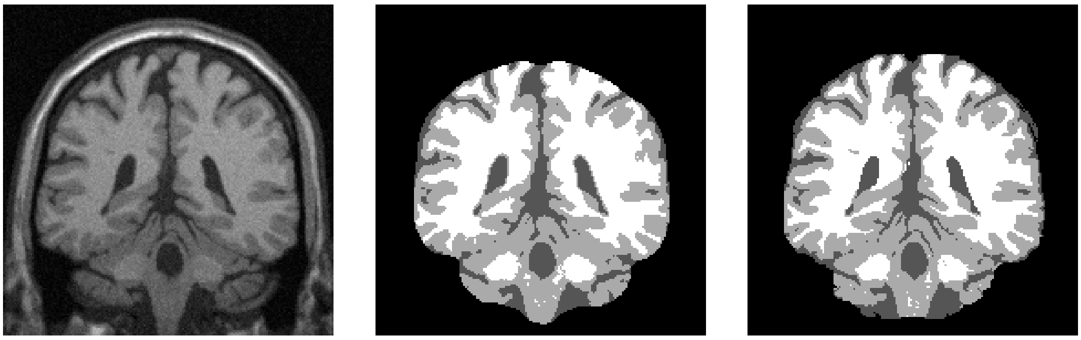
6 Results and Validation

This section gives validation results on real and synthetic brain-MR images along with the analysis of the method's behavior. It also provides quantitative comparisons with a current state-of-the-art classification method. The proposed method sets $|A_t|$, for all voxels t, to be about 500, based on the method explained in Section 5.5.3. For each iteration, it has a computational complexity of $O(K|A_t||T|)$ and, with $|A_t| = 500$, it takes about 45 minutes for processing a 181 × 217 × 181 volume on a standard single Pentium processor. The algorithm scales linearly with the number of processors on a shared-memory, e.g. dual-processor, machine. The classification typically takes about 4 to 7 iterations depending on the noise/bias level. The implementation in this paper relies on the Insight Toolkit [43].

Section 3.3.1 described a method to obtain an optimal Parzen-window kernel parameter σ by minimizing the entropy of Z. This parameter σ, essentially controls the smoothing on the data in the high-dimensional space (7-dimensional in our case) of neighborhood intensity vectors. The method of Parzen-window density estimation with single-scale isotropic kernels is, perhaps, one of the simplest such schemes. Empirical studies suggest that such a σ can be too small for the purpose of MR-image classification, splitting the high-dimensional space and, in turn, the image into many more spatial regions T_k than what may be appropriate. Table 1, later, shows that the performance is poor. Hence, in this paper, we multiply the σ obtained after the optimization by a factor of ten. Better techniques for Parzen-window density estimation that choose kernels adaptively to accommodate the signal or noise might alleviate the need for such a factor. The choice of the precise value of this *multiplicative factor* is not critical and Table 1 in the next section confirms that the algorithm is quite robust to small changes in this parameter.

6.1 Validation on Simulated MR Data

This section validates the proposed approach on simulated brain-MR images with a known ground truth. We use 1 mm isotropic T1-weighted images from the BrainWeb simulator [44] with varying amounts of noise and bias field. Figure 1 shows some data along with the classification and the ground truth. Leemput et al. [3] use the Dice metric [45] to evaluate the classification performance of their state-of-the-art approach, based on expectation maximization and Markov random fields, on images from the BrainWeb simulator. For a direct comparison, we use the same metric. Let $\{\tilde{T}_k\}_{k=1}^K$ denote the ground-truth classification and $\{T_k^*\}_{k=1}^K$ denotes the classification obtained from the proposed method. Then, the Dice metric D_k that quantifies the quality of the

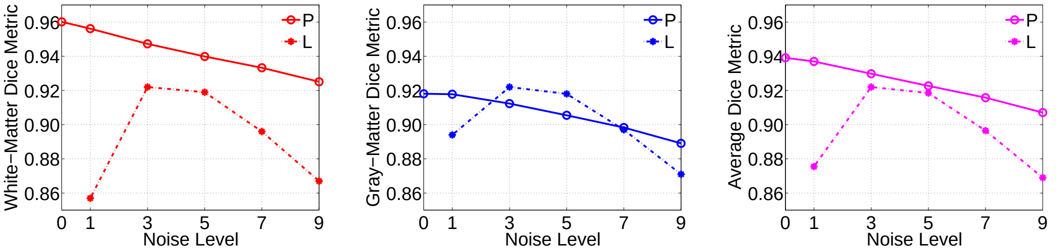


(a) (b) (c)

Fig. 1. Qualitative analysis of the proposed algorithm with BrainWeb data [44] with 5% noise and a 40% bias field. (a) A coronal slice of the data. (b) The classification produced by the proposed method. (c) The ground truth.

classification for class k is $2|T_k^* \cap \tilde{T}_k|/(|T_k^*| + |\tilde{T}_k|)$, where $|\cdot|$ is an operator giving the cardinality of sets.

We first validate on simulated T1-weighted data without any bias field and with noise levels varying from 0% to 9% . We use the *2-class* prior. The BrainWeb simulator defines the noise-level percentages with respect to the mean intensity of the brightest tissue class. Figures 2(a) and 2(b) plot the Dice metrics for gray-matter (D_{gray}) and white-matter (D_{white}) classifications for the proposed algorithm and compare them with the corresponding values for the current state-of-the-art [3]. We see that the proposed method is consistently better for the white matter. For a few noise levels for the gray matter, its performance level is slightly below the state-of-the-art. We have found that this is due to the effect of using the *2-class* prior which biases the results against the gray matter, as compared to the *scaled-atlas* prior. With the *scaled-atlas* prior the results are consistently better than the state-of-the-art for all noise levels. The next section describes that both priors perform equally well as measured by the average of the Dice metric for the white matter and gray matter, i.e.



(a) (b) (c)

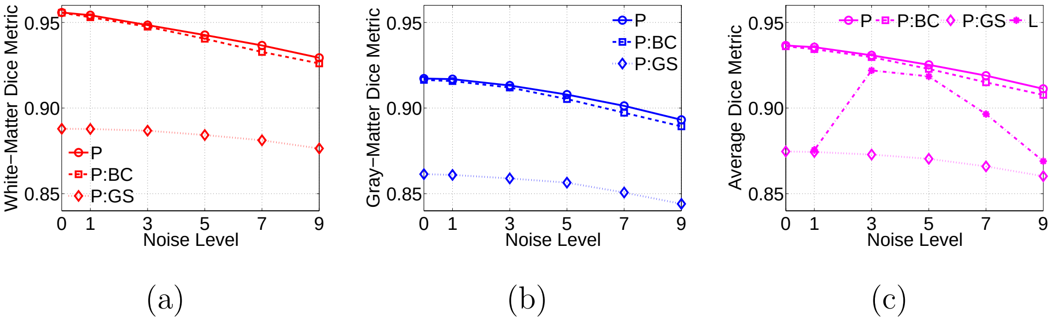
Fig. 2. Validation, and comparison with the state-of-the-art [3], on simulated T1-weighted data without any bias and varying noise levels. Here, the proposed method uses the *2-class* prior. Dice metrics for (a) white matter: D_{white}, (b) gray matter: D_{gray}, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, P: Proposed method, L: Leemput et al.'s state-of-the-art method [3].

$(D_{\text{white}} + D_{\text{gray}})/2$.

Figure 2(c) shows that for the average Dice metric, the proposed algorithm performs consistently better than the state-of-the-art at all noise levels for gray matter and white matter. Furthermore, it exhibits a slower performance degradation with increasing noise levels than the state-of-the-art method. For 3% noise, which is typical for real MRI [3], the improvement in the average Dice metric is approximately 1.1 %. The performance gain at 9 % noise is 3.8%. The larger gain over the state-of-the-art for large noise levels should prove useful for classifying noisier clinical fast-acquisition MRI.

Figure 2 shows that for low noise levels, the performance of the parametric EM-based algorithm drops dramatically. This is because it systematically assigns voxels close to the interface between gray matter and white matter to the class which happens to have a larger intensity variability [3]. This class is, inherently, the gray matter class. It turns out that, in such low-noise cases, partial voluming seems to dictate the MR-tissue intensity model which deviates significantly from the assumed Gaussian [3]. Hence, approaches enforcing Gaussian intensity PDFs on the classes, such as [3,14], would face a serious challenge in this case. In contrast, the proposed adaptive modeling strategy, which is based on nonparametric density estimation, does not suffer from this drawback. Figure 2 clearly depicts this advantage of the proposed method.

Figure 3 shows the validation results with the BrainWeb data having a 40% bias field with varying noise levels. Even in the absence of an explicit bias-correction scheme, the method performs equally well on both biased and unbiased MR-data (Figure 2). This is because of the adaptive model of higher-order statistics underlying the method, as explained before in Section 5.2. To confirm the vital role that the *local sampling* Parzen-window density estimation strategy plays in enabling the automatic learning of the bias field, we perform



(a) (b) (c)

Fig. 3. Validation, and comparison with the state-of-the-art [3], on simulated T1-weighted data with 40% bias and varying noise levels. We compare the performance by incorporating explicit bias correction and *global sampling* (see text). Dice metrics for (a) white matter: D_{white}, (b) gray matter: D_{gray}, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, P: Proposed method, BC: Bias correction, GS: *Global sampling*, L: Leemput et al.'s state-of-the-art method [3].

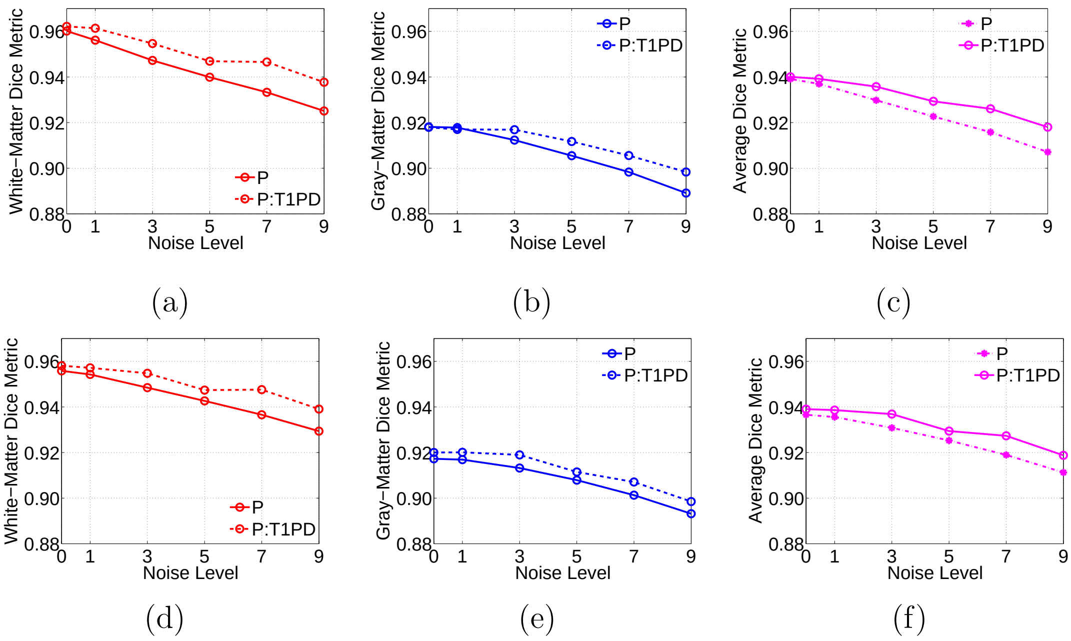
two more experiments. In the first experiment, we use explicit bias correction with the proposed method (degree-4 polynomial fit to the white-matter intensities iteratively). Figure 3 shows that this method performs approximately as well, but not any better than without the bias correction. The second experiment replaced the *local sampling* scheme with a *global sampling* scheme that chooses the random Parzen-window sample uniformly over the image as was done in our previous work [5]. Figure 3 shows that this scheme performs significantly worse at all noise levels in the absence of bias correction. These results empirically justify the choice of changing the sampling strategy from global to local as discussed in Section 5.2.

To study the sensitivity of the variance parameter for the local-sampling Gaussian associated with Parzen-window density estimation and the Parzen-window σ multiplicative factor, we measure the Dice metrics for the white matter and gray matter over a range of values. We use the BrainWeb T1 data with 5% noise and a 40% bias field. Table 1 gives the results confirming that the classification performance is fairly robust to changes in the values of these two parameters, as explained before in Section 5.5.3.

We can extend the proposed method in a straightforward manner to deal with multimodal data. Multimodal segmentation entails classification using MR images of multiple modalities, e.g. T1 and PD. It treats the combination of images as an image of vectors with the associated PDFs in the *combined* probability space. Figure 4 shows the classification results for multimodal data using T1

Table 1
The proposed method is fairly robust to changes in the values of the local-sampling Gaussian variance parameter and the Parzen-window σ multiplicative factor. This table gives the Dice metrics for the BrainWeb T1 data with 5% noise and a 40% bias field.

Local-sampling Gaussian variance	Gray matter	White matter
100	0.9033	0.9386
225	0.9079	0.9427
400	0.9082	0.9422
625	0.9043	0.9368
Parzen-window σ *multiplicative factor*	Gray matter	White matter
1.0	0.7634	0.9105
2.5	0.8988	0.9502
5.0	0.9106	0.9487
7.5	0.9095	0.9451
10.0	0.9079	0.9427
12.5	0.9066	0.9411
15.0	0.9058	0.9402



(a) (b) (c)

(d) (e) (f)

Fig. 4. Validation on simulated multimodal (T1 and PD) data with varying noise levels. Dice metrics for (a) white matter: 0% bias, (b) gray matter: 0% bias, and (c) their average: 0% bias. Dice metrics for (d) white matter: 40% bias, (e) gray matter: 40% bias, and (f) their average: 40% bias. *Note*: In the graphs, P: Proposed method, T1PD: Using both T1 and PD images.

and PD images, both with and without a bias field. The results demonstrate that incorporating more information in the classification framework, via images of two modalities T1 and PD, produces consistently better results than using T1 images alone.

6.2 Validation on Real MR Data

The section shows validation results with real expert-classified MR data. We obtained this data set from the IBSR website [46]. The data set comprises T1-weighted brain-MR images for 18 subjects. Figure 5 shows an example from the data set. We observe that the data has lower contrast and possesses certain acquisition-related artifacts that makes the classification task more challenging than that for the BrainWeb dataset. Figure 5 also shows an example of a classification generated by the proposed method and compares it to the ground truth.

Figure 6 compares the performance of the proposed method using the two different atlas-based priors. Figure 6(a) shows that the *2-class* prior, relative to the *scaled-atlas* prior, biases the classification more in favor of the white matter. With the *2-class* prior, which gives equal weight to all three brain-tissue types, the Dice metric for the white matter is better than that for the

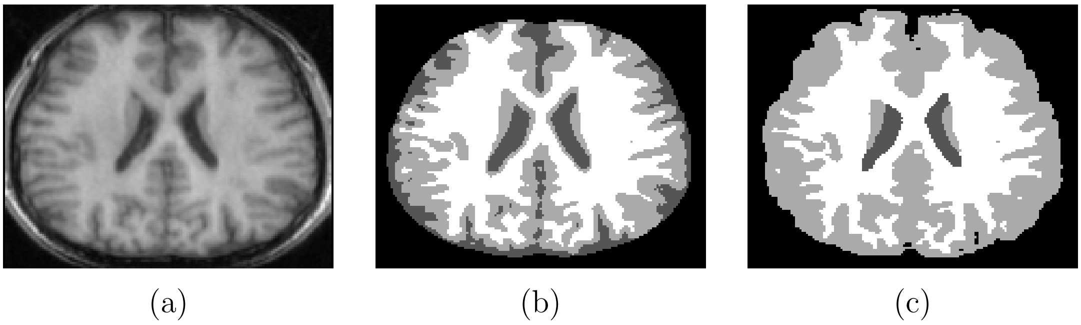


Fig. 5. Qualitative analysis of the proposed algorithm with IBSR data [46]. The voxel size for this image is 0.9375×0.9375×1(coronal) (a) An axial slice of the data. (b) The classification produced by the proposed method. (c) The expert-classified ground truth.

gray matter because of lower inherent variability of the intensities in the white matter. The *scaled-atlas* prior imposes a stronger constraint which tends to shift this bias, as seen in Figure 6(b). Empirical evidence confirms that as the parameter v varies from 0.0 to 1.0 the bias shifts away from white matter towards gray matter. Nevertheless, with the average Dice metric, Figure 6(c) shows that both priors perform equally well.

For the proposed algorithm using the *2-class* prior, Table 2 gives the mean, median, and the standard deviation for the Dice metrics over the entire dataset. The proposed method yields a higher mean (by a couple of percent) and lower

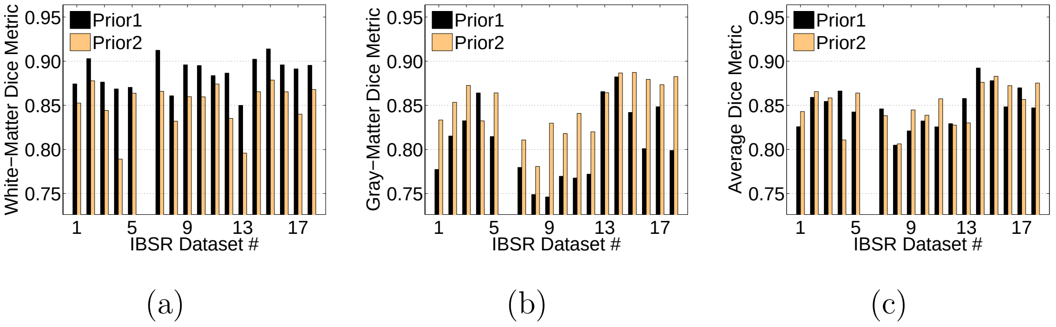


Fig. 6. Validation, of the proposed method with two different atlas-based priors, on IBSR data. Dice metrics for (a) white matter: D_{white}, (b) gray matter: D_{gray}, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, Prior1: *2-class* prior, Prior2: *scaled-atlas* prior.

Table 2
Mean, median, and standard deviation for the gray-matter and white-matter tissue classes in the IBSR data set using the proposed method with the *2-class* prior.

Statistical measure	White matter	Gray matter
Mean	0.8868	0.8074
Median	0.8913	0.8009
Standard deviation	0.0179	0.0426

standard deviation for the Dice metrics over both white matter and gray matter classes, as compared to the results reported by Ruf et al. [14] for Leemput et al.'s state-of-the-art method [3] as well as their own method.

7 Conclusions and Discussion

This paper presents a novel method for unsupervised brain-MRI tissue classification by adaptively learning the higher-order image statistics via data-driven nonparametric density estimation. It also describes the essential theoretical aspects underpinning adaptive, nonparametric Markov modeling and the theory behind the consistency of such a model. The proposed method relies on the information content of input data for setting important parameters, and does not require significant parameter tuning. Moreover, it does not rely on any kind of training. The adaptive image model enables the method to implicitly account for the bias field and perform equally well on both biased and unbiased MR-data without requiring any bias correction. Incorporating the information content in neighborhoods in the classification process virtually eliminates the need for explicit smoothness constraints on the classification, and provides optimal regularization.

The results in the paper empirically confirm that the piecewise-stationary Markov model conforms well to brain-MR images. It shows that it is possible to construct nonparametric density estimations in the high-dimensional spaces of MR-image neighborhoods. These results also suggest that the statistical structure in these spaces capture important tissue properties in brain-MR images. The formulation underlying the proposed method generalizes in several different ways. The statistical and engineering components in this paper are appropriate for any kind of densely sampled medical data. This includes images with higher-dimensional domains (e.g. sequence of volumetric MR images over time) and vector-valued data (e.g. multimodal MR data).

The proposed method can be further improved via some engineering advances. For instance, the method of density estimation with single-scale isotropic Parzen-window kernels is, perhaps, one of the simplest such schemes. Parzen-window density estimation can improve by choosing kernels adaptively to accommodate the signal or noise. The Markov neighborhood in the current algorithm comprises only first-order neighbors. Using larger neighborhoods could, potentially, improve the results, but entails longer computation times. Of course, very large neighborhoods are infeasible because of unavailability of sufficiently many observations in the higher-dimensional space for the Parzen-window density estimation. The computation times for the implementation are impractical for most applications. Improving the computational scheme, e.g. using methods based on fast Gauss transforms [47], is an important area of future work.

References

[1] C. Cocosco, A. Zijdenbos, A. Evans, A fully automatic and robust brain MRI tissue classification method, Medical Image Analysis 7 (4) (2003) 513–527.

[2] R. Nowak, Wavelet-based rician noise removal for magnetic resonance imaging, IEEE Trans. Imag. Proc. 8 ('99) 1408–1419.

[3] K. V. Leemput, F. Maes, D. Vandermeulen, P. Seutens, Automated model-based tissue classification of mr images of the brain, IEEE Tr. Med. Imaging 18 (1999) 897–908.

[4] M. Prastawa, J. H. Gilmore, W. Lin, G. Gerig, Automatic segmentation of neonatal brain MRI, in: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 2004, pp. 10–17.

[5] T. Tasdizen, S. P. Awate, R. T. Whitaker, N. L. Foster, MRI tissue classification with neighborhood statistics: A nonparametric, entropy-minimizing approach., in: Proc. Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI), Vol. 3750, 2005, pp. 517–525.

[6] G. Gerig, O. Kubler, R. Kikinis, F. A. Jolesz, Nonlinear anisotropic filtering of MRI data, IEEE Tr. Med. Imaging 11 (2) (1992) 221–232.

[7] W. M. Wells, W. E. L. Grimson, R. Kikinis, F. A. Jolesz, Adaptive segmentation of MRI data, IEEE Tr. Med. Imaging 15 (4) (1996) 429–443.

[8] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39 (1977) 1–38.

[9] K. V. Leemput, F. Maes, D. Vandermeulen, P. Seutens, Automated model-based bias field correction of mr images of the brain, IEEE Tr. Med. Imag. 18 (1999) 885–896.

[10] T. Kapur, W. E. L. Grimson, W. M. Wells, R. Kikinis, Segmentation of brain tissue from magentic resonance images, Med. Im. An. 1 (1996) 109–127.

[11] K. Held, E. R. Kops, B. J. Krause, W. M. Wells, R. Kikinis, H.-W. Muller-Gartner, Markov random field segmentation of brain mr images, IEEE Tr. Med. Imaging 16 (6) (1997) 878–886.

[12] C. Pachai, Y. M. Zhu, C. R. G. Guttmann, R. Kikinis, F. A. Jolesz, G. Gimenez, J.-C. Froment, C. Confavreux, S. K. Warfield, Unsupervised and adaptive segmentation of multispectral 3d magnetic resonance images of human brain: a generic approach, in: Proc. Int. Conf. Medical Image Computing and Computer Assisted Intervention, 2001, pp. 1067–1074.

[13] Y. Zhang, M. Brady, S. Smith, Segmentation of brain mr images through a hidden markov random field model and the expectation maximization algorithm, IEEE Tr. Med. Imaging 20 (2001) 45–57.

[14] A. Ruf, H. Greenspan, J. Goldberger, Tissue classification of noisy mr brain images using constrained gmm., in: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 2005, pp. 790–797.

[15] C. Davatzikos, J. Prince, An active contour model for mapping the cortex, IEEE Trans. Medical Imaging 14 (1) (1995) 65–80.

[16] R. Valds-Cristerna, V. Medina-Bauelos, O. Yez-Surez, Coupling of radial-basis network and active contour model for multispectral brain MRI segmentation, IEEE Trans. Biomedical Engineering 51 (3) (2004) 459–470.

[17] A. Toga, Brain Warping, Academic Press, 1999.

[18] B. A. C. O. V. J. Bach Cuadra M, Pollo C, T. J, Atlas-based segmentation of pathological MR brain images using a model of lesion growth, IEEE Transactions on Medical Imaging 23 (10) (2004) 1301– 1314.

[19] T. Rohlfing, C. R. M. Jr., Multi-classifier framework for atlas-based image segmentation., in: IEEE Int. Conf. Comp. Vis. Pattern Recog., 2004, pp. 255–260.

[20] A. Lee, K. Pedersen, D. Mumford, The nonlinear statistics of high-contrast patches in natural images, Int. J. Comput. Vision 54 (1-3) (2003) 83–103.

[21] V. de Silva, G. Carlsson, Topological estimation using witness complexes, Symposium on Point-Based Graphics.
URL `http://math.stanford.edu/comptop/preprints/`

[22] K. Popat, R. Picard, Cluster based probability model and its application to image and texture processing., IEEE Trans. Image Processing 6 (2) (1997) 268–284.

[23] S. P. Awate, R. T. Whitaker, Higher-order image statistics for unsupervised, information-theoretic, adaptive, image filtering, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2005, pp. 44–51.

[24] S. P. Awate, R. T. Whitaker, Unsupervised, information-theoretic, adaptive image filtering for image restoration, To Appear, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), March 2006 (estimated).

[25] A. Efros, T. Leung, Texture synthesis by non-parametric sampling, in: Int. Conf. Computer Vision, 1999, p. 1033.

[26] L. Wei, M. Levoy, Order-independent texture synthesis, Stanford University Computer Science Department Tech. Report TR-2002-01.
URL `http://graphics.stanford.edu/projects/texture/`

[27] R. Paget, Strong markov random field model, IEEE Trans. Pattern Anal. Mach. Intell. 26 (3) (2003) 408–413.

[28] E. Levina, Statistical issues in texture analysis, Ph.D. Dissertation, Department of Statistics, University of California, Berkely.

[29] E. Dougherty, Random Processes for Image and Signal Processing, Wiley, 1998.

[30] B. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.

[31] D. W. Scott, Multivariate Density Estimation, Wiley, 1992.

[32] E. Parzen, On the estimation of a probability density function and the mode, Annals of Math. Stats. 33 (1962) 1065–1076.

[33] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, 2001.

[34] P. Viola, W. Wells, Alignment by maximization of mutual information, in: Int. Conf. Comp. Vision, 1995, pp. 16–23.

[35] J. Besag, Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society, series B 36 (2) (1974) 192–236.

[36] J. Besag, Statistical analysis of non lattice data, Journal of the Royal Statistical Society 24 (1975) 179–195.

[37] S. Geman, C. Graffigne, Markov random field image models and their applications to computer vision, in: Proc. Int. Congress of Mathematicians, 1986, pp. 1496–1517.

[38] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley, 1991.

[39] J. Kim, J. W. Fisher, A. Yezzi, M. Cetin, A. S. Willsky, Nonparametric methods for image segmentation using information theory and curve evolution, in: Proc. IEEE Int. Conf. on Image Processing, 2002, pp. 797–800.

[40] S. Jehan-Besson, M. Barlaud, G. Aubert, Dream2s: Deformable regions driven by an eulerian accurate minimization method for image and video segmentation, in: Proc. European Conf. on Computer Vision-Part III, 2002, pp. 365–380.

[41] D. E. Rex, J. Q. Ma, A. W. Toga, The LONI pipeline processing environment, NeuroImage 19 (2003) 1033–1048.

[42] S. S. Rao, Engineering Optimization, Theory and Practice, Wiley, 1996.

[43] NLM Insight Segmentation and Registration Toolkit (ITK).
URL `http://www.itk.org`

[44] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, A. C. Evans, Design and construction of a realistic digital brain phantom., IEEE Trans. Med. Imag. 17 (3) (1998) 463–468.

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

[46] Internet Brain Segmentation Repository (IBSR).
URL `http://www.cma.mgh.harvard.edu/ibsr/`

[47] C. Yang, R. Duraiswami, N. Gumerov, L. Davis, Improved fast gauss transform and efficient kernel density estimation, in: Int. Conf. Comp. Vision, 2003, pp. 464–471.