Available online at www.sciencedirect.com
ScienceDirect
Medical Image Analysis 10 (2006) 726–739
MEDICAL IMAGE ANALYSIS
www.elsevier.com/locate/media

Adaptive Markov modeling for mutual-information-based, 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 South Central Campus Drive, 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

Received 23 January 2006; received in revised form 4 July 2006; accepted 10 July 2006
Available online 21 August 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 Markov 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 reasonably well on T1-weighted MR data 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.

Keyword: Adaptive image modeling

1. Introduction

Tissue classification in magnetic resonance (MR) images of human brains is an important problem in medical image analysis. The fundamental task in tissue classification is to classify the voxels in the volumetric (three-dimensional/3D) 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 segmentation or classification of high-resolution three-dimensional images is a tedious task, which is 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 (Cocosco et al., 2003). *Fully automatic* or *unsupervised* methods, on the other hand, virtually eliminate the need for manual interaction, and thus such methods for brain-tissue classification have received significant attention in the literature.

Current state-of-the-art methods for automatic brain-tissue classification typically incorporate the following strategies: (a) parametric statistical modeling, e.g. Gaussian, of

* Corresponding author. Tel.: +1 801 585 1867.
E-mail addresses: suyash@cs.utah.edu (S.P. Awate), tolga@cs.utah.edu (T. Tasdizen), Norman.Foster@hsc.utah.edu (N. Foster), whitaker@cs.utah.edu (R.T. Whitaker).
URL: http://www.cs.utah.edu/~suyash (S.P. Awate).

1361-8415/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2006.07.002

voxel grayscale intensity for each tissue class, (b) Markov-random-field (MRF) modeling to enforce spatial smoothness on the classification, (c) methods to explicitly correct for the inhomogeneities inherent in MR images, and (d) probabilistic-brain-atlas information in the classification method. Several factors, however, continue to pose significant challenges to the state-of-the-art. These include:

- The intensities and contrast in MR images varies significantly with the pulse sequence, and several other scanner parameters. The quality of MR data also shows a certain amount of variation when produced at multiple sites with different MR scanners.
- Magnetic resonance imaging (MRI) acquisition artifacts, which include the Rician nature of the noise in magnitude-MR data (Nowak, 1999) and partial voluming effects (Leemput et al., 1999b), can cause the data to significantly deviate from the Gaussian models, thereby compromising the quality of the classification.
- Many methods treat the inhomogeneity as multiplicative noise (bias field) and explicitly correct the MR intensities to reduce its effect. There are, however, cases, where the noise is no longer multiplicative. For certain kinds of coil configurations or applications, such as neonatal brain MRI, however, inhomogeneities do not adhere to standard multiplicative models (Prastawa et al., 2004).

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. Based on this key idea, this paper presents a novel method for MRI brain-tissue classification that incorporates an adaptive nonparametric model of neighborhood/Markov statistics. The method incorporates the information content in the neighborhoods in the classification process. Together with a weak smoothness constraint on the estimated Markov statistics, it virtually eliminates the need for explicit smoothness constraints on the class-label image. The method produces an optimal classification by iteratively maximizing a mutual-information metric that relies on Markov probability density function (PDF). The algorithm 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 *a priori* information in probabilistic-brain-atlases in a coherent manner via a Bayesian formulation. Experiments on real, simulated, and multimodal data demonstrate the significant advantages of the method over the current state-of-the-art. The method also performs reasonably well without any explicit inhomogeneity correction.

The rest of the paper is organized as follows. Section 2 discusses works in MR-image classification and Markov modeling along with their relationships to the proposed method. Section 3 presents the mathematical underpinnings of the proposed method, which relies on an adaptive, MRF image model. Section 4 formulates the classification as an optimal-segmentation problem associated with an information-theoretic goodness measure on Markov image statistics. Section 5 focuses on the application of the proposed method to brain-tissue classification. It explains why the method performs reasonably well in the absence of explicit inhomogeneity correction, describes a strategy for data-driven choice of important internal parameters, 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 MRI brain-tissue 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 various classification strategies have evolved, including (a) partitioning based on grayscale voxel-intensity data, (b) regularization schemes based on local interactions among class labels, and (c) spatial priors based on probabilistic and anatomical atlases.

Wells et al. (1996) present a method that couples tissue classification with inhomogeneity correction based on maximum-likelihood parameter estimation. They use the expectation-maximization (EM) algorithm of Dempster et al. (1977) to simultaneously estimate the unknown bias field and the classification. Leemput et al. (1999a,b) 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 each tissue-intensity distribution conforms to a parametric Gaussian PDF whose parameters are obtained via the EM algorithm. 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 the data using neighborhood sampling and nonparametric density estimation.

The EM-classification algorithm does not impose any smoothness constraint on the classification and it is therefore susceptible to outliers in the tissue intensities. Some approaches for tissue classification do not explicitly account for noise, but employ image-denoising methods as a preprocessing step (Gerig et al., 1992; Lysaker et al., 2003). Many subsequent works incorporate noise models into the classification without such preprocessing. Several authors (Kapur et al., 1996; Held et al., 1997; Leemput et al., 1999a,b; Pachai et al., 2001; Zhang et al., 2001) have extended the EM-classification algorithm to incorporate spatial smoothness via Gibbs/Markov priors on the label image. For instance, Kapur et al. (1996) 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, making them more likely. Such strong Markov models, however, 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 (Held et al., 1997; Leemput et al., 1999a,b). Ruf et al. (2005) 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.

This tissue-classification work dovetails with the mainstream image-processing literature, which presents a variety of algorithms that rely on MRF models of images (Geman and Geman, 1984; Besag, 1986; Owen, 1989; Li, 1995; Stark and Woods, 2001). Such methods typically involve iterative stochastic-relaxation schemes that compute local image updates based on random sampling from local conditional PDFs. These conditional PDFs on neighborhood configurations define an energy that is progressively reduced. Typically, the methods specify the conditional PDFs in parametric forms, e.g. Gaussian (Li, 1995). In this way, they encode a set of probabilistic assumptions (priors) about the geometric/statistical properties of the image data, and thus they are effective only when the data conforms sufficiently well to the prior. Furthermore, the previous work on MRI classification models each tissue class with Gaussian-mixture models, which is homogeneous across the image. The proposed method – rather than enforcing a particular Markov prior on the data – *learns* the relevant Markov statistics nonparametrically from the input data and bases the classification on this adaptive model.

Researchers have also used active contour models (Davatzikos and Prince, 1995; ValdTs-Cristerna et al., 2004) to impose smoothness constraints for segmentation. These methods typically attempt to minimize the area of the segmentation boundary (smoothness) simultaneously with proper fidelity to the data. These models produce results that can be quite sensitive to the contour parameters that control the influence of the data and the smoothness. Hence, these methods typically require careful manual parameter-tuning. The proposed method, on the other hand, sets its important internal free parameters via data-driven techniques using information-theoretic optimality criteria. As a result, it easily applies to a wide spectrum of data with little parameter tuning.

An important component in MRI brain tissue classification is the correction of intensity inhomogeneities or bias fields. Several approaches propose an approach that couples iterative updates of the class labels with the bias-field correction based on polynomial least-squares fitting (Wells et al., 1996; Guillemaud and Brady, 1997; Leemput et al., 1999a). Although, the focus of this paper is not on inhomogeneity correction, it is compatible with all such schemes. The literature also presents many methods that aim at *implicitly* dealing with inhomogeneities in the classification method itself (Yan and Karp, 1995a,b; Lee and Vannier, 1996; Rajapakse et al., 1997; Nocera, and Gee, 1997). For instance, Yan and Karp (1995b) employ an adaptive K-means clustering strategy that, over many iterations, gradually takes the feature-space points from increasingly-local neighborhoods. The initial segmentation uses all points in the image but the final segmentation implicitly accounts for local intensity variations such as those cause by the inhomogeneity field.

More recently, researchers have realized the importance of the *nonstationarity* of head images in tissue classification, and several authors introduce global information in the form of anatomical atlases (Toga, 1999; Cuadra et al., 2004; Rohlfing and Maurer, 2004). Typically, they use atlases in one of the two ways. First is to 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 based on the labels in the atlas. Several authors use *probabilistic atlases*, which are generated from ensembles of head images. These atlases encode tissue probabilities (rather than discrete label values) at each voxel, and are used as a prior in the EM estimation described previously (Craene et al., 2004). The proposed method uses probabilistic atlases for the initialization, which is important to the success of the algorithm, and can include probabilities from atlases in the posterior estimation.

The proposed method learns Markov statistics nonparametrically entailing 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 seven-dimensional space (center voxel along with its neighbors). High-dimensional spaces are notoriously challenging for data analysis because they are so sparsely populated. This is one of the effects of *the curse of dimensionality* (Silverman, 1986; Scott, 1992). Despite theoretical arguments suggesting that density estimation beyond a few dimensions is impractical due to the unavailability of (theoretically) sufficient data, the empirical evidence from the literature is more optimistic (Scott, 1992; Popat and Picard, 1997). Indeed, the results in this paper confirm that observation. Furthermore, the proposed method relies on a stationary Markov model implying that the neighborhood random vector has identical marginal PDFs, thus lending itself to more accurate density estimates (Scott, 1992; Silverman, 1986). Researchers analyzing the statistics of natural images in terms of local neighborhoods describe results that are consistent with Markov image models. For instance, Lee et al. (2003) as well as de Silva and Carlsson (2004) analyze the statistics of 3-pixel $\times$ 3-pixel neighborhoods in images, in the corresponding nine-dimensional spaces, and find the data to be concentrated in clusters and low-dimensional manifolds exhibiting nontrivial topologies. Therefore, in

the feature space, locally, the Markov PDFs are lower-dimensional entities that lend themselves to better density estimation.

The literature presents some examples of algorithms that empirically learn the Markov image statistics. Popat and Picard (1997) were among the pioneers to use nonparametric Markov sampling in images. They model the Markov image statistics via cluster-based nonparametric density estimation, unlike the Parzen-window scheme described in this paper. They exploit their nonparametric Markov model for image restoration, image compression, and texture classification. Their learning approach, however, relies on training data, which limits its practical use. In contrast, the proposed method learns the Markov statistics of the image directly from the input data. Several researchers, mostly in the computer-graphics literature, have proposed texture-synthesis algorithms that rely on learning Markov statistics from a sample texture image to construct new images having the same Markov statistics as the input texture (Efros and Leung, 1999; Wei and Levoy, 2002; Paget, 2003).

The method described in this paper, an extension of Tasdizen et al. (2005), builds on the previous work by the authors in Awate and Whitaker (2005) and Awate and Whitaker (2006), which lays down the building blocks for unsupervised learning of Markov image statistics and proposes entropy reduction on Markov 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.

3. Adaptive image modeling via nonparametric Markov random fields

The proposed method constructs a segmentation strategy based on a Markov statistical image model (Li, 1995) that it *learns* automatically from the input data. It formulates the segmentation problem as an optimization problem to maximize the dependency or *mutual information* (Cover and Thomas, 1991) between the segmentation labels and the Markov image statistics (see Section 4). This section presents the statistical theory behind the novel adaptive-MRF image model underpinning the proposed classification approach.

A *random field* (Dougherty, 1998; Stark and Woods, 2001) is a family of random variables $X(\Omega; T)$, for some index set T, where for each $t \in T$, 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 choose one $\omega \in \Omega$, we have a realization of the random field called the *digital image*, $X(\omega, T)$. For 3D images, t is a 3-tuple and T is the set of voxels in the image. 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* (Dougherty, 1998; Stark and Woods, 2001). 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 local statistical dependence for voxel intensities during image formation. In this way, 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 $N_t \subset T$, and $u \in N_t$ if and only if $t \in N_u$, then N is a *neighborhood system* for the set T. Voxels in N_t – can include t itself – are within the neighborhood of voxel t. Section 5.6.1 discusses the neighborhood shape used in the paper. We define a random vector $Z(t) = \{X(t)\}_{t \in N_t}$. In this paper, we refer to the PDFs $P(Z(t))$ as Markov PDFs.

3.1. Unsupervised learning of Markov statistics

The proposed method exploits the Markovity property in images, but we know neither the functional forms nor the parameter values for the Markov model, i.e. the PDFs $P(Z(t))$. Images obtained by varying MRI-parameter values, e.g. T1, T2, and PD, or varying noise and bias fields represent distinct Markov models. For a segmentation method to be effective in all such cases, we propose an *adaptive* Markov model that derives from the input data.

A statistical *model* is a set of PDFs on the sample space associated with the data. Parametric statistical modeling parameterizes this set using a few control variables. An inherent difficulty with this approach is to find suitable parameters such that the model is well-suited for the data. For instance, most parametric PDFs are unimodal, whereas typical practical problems involve multimodal PDFs. *Nonparametric* statistical modeling (Duda et al., 2001) fundamentally differs from this approach by not imposing strong parametric models on the data. It provides the power to model and learn arbitrary PDFs via data-driven strategies.

In order to rely on image samples to produce nonparametric estimates of Markov statistics, we must assume that different neighborhood-intensities in the image are derived from the same PDF. Mathematically, this is the notion of *stationarity* associated with a random field. A stationary region $R \subset T$ is one, where the Markov PDFs $P(Z(t))$ are exactly the same for all voxels t in that region (Dougherty, 1998; Stark and Woods, 2001), i.e.

$$\forall t \in R, \quad P(Z(t)) = P(Z). \tag{1}$$

In other words, the Markov statistics are shift invariant. For brain-MR images, the Markov PDFs at voxels in individual parts of the brain, such as white matter or gray matter, are similar and, hence, the piecewise-stationary model holds to some degree. Stationarity provides many observations $\{z(t)\}_{t \in R}$, all derived from $P(Z)$.

Stationarity alone, however, is not sufficient to provide accurate estimates of the Markov PDFs from a single observed image. To do this, we must rely on another statistical property, namely *ergodicity*. Essentially, ergodicity guarantees accurate estimation of certain *ensemble* properties of the random field, e.g. the Markov PDFs $P(Z)$, from observations $\{z(t)\}_{t\in R}$ in a *single* realization of the stationary random field, i.e. the observed image. Mathematically, it guarantees that, for certain quantities associated with $P(Z)$, the spatial averages (i.e. over R) converge to the ensemble averages (i.e. over z) as the size of the image $|R|$ tends to infinity (Stark and Woods, 2001). It does so by ensuring that: (a) the random variables become progressively more independent with increasing spatial distance at a sufficiently-rapid rate and (b) random variables become independent as the shift between them approaches infinity. Therefore, spatial averages over sufficiently-large regions appear as averages of nearly-independent random variables and, subsequently, the weak law of large numbers (Stark and Woods, 2001) ensures the convergence of such averages to the desired ensemble average, as described in more detail in Section 3.2.

To represent the PDFs of image neighborhoods, $P(Z)$, we use the nonparametric *Parzen-window* technique (Parzen, 1962; Duda et al., 2001). The Parzen-window probability estimate for $P(Z=z)$ is defined as the ensemble average

$$P(Z=z)=\frac{1}{|S|}\sum_{y\in S}G_n(z-y;\Psi_n), \tag{2}$$

where S is a *random sample* (Dougherty, 1998; Stark and Woods, 2001) drawn from the PDF $P(Z)$, $n=|N_t|$ is the neighborhood size, and $G_n(z;\Psi_n)$ is the n-dimensional Gaussian kernel with zero mean and covariance matrix Ψ_n. Having no *a priori* information on the structure of $P(Z)$, we choose an isotropic Gaussian kernel $\Psi_n=\sigma I_n$, where I_n is the $n\times n$ identity matrix. Ergodicity enables us to approximate the ensemble average as a spatial average

$$P(Z=z)\approx\frac{1}{|A|}\sum_{t\in A}G_n(z-z(t);\Psi_n), \tag{3}$$

where the set A is a small subset of R. Taking $A=R$ increases the algorithmic complexity of the scheme. Section 5.4 describes an effective technique of choosing this Parzen-window sample. The density estimate varies with the kernel-parameter σ value. Properly tuning σ is especially critical in high-dimensional spaces because of the relatively-high sparseness of the spaces; Section 5.5 describes a data-driven technique to estimate an optimal kernel-parameter σ value.

3.2. Consistency for the nonparametrically-estimated Markov model

The power of the Markov model on the random field and nonparametric density estimation comes with some additional theoretical constraints that warrant mention. In order for the Parzen-window estimation to converge (Parzen, 1962; Duda et al., 2001) the kernel-parameter σ must decrease with the increasing number of samples. This relationship can be derived from the data itself, and several authors have proposed maximum-likelihood based schemes for estimating σ (Besag, 1975; Geman and Graffigne, 1986). We have found that a constant multiple of the maximum-likelihood σ works well for MRI classification. Section 5.5 discusses the practical issues in more detail.

Another important issue is *consistency*. A consistent system is one, where the joint PDF $P(\{X(t)\}_{t\in T})$ of all the random variables gives, using rules of probabilistic inference, each conditional PDF uniquely. Besag's proof of the Hammersely-Clifford theorem (Besag, 1974), also known as the Markov-Gibbs equivalence theorem, shows 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.

The 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, because of the convergence of the Parzen-window density estimate. This convergence holds only when the *observations* in the sample are independently generated from a single underlying PDF. The stationarity of the Markov random field implies that all observations are derived from a single PDF. However, in our case, these observations are the neighborhood-intensity vectors, which share neighboring voxel values. Independence requires sampling from a subset U of the entire voxel-set T, such that no two voxels in the subset have overlapping neighborhoods ($\forall a,b\in U: N_a\cap N_b=\phi$). The constraint of nonoverlapping neighborhoods leads to a wastage of a large amount of data ($\{z(t)\}_{t\in T\setminus U}$) (Besag, 1974), which would, in practice, lead to too few image samples. However, Levina (1997) shows that in this particular situation (overlapping neighborhoods), converge holds even in the case of overlapping data, and thus it is appropriate to sample from the entire set of image neighborhoods.

4. Optimal segmentation via mutual-information maximization on Markov statistics

This section formulates the classification problem as an optimal-segmentation problem using with an information-theoretic goodness measure associated with the Markov PDFs. It begins by forming a connection between information-theoretic measures, such as mutual information, entropy (Cover and Thomas, 1991), and classification.

Loosely speaking, the mutual information between two random variables quantifies the degree of *functional dependence* between them. For functionally-dependent random variables, each variable uniquely determines the other, and the mutual information is maximized. On the other hand, independent random variables convey no information

about each other, and their mutual information is zero (minimal). For image segmentation (Kim et al., 2005), we can say that a good segmentation is one in which the voxel-neighborhood-intensity values provide the most information about the class labels. Likewise, knowing the voxel class should provide the most reliable estimate of the voxel neighborhood. Clearly, there is no strict functional dependence and images are inherently stochastic, but mutual information provides a well-founded mechanism for quantifying the degree to which these properties hold.

For the Markov image model, we consider a discrete random variable L that maps each voxel t to the class to which it belongs, i.e. $L(t) = k$ if voxel t is in class k. Let $\{T_k\}_{k=1}^K$ denote a mutually-exclusive and collectively-exhaustive decomposition of the image domain T into K stationary-ergodic MRFs such that $T_k = \{t \in T : L(t) = k\}$. Stationarity implies that for each class k, the conditional PDFs $P(Z(t)|L(t) = k)$ are the same $\forall t \in T_k$. For notational simplicity, we refer to these conditional PDFs, one for each class k, as $P(Z|L = k) = P_k(Z)$. Using these conditional PDFs, we can also define a joint PDF $P(L, Z)$ between L and Z. At each voxel t, an instance $(l(t), z(t))$ is drawn from the joint PDF. What we observe are, however, only the intensity vectors $z(t)$. The labels $l(t)$ remain unknown and those are precisely what we want to recover. We define the optimal segmentation as the one that maximizes the mutual information between L and Z, i.e.

$$I(L,Z) = h(Z) - h(Z|L) = h(Z) - \sum_{k=1}^{K} P(L=k)h(Z|L=k), \tag{4}$$

where $I(\cdot)$ is the mutual information function and $h(\cdot)$ is the entropy (or differential entropy for continuous random variables). Entropy is a measure of randomness or uncertainty associated with a PDF (Cover and Thomas, 1991), and regions T_k having low entropies $h(Z|L = k)$ for Markov PDFs exhibit a high degree of predictability in their neighborhoods.

The entropy of class k is

$$h(Z|L=k) = -\int_{\Re^{|N_t|}} P_k(Z=z)\log P_k(Z=z)\,\mathrm{d}z, \tag{5}$$

where $|N_t|$ is the neighborhood size and $P_k(Z = z)$ is the probability of observing a neighborhood-vector z in class k.

The entropy of the Markov PDF associated with the entire image, $h(Z)$, is independent of the label assignment L and we can ignore it during the optimization. Thus, Eq. (4) implies that the optimal segmentation is the one that minimizes a weighted average of entropies $h(Z|L = k)$ of the K Markov PDFs associated with the K stationary-ergodic regions. The present mutual-information-based energy gives more importance, or weight, to reducing entropies of larger regions in the image in direct proportion to their size – the weights are the probability of occurrence of the classes $P(L = k)$ in the image. Rewriting $I(L,Z) = h(L) - h(L|Z)$ provides more insight into this optimality metric. We see that the metric encourages segmentations with equal voxel counts for the classes (uniform PDF for L implying maximal $h(L)$) while demanding high predictability of the label at each voxel t given its neighborhood intensities $z(t)$ (low $h(L|Z = z(t))$ leading to low $h(L|Z)$).

Eqs. (4) and (5) give the optimal segmentation as:

$$\{T_k^*\}_{k=1}^K = \operatorname*{argmin}_{\{T_k\}_{k=1}^K} \left(\sum_{k=1}^{K} P(L=k)h(Z|L=k) \right) \tag{6}$$

$$= \operatorname*{argmin}_{\{T_k\}_{k=1}^K} \left(-\sum_{k=1}^{K} P(L=k) \times \int_{\Re^{|N_t|}} P_k(Z=z)\log P_k(Z=z)\,\mathrm{d}z \right). \tag{7}$$

Treating entropy as the expectation of negative log-probability and approximating the expectation, in turn, by the sample mean (Cover and Thomas, 1991), we get

$$\{T_k^*\}_{k=1}^K = \operatorname*{argmin}_{\{T_k\}_{k=1}^K} \left(-\sum_{k=1}^{K} P(L=k) E_{P_k(Z)}[\log P_k(Z)] \right) \tag{8}$$

$$\approx \operatorname*{argmin}_{\{T_k\}_{k=1}^K} \left(-\sum_{k=1}^{K} P(L=k) \frac{1}{|S_k|} \sum_{z \in S_k} \log P_k(Z=z) \right), \tag{9}$$

where S_k is a *random sample* (Dougherty, 1998; Stark and Woods, 2001) derived from the PDF $P_k(Z)$. Assuming ergodicity (Dougherty, 1998), in addition to stationarity, enables us to approximate ensemble averages using S_k with spatial averages using T_k. Hence, we have

$$\{T_k^*\}_{k=1}^K \approx \operatorname*{argmin}_{\{T_k\}_{k=1}^K} \left(-\sum_{k=1}^{K} P(L=k) \frac{1}{|T_k|} \sum_{t \in T_k} \log P_k(Z=z(t)) \right). \tag{10}$$

To estimate $P(L = k)$ from the data, we observe that the discrete random variable L can take only K possible values. Furthermore, $|T_k|$ voxels, out of a total of $|T|$ voxels, have $L(t) = k$. Thus,

$$P(L=k) = \frac{|T_k|}{|T|}. \tag{11}$$

Substituting Eq. (11) in Eq. (10) gives

$$\{T_k^*\}_{k=1}^K \approx \operatorname*{argmin}_{\{T_k\}_{k=1}^K} \left(-\frac{1}{|T|} \sum_{k=1}^{K} \sum_{t \in T_k} \log P_k(Z=z(t)) \right). \tag{12}$$

The probabilities $P_k(Z = z(t))$ are given by the Parzen-window density estimate in Eq. (3), i.e.

$$P_k(Z=z(t)) \approx \frac{1}{|A_t|} \sum_{u \in A_t} G_n(z(t) - z(u); \Psi_n), \tag{13}$$

where the set A_t is a small subset of T_k. Section 5.4 describes how to construct A_t, unique for each voxel t, to efficiently estimate the probability.

So far, we have not taken into account any *a priori* information in the segmentation process and we have derived all probabilities solely from the data. The formulation, however, extends in a straightforward manner to include *a priori* information using standard Bayesian strategies followed by optimization involving the resulting posterior probabilities. Section 5.3 discusses how to integrate *a priori* information in the form of brain-tissue probabilistic atlases into the proposed method. For the minimization in Eq. (12), we manipulate the regions T_k using an iterative gradient-descent optimization strategy, as discussed in Section 5.2.

5. MR-image brain-tissue classification

For brain-MR images, the goal is to segment the image into $K = 4$ regions corresponding to the (a) white matter, (b) gray matter, (c) cerebrospinal fluid, and (d) all other tissue types. This section starts by giving a high-level version of the proposed iterative classification algorithm along with an initialization strategy. It gives a few ways of incorporating *a priori* information in the probabilistic atlases into the proposed method. It describes the details of an efficient strategy for choosing the Parzen-window sample A_t, explains why the method performs reasonably well without explicit inhomogeneity correction, and describes a optimal data-driven choice of important internal parameters.

5.1. Initial classification using probabilistic atlases

The proposed classification algorithm seeks local optima of mutual information from an initial assignment of class labels, $\{T_k^0\}_{k=1}^K$. These labels must be sufficiently close to the solution to provide distinct density estimates for the different classes. For this, we use co-registered probabilistic atlases for the white matter, gray matter, and cerebrospinal fluid. We obtain these atlases from the ICBM repository (Rex et al., 2003), 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. We first register the average-T1 image to the data using an affine transformation and then use the transformation to resample the three probability images. The initialization is therefore:

(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 tth voxel belonging to the kth tissue type.
(3) Compute the probabilities for the class (say class $k = 4$) comprising all the nonbrain tissue types:

$$\forall t \in T : P_4^a(t) = 1 - \sum_{k=1}^{3} P_k^a(t). \tag{14}$$

(4) Assign the initial class labels:

$$\forall t \in T : L^0(t) = \operatorname*{argmax}_k P_k^a(t). \tag{15}$$

5.2. Classification algorithm

From the Markov PDFs, which are estimated from the initial classification, we reassign voxels based on optimizing the information content of the labels. We observe that the energy in Eq. (12) can be reduced if each voxel t is assigned to the class k that maximizes the probability $P_k(Z = z(t))$. This is an iterative process, where the Markov 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, the classification. We have found this to be an acceptable approximation, although some recent work (Jehan-Besson, 2002) introduces some additional terms in the update rule to avoid this lag.

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) For $k = 1,2,3,4$, $\forall t \in T$, estimate $P_k^m(Z = z(t))$ nonparametrically, as described in Section 4.
(2) Update the classification labels:

$$\forall t \in T : L^{m+1}(t) = \operatorname*{argmax}_k P_k^m(Z = z(t)). \tag{16}$$

(3) Stop upon convergence, i.e. when $\|L^{m+1} - L^m\|_2 < \delta$, where δ is a small threshold.

5.3. Bayesian classification using probabilistic-atlas driven priors

The registered, probabilistic atlas plays another role in the proposed classification algorithm. Instead of using data-driven probabilities alone for the classification updates, we can employ a Bayesian estimation strategy to compute the probabilities. The likelihood terms are the data-driven probabilities $P_k(Z = z(t))$ that we have computed via Parzen-window density estimation. The posterior is therefore the likelihood multiplied by the prior $P_k^a(t)$, which we derive from the probabilistic atlas.

For the proposed method, empirical evidence suggests that using the atlas directly as a prior can strongly dominate over the likelihood and introduce systematic biases in the classification (Pohl et al., 2004). 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 may be very much like the initialization. Section 6.2 discusses empirical results and the effect of different priors on the proposed method in more detail. Such behavior is likely an artifact from either the limited vari-

ability in the atlas, due to a limited population and construction, or the degree of misfit that remains after an affine registration. In practice, the prior strictly interpreted from the atlas is too strong, and we have investigated two ways of weakening its affect on the final solution. Section 6.2 (see Fig. 5) demonstrates the performance with both these priors.

One way of weakening the atlas prior is to use the atlas for discriminating only between two tissue types, namely the brain and nonbrain tissue. In this way, the prior does not interfere with the more subtle distinctions between the different brain tissues. For this, 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. This is equivalent to redefining $P_k^a(t), \forall t \in T$ as

$$\text{For } k = 1,2,3, \quad \forall t \in T : P_k^a(t) = 1 - P_4^a(t) \tag{17}$$

We call this the *2-class* prior.

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:

$$\text{For } k = 1,2,3, \quad \forall t \in T : P_k^a(t) = \frac{1-v}{4} + vP_k^a(t), \tag{18}$$

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

5.4. Parzen-window sampling and implicit models of inhomogeneity

This section discusses effective strategies for choosing the sample A_t during the Parzen-window density estimation of the probability $P_k(Z = z(t))$ of observation $z(t)$ at voxel t. We construct A_t as a small *randomly-chosen* subset of neighborhoods throughout T_k. 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 (Viola and Wells, 1995). MRI head images are not truly stationary and we have found that, in practice, image statistics are more consistent in proximate regions in the image than between distant regions, either because of piecewise stationarity or continuity in image statistics. To account for this, we use a *local* sampling strategy. In this local-sampling framework, for each voxel t, we define a unique sample A_t as a *random sample* (Dougherty, 1998; Stark and Woods, 2001) drawn from an isotropic 3D Gaussian PDF, defined on the image-coordinate space, with mean at the voxel t and variance $\sigma^2_{\text{spatial}}$. Thus, the sample A_t is biased and contains more voxels near the voxel t being processed. We have found that the method performs well for any choice of σ_{spatial} that encompasses more than several hundred voxels. The empirical results in Table 1 (shown later in Section 6.1) confirm that the performance of the proposed method degrades gracefully for suboptimal values of this parameter. For all of the results in this paper, we use $\sigma_{\text{spatial}} = 15$ voxels along each cardinal direction. This local-sampling strategy also plays an important role in implicit inhomogeneity handling. The local-sampling strategy enables the method to subsume the bias field in the estimated Markov statistics that determine the segmentation.

Table 1
The proposed method is fairly robust to changes in the values of the local-sampling Gaussian variance parameter and the Parzen-window σmultiplicative factor

	Gray matter	White matter
Local-sampling Gaussian standard deviation: σ_{spatial}		
10	0.9033	0.9386
15	0.9079	0.9427
20	0.9082	0.9422
25	0.9043	0.9368
Parzen-window σ *multiplicative factor*: α		
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

This table gives the Dice metrics for the BrainWeb T1 data with 5% noise and a 40% bias field.

5.5. Parzen-window kernel parameter

The nonparametric Parzen-window scheme for estimating Markov PDFs entails setting an appropriate value for the kernel-parameter σ. Section 3.2 described a maximum-likelihood based estimate for this parameter and discussed the theoretical advantages of such a strategy. A maximum-likelihood estimate for σ is equivalent to the choice that minimizes the entropy of the Markov statistics assuming the entire image was derived from a single stationary-ergodic random field. That is,

$$\begin{aligned}
&\operatorname*{argmax}_{\sigma} \left(\mathbf{\Pi}_{t\in T} P(Z = z(t); \sigma)\right) \\
&= \operatorname*{argmin}_{\sigma} \left(-\sum_{t\in T} \log P(Z = z(t); \sigma) \right) \\
&\approx \operatorname*{argmin}_{\sigma} \left(\sum_{z\in S_\sigma} [-\log P(Z = z; \sigma)] \right) \\
&= \operatorname*{argmin}_{\sigma} \left(E_{P(Z;\sigma)}[-\log P(Z; \sigma)] \right) \\
&= \operatorname*{argmin}_{\sigma} h(Z; \sigma),
\end{aligned} \tag{19}$$

where S_σ is a *random sample* (Dougherty, 1998; Stark and Woods, 2001) derived from the PDF $P(Z;\sigma)$, and $h(Z;\sigma)$ is the σ-dependent entropy of the random variable Z. Indeed, the relationship between log-likelihood and entropy is well-documented in the literature (Viola and Wells, 1995). We use the iterative Newton–Raphson optimization scheme (Rao, 1996) to find the optimal σ value.

The Parzen-window parameter σ, essentially controls the smoothing on the data in the feature space (seven-dimensional in our case) of neighborhood-intensity vectors. However, σ must be commensurate with the number and density of samples in that space, and thus it should adapt to different sampling strategies and applications. We have found that the optimal (maximum-likelihood) σ, estimated from limited data, does not properly "connect" all of the configurations of gray matter neighborhoods in feature space, thereby breaking the manifold into many distinct pieces prone to misclassification. In practice, to obtain desirable results with finite data, we impose a weak smoothness constraint on the Markov PDFs of each class, by multiplying the optimal σ by a factor α larger than unity. The choice of the precise value of this *multiplicative factor* α is not critical and Table 1 in the following section confirms that the algorithm is quite robust to small changes in α, i.e. α varying between 5 and 10. All of the results in this paper employ $\alpha = 10$.

5.6. Implementation issues

5.6.1. Neighborhood size and shape

In this paper, while working with three-dimensional MR data, we use a neighborhood comprising seven 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 isotropic. We incorporate such fuzzy weights by using an anisotropic feature-space distance metric, $\|z\|_M = \sqrt{z'Mz}$, where z' is the transpose of the vector z and M is a diagonal matrix with the diagonal elements being the appropriate weights on the influence of the neighbors on the center voxel. We select the weight for each neighbor to be reciprocal of the grid spacing along its associated axis.

5.6.2. Data-driven choice for the Parzen-window sample size

Section 5.5 described that we choose the maximum-likelihood (or, equivalently, minimal-entropy) based value of the Parzen-window Gaussian standard-deviation kernel-parameter σ. We have found (Awate and Whitaker, 2005; Awate and Whitaker, 2006) 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. Thus, given the Markov neighborhood and the local-sampling Gaussian variance σ_{spatial}, 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 5.6.2. The proposed method sets $|A_t|$, for all voxels t, to be about 500, based on the method explained in Section 5.6.2. The computation at each iteration is $O(K|A_t||T|)$, and the classification typically takes about 4–8 iterations depending on the noise/bias level. For $|A_t| = 500$, it takes about 45 min to process a $181 \times 217 \times 181$ volume on a single Pentium-IV 2.8 GHz processor. The algorithm scales linearly with the number of processors on a shared-memory, e.g. dual-processor Pentium, machine. The implementation in this paper relies on the Insight Toolkit (NLM Insight Segmentation and Registration Toolkit (ITK)).

Leemput et al. (1999b) use the Dice metric (Dice, 1945) to evaluate the classification performance of their state-of-the-art approach, based on expectation maximization and Gibbs/Markov priors on the segmentation labels. 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 classification for class k is $2|T_k^* \cap \tilde{T}_k|/(|T_k^*| + |\tilde{T}_k|)$, where the $|\cdot|$ operator gives the cardinality of sets.

6.1. Validation on simulated MR images

This section describes the behavior of 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 (Collins et al., 1998) with varying amounts of noise and bias field. Fig. 1 shows some data along with the classification and the ground truth.

We first show results 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. Figs. 2a and 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 (Leemput et al., 1999b). 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 caused by 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. Section 6.2 (see Fig. 5) describes that both priors perform equally well as measured by the average of the

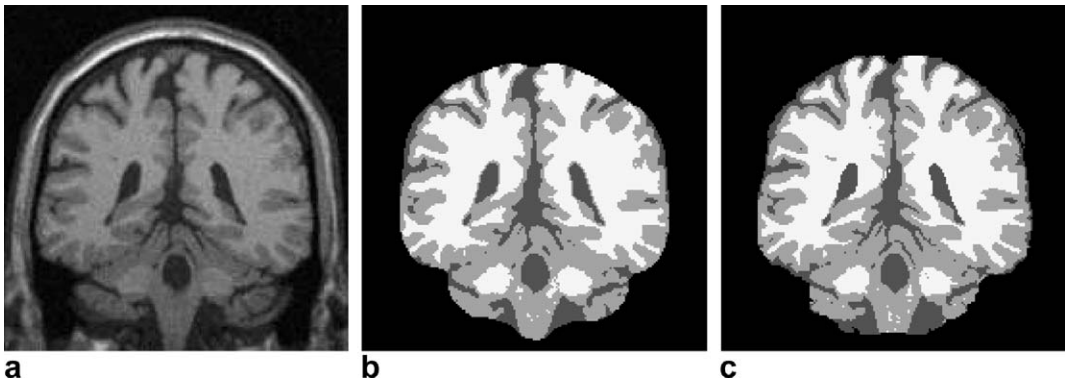

Fig. 1. Qualitative analysis of the proposed algorithm with BrainWeb data (Collins et al., 1998) with 5% noise and a 40% bias field: (a) a coronal slice of the data; (b) the classification produced by the proposed method and (c) the ground truth.

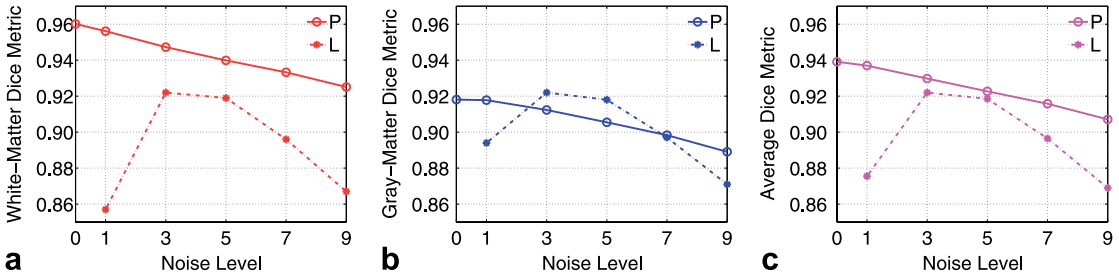

Fig. 2. Validation, and comparison with the state-of-the-art (Leemput et al., 1999b), on simulated T1-weighted data without any bias field and varying noise levels. Here, the proposed method uses the *2-class* prior. Dice metrics for: (a) white matter: D_{white}, (b) gray matter: D_{gray}, and (c) their 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 (Leemput et al., 1999b).

Dice metric for the white matter and gray matter, i.e. $(D_{\text{white}} + D_{\text{gray}})/2$.

Fig. 2c 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 (Leemput et al., 1999b), 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 fast-acquisition clinical MRI.

Fig. 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 (Leemput et al., 1999b). 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 (Leemput et al., 1999b). Hence, approaches enforcing Gaussian intensity PDFs on the classes, such as Leemput et al. (1999b) and Ruf et al. (2005), 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. Fig. 2 clearly depicts this advantage of the proposed method.

Strictly speaking, all methods trying to classify partial-volume voxels to one specific class are, in a way, fundamentally flawed. The proposed method, however, approaches this problem in a relatively more principled manner as compared to the EM-based method (Leemput et al., 1999b). A partial-volume voxel t comprising a larger contribution from tissue-class k will produce a $z(t)$ lying "closer" to the feature-space distribution of class k. The results show that the data-driven nonparametric estimation of all tissue-class PDFs, employing the same Parzen-window σ for each class, prevents any undesirable biases (unlike Leemput et al., 1999b) in the classification.

Fig. 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 quite well on biased BrainWeb MR data (Fig. 2). This is because of the adaptive model of Markov statistics underlying the method, as explained before in Section 5.4. To confirm the important role that the *local-sampling* Parzen-window density estimation strategy plays in enabling the automatic learning of the bias field, we perform two more experiments. In the first experiment, we use explicit bias correction with the proposed method (degree-4 polynomial fit, Leemput et al., 1999a, to the white matter intensities iteratively). Fig. 3 shows that this method performs approximately as well, but not significantly better

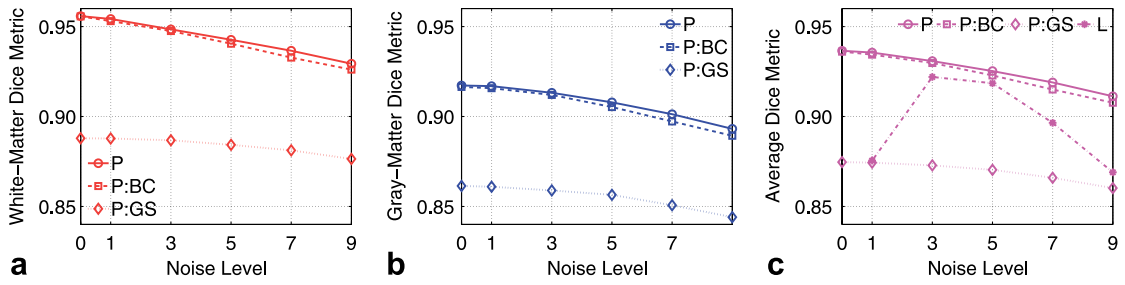

Fig. 3. Validation, and comparison with the state-of-the-art (Leemput et al., 1999b), on simulated T1-weighted data with 40% bias field and varying noise levels. We compare the performance by incorporating explicit bias correction and *global sampling: same sample size* $|A_t|$ (see text). Dice metrics 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: same sample size* $|A_t|$, L: Leemput et al.'s state-of-the-art method (Leemput et al., 1999b).

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 (with the same sample size $|A_t|$) uniformly over the image as was done in our previous work (Tasdizen et al., 2005). Fig. 3 shows that this scheme performs significantly worse at all noise levels in the absence of bias correction.

To study the sensitivity of the variance parameter $\sigma^2_{\text{spatial}}$ for the local-sampling Parzen-window Gaussian and the Parzen-window σ multiplicative factor α, we measure the Dice metrics for the white matter and gray matter over a range of parameter 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 5.6.2.

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. Fig. 4 shows the classification results for multimodal data using T1 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 those using T1 images alone.

6.2. Validation on real MR images

The section shows validation results with real expert-classified MR images. We obtained this data set from the IBSR website (IBSR). The data set comprises T1-weighted brain-MR images for 18 subjects. Fig. 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

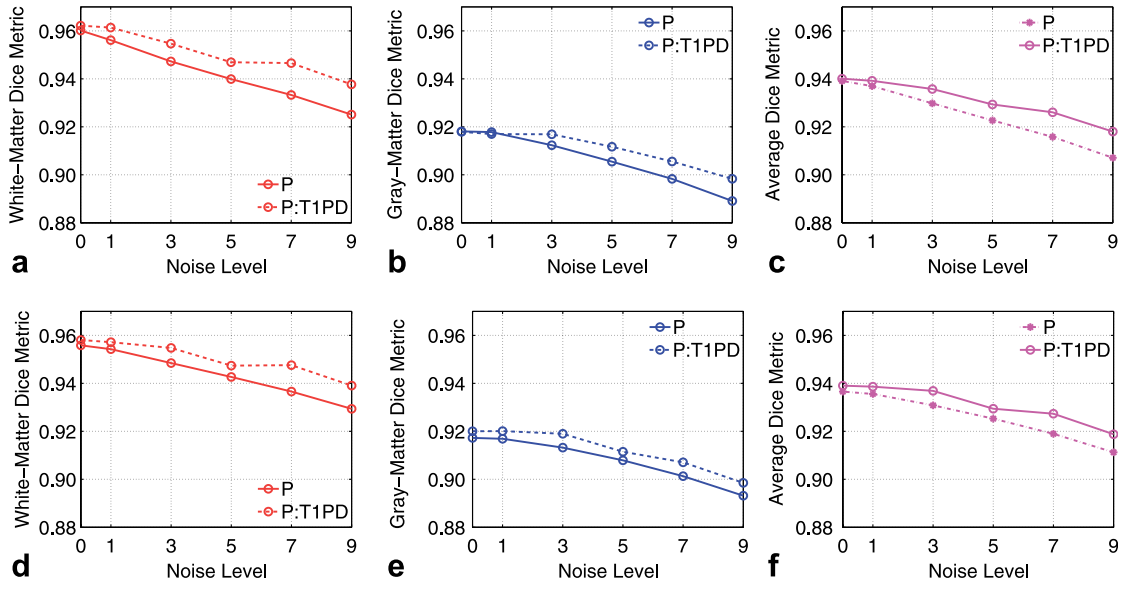

Fig. 4. Validation on simulated multimodal (T1 and PD) data with varying noise levels. Dice metrics for: (a) white matter: 0% bias, (b) gray matter: 0% bias, and (c) their average: 0% bias. Dice metrics for: (d) white matter: 40% bias, (e) gray matter: 40 % bias, and (f) their average: 40% bias. *Note*: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

Fig. 5. Qualitative analysis of the proposed algorithm with IBSR data (IBSR). The voxel size for this image is $0.9375 \times 0.9375 \times 1$(coronal): (a) an axial slice of the data; (b) the classification produced by the proposed method; and (c) the expert-classified ground truth.

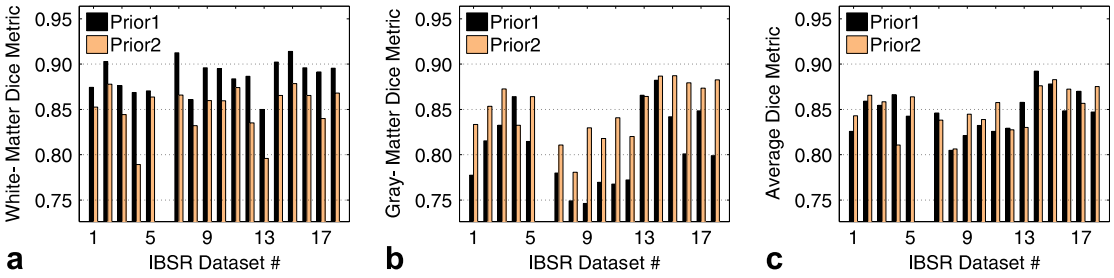

Fig. 6. Validation, of the proposed method with two different atlas-based priors, on IBSR data. Dice metrics for: (a) white matter: D_{white}, (b) gray matter: D_{gray}, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, Prior1: *2-class* prior, Prior2: *scaled-atlas*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

the BrainWeb dataset. Fig. 5 also shows an example of a classification generated by the proposed method and compares it to the ground truth.

Fig. 6 compares the performance of the proposed method using the two different atlas-based priors. Fig. 6a 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 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 Fig. 6b. 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, Fig. 6c 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 standard deviation for the Dice metrics over both white matter and gray matter classes, as compared to the results reported by Ruf et al. (2005) for Leemput et al.'s state-of-the-art method (Leemput et al., 1999b) as well as their own method.

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

7. Discussion and conclusions

This paper presents a novel method for unsupervised brain-MRI tissue classification by adaptively learning the image-neighborhood 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 tuning several important parameters, and therefore can operate on a moderately wide range of images without parameter retuning. Moreover, the proposed method does not rely on training for class intensities or neighborhood configurations, but adapts to the data given by an initial configuration that is generated from an atlas of labels. The adaptive image model enables the method to implicitly account for the bias field and perform reasonably well on biased MR-data without requiring explicit bias correction. By incorporating the information content in the neighborhoods in the classification process and imposing a weak smoothness constraint on the Markov statistics, the proposed method eliminates the need for explicit smoothness constraints on the class-label image.

The results in the paper empirically confirm that the piecewise stationary-ergodic Markov model conforms well to brain-MR images. It shows that it is possible to learn these models via nonparametric density estimation 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 mathematical and engineering components in this paper are appropriate for any kind of densely-sampled medical data, including vector-valued images (e.g. multimodal MR data) and images with higher-dimensional domains (e.g. a sequence of volumetric MR images over time).

The proposed method does face some limitations. For instance, results from previous work (Awate and Whitaker, 2006) on density estimation of image neighborhoods show that the this strategy fails for image features that do not occur with sufficient frequency. The proposed method might be further improved via some modeling and engineering advances. For instance, the use of single isotropic Parzen-window kernels is the simplest of such schemes. Parzen-window density estimation could improve, given sufficiently-large amounts of data, by choosing kernels adaptively to accommodate the signal or noise (Vincent and Bengio, 2002; Bengio et al., 2005). The Markov neighborhood in the current algorithm comprises only nearest neighbors. Using larger neighborhoods might also improve the results. However, this will entail significantly longer computation times, driven by the increased computation of distances in the higher-dimensional spaces and the larger number of samples needed to establish reliable statistics in those spaces. Improving the computational scheme, e.g. via parallelization or improved methods of density approximation (Yang et al., 2003), is an important area of future work.

Acknowledgments

This work was supported by NIH NCRR P41 RR12553-04, NSF EIA 0313268, NIH U01-AG024904, and the Louise Madsen Memorial Fund and the Rolan K. Schuholz Research Fund at the University of Michigan.

References

Awate, S.P., Whitaker, R.T., 2005. Higher-order image statistics for unsupervised, information-theoretic, adaptive, image filtering. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 44–51.

Awate, S.P., Whitaker, R.T., 2006. Unsupervised, information-theoretic, adaptive image filtering for image restoration. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 28 (3), 364–376.

Bengio, Y., Larochelle, H., Vincent, P., 2005. Non-local manifold parzen windows. In: Proceedings of the Advances in Neural Information Processing Systems.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statis. Soc., Ser. B 36 (2), 192–236.

Besag, J., 1975. Statistical analysis of non lattice data. J. Roy. Statis. Soc. 24, 179–195.

Besag, J., 1986. On the statistical analysis of dirty pictures. J. Roy. Statis. Soc., Ser. B 48, 259–302.

Cocosco, C., Zijdenbos, A., Evans, A., 2003. A fully automatic and robust brain MRI tissue classification method. Med. Image Anal. 7 (4), 513–527.

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

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

Craene, M.D., du Bois d'Aische, A., Macq, B., Warfield, S.K., 2004. Multi-subject registration for unbiased statistical atlas construction. In: Proceedings of the MICCAI, pp. 655–662.

Cuadra, M.B., Pollo, C., Bardera, A., Cuisenaire, O., Villemure, J., Thiran, J., 2004. Atlas-based segmentation of pathological MR brain images using a model of lesion growth. IEEE Trans. Med. Imaging 23 (10), 1301–1314.

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

de Silva, V., Carlsson, G., 2004. Topological estimation using witness complexes. In: Proceedings of the Symposium on Point-Based Graphics.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statis. Soc. B39, 1–38.

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

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

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

Efros, A., Leung, T., 1999. Texture synthesis by non-parametric sampling. In: Proceedings of the International Conference on Computer Vision, pp. 1033.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.

Geman, S., Graffigne, C., 1986. Markov random field image models and their applications to computer vision. In: Proceedings of the International Congress of Mathematicians, pp. 1496–1517.

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

Guillemaud, R., Brady, M., 1997. Estimating the bias field of MR images. IEEE Trans. Med. Imaging 16 (3), 238–251.

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

Internet Brain Segmentation Repository (IBSR). Available from: http://www.cma.mgh.harvard.edu/ibsr.

Jehan-Besson, S., Barlaud, M., Aubert, G., 2002. Dream2s: Deformable regions driven by an eulerian accurate minimization method for image and video segmentation. In: Proceedings of the European Conference on Computer Vision–Part III, pp. 365–380.

Kapur, T., Grimson, W.E.L., Wells, W.M., Kikinis, R., 1996. Segmentation of brain tissue from magnetic resonance images. Med. Image Anal. 1, 109–127.

Kim, J., Fisher, J.W., Yezzi, A.J., Cetin, M., Willsky, A.S., 2005. A nonparametric statistical method for image segmentation using information theory and curve evolution. IEEE Trans. Image Process., 14 (10), 1486–1502.

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

Lee, S., Vannier, M., 1996. Post-acquisition correction of MR inhomogeneities. Magn. Reson. Med. 36 (2), 275–286.

Leemput, K.V., Maes, F., Vandermeulen, D., Seutens, P., 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18, 885–896.

Leemput, K.V., Maes, F., Vandermeulen, D., Seutens, P., 1999b. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18, 897–908.

Levina, E., 1997. Statistical issues in texture analysis. Ph.D. Dissertation, Department of Statistics, University of California, Berkeley.

Li, S.Z., 1995. Markov Random Field Modeling in Computer Vision. Springer.

Lysaker, M., Lundervold, A., Tai, X., 2003. Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time. IEEE Trans. Imaging Process.

NLM Insight Segmentation and Registration Toolkit (ITK). Available from: http://www.itk.org.

Nocera, L., Gee, J., 1997. Robust partial volume tissue classification of cerebral MRI scans. In: Proceedings of the SPIE Medical Imaging: Image Processing, pp. 312–322.

Nowak, R., 1999. Wavelet-based rician noise removal for magnetic resonance imaging. IEEE Trans. Imaging Process. 8, 1408–1419.

Owen, A., 1989. Image segmentation via iterated conditional expectations. Technical Report, Department of Statistics, University of Chicago.

Pachai, C., Zhu, Y.M., Guttmann, C., Kikinis, R., Jolesz, F.A., Gimenez, G., Froment, J.-C., Confavreux, C., Warfield, S.K., 2001. Unsupervised and adaptive segmentation of multispectral 3d magnetic resonance images of human brain: a generic approach. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 1067–1074.

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

Parzen, E., 1962. On the estimation of a probability density function and the mode. Ann. Math. Stats. 33, 1065–1076.

Pohl, K., Grimson, W.E.L., Bouix, S., Kikinis, R., 2004. Anatomical guided segmentation with non-stationary tissue class distributions in an expectation-maximization framework. In: Proceedings of the International Symposium on Biomedical Imaging, pp. 81–84.

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

Prastawa, M., Gilmore, J.H., Lin, W., Gerig, G., 2004. Automatic segmentation of neonatal brain MRI. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 10–17.

Rajapakse, J., Giedd, J., Rapoport, J., 1997. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans. Med. Imag. 16 (2), 176–186.

Rao, S.S., 1996. Engineering Optimization, Theory and Practice. Wiley.

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

Rohlfing, T., Maurer Jr., Calvin R., 2004. Multi-classifier framework for atlas-based image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision Pattern Recognition., pp. 255–260.

Ruf, A., Greenspan, H., Goldberger, J., 2005. Tissue classification of noisy MR brain images using constrained gmm. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 790–797.

Scott, D.W., 1992. Multivariate Density Estimation. Wiley.

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

Stark, H., Woods, J.W., 2001. Probability and Random Processes with Applications to Signal Processing. Prentice Hall.

Tasdizen, T., Awate, S.P., Whitaker, R.T., Foster, N.L., 2005. MRI tissue classification with neighborhood statistics: a nonparametric, entropy-minimizing approach. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), vol. 3750, pp. 517–525.

Toga, A., 1999. Brain Warping. Academic Press.

ValdTs-Cristerna, R., Medina-Bauelos, V., Yez-Surez, O., 2004. Coupling of radial-basis network and active contour model for multispectral brain MRI segmentation. IEEE Trans. Biomed. Eng. 51 (3), 459–470.

Vincent, P., Bengio, Y., 2002. Manifold parzen windows. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 825–832.

Viola, P., Wells, W., 1995. Alignment by maximization of mutual information. In: Proceedings of the International Conference on Computer Vision, pp. 16–23.

Wei, L., Levoy, M., 2002. Order-independent texture synthesis. Stanford University Computer Science Department Technical Report TR-2002-01.

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

Yan, M., Karp, J., 1995a. An adaptive bayesian approach to three-dimensional MR brain segmentation. In: Proceedings of Information processing in Medical Imaging, pp. 201–213.

Yan, M., Karp, J., 1995b. Segmentation of 3D brain MR using an adaptive *k*-means clustering algorithm. In: Proceedings of the 1994 Nuclear Science Symposium and Medical Imaging Conference, pp. 1529–1533.

Yang, C., Duraiswami, R., Gumerov, N., Davis, L., 2003. Improved fast gauss transform and efficient kernel density estimation. In: International Conference on Computer Vision, pp. 464–471.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.