Comparison and Validation of Tissue Modelization and Statistical Classification Methods in T1-Weighted MR Brain Images

Meritxell Bach Cuadra, Leila Cammoun, Torsten Butz, Olivier Cuisenaire, *Member, IEEE*, and Jean-Philippe Thiran, *Senior Member, IEEE*

***Abstract*—This paper presents a validation study on statistical nonsupervised brain tissue classification techniques in magnetic resonance (MR) images. Several image models assuming different hypotheses regarding the intensity distribution model, the spatial model and the number of classes are assessed. The methods are tested on simulated data for which the classification ground truth is known. Different noise and intensity nonuniformities are added to simulate real imaging conditions. No enhancement of the image quality is considered either before or during the classification process. This way, the accuracy of the methods and their robustness against image artifacts are tested. Classification is also performed on real data where a quantitative validation compares the methods' results with an estimated ground truth from manual segmentations by experts. Validity of the various classification methods in the labeling of the image as well as in the tissue volume is estimated with different local and global measures. Results demonstrate that methods relying on both intensity and spatial information are more robust to noise and field inhomogeneities. We also demonstrate that partial volume is not perfectly modeled, even though methods that account for mixture classes outperform methods that only consider pure Gaussian classes. Finally, we show that simulated data results can also be extended to real data.**

***Index Terms*—Brain tissue models, hidden Markov random fields models, magnetic resonance imaging, partial volume, statistical classification, validation study.**

I. Introduction

ACCURATE and robust brain tissue segmentation from magnetic resonance (MR) images is a key issue in many applications of medical image analysis for quantitative studies and particularly in the study of several brain disorders such as Alzheimer's disease or Schizophrenia [1]–[4]. Moreover, brain tissue segmentation can also be required as preliminary step of image processing algorithms such as, for instance, voxel-based morphometry [5] or image registration [6]. Manual tracing by an expert of the three brain tissue types—white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF)—is exceedingly time consuming as the volume of data involved in magnetic resonance imaging (MRI) studies is large. On the other hand, automated and reliable tissue classification is a challenging task as the intensity representation of the data typically does not allow a clear delimitation of the different tissue types present in a MRI, because of partial volume (PV) effect, image noise and intensity nonuniformities caused by magnetic field inhomogeneities.

Numerous approaches have been proposed for MRI brain tissue classification. They can be divided into two main groups: supervised classification explicitly needs user interaction while nonsupervised classification is completely automatic. An exhaustive review of these classification methods is beyond the scope of this paper but we refer the interested reader to [7]–[9]. In this paper, we focus on statistical nonsupervised methods only. While this choice limits the scope of the paper, it allows us to create a homogeneous scenario in which we can compare the different hypotheses about the intensity distribution, the number of classes and the use of a spatial prior.

A. State-of-the-Art

Statistical classification methods usually solve the estimation problem by either assigning a class label to a voxel or by the estimation of the relative amounts of the various tissue types within a voxel [10]–[12]. *Finite Gaussian Mixture (FGM)* models, that assume a Gaussian distribution for the image intensities, are widely used and their parameter estimation problem is typically solved in an expectation-maximization (EM) framework [2], [13], [14]. Other algorithms [10], [15] add separate classes to take into account the PV voxels and model them also by independent Gaussian densities. A more realistic model of PV than Gaussian is proposed by Santago *et al.* [16], [17] and it is extensively used by other authors [11], [18]–[21]. However, some *finite mixture (FM)* models have the limitation of not considering the spatial information. That is why increasing attention has been paid recently to methods that model the spatial information by a Markov random field (MRF) [19], [22]–[25]. Finally, nonparametric classification techniques can be considered when no well justified parametric model is known [26], [27].

The assessment of brain tissue classification is a complex issue in medical image processing. Visual inspection and comparison with manual segmentation are labor intensive and not reliable since the amount of data to deal with is usually large.

Manuscript received April 29, 2005; revised August 15, 2005. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was F. Pernuš. *Asterisk indicates corresponding author.*

*M. B. Cuadra is with the Signal Processing Institute (ITS), Ecole Polytechnique Fédérale Lausanne (EPFL), CH-1015 Lausanne, Switzerland (e-mail: Meritxell.Bach@epfl.ch, http://lts5www.epfl.ch).

L. Cammoun, O. Cuisenaire, and J.-P. Thiran are with the Signal Processing Institute (ITS), Ecole Polytechnique Fédérale Lausanne (EPFL), CH-1015 Lausanne, Switzerland (e-mail: Leila.Cammoun@epfl.ch, http://lts5www.epfl.ch; Olivier.Cuisenaire@epfl.ch, http://lts5www.epfl.ch; JP.Thiran@epfl.ch, http://itswww.epfl.ch).

T. Butz is with ImaSys SA, PSE, CH-1015 Lausanne, Switzerland.

Digital Object Identifier 10.1109/TMI.2005.857652

Tissue classification methods can also be assessed by using synthetic data even if these kinds of images can hardly capture the complexity and the artifacts present in a MRI. There is however the possibility to validate brain tissue segmentation methods on a brain *simulated* data set as the one proposed by the *Brain Web* MR simulator [28], [29]. Their data is well-suited for this purpose since a ground-truth classification is known while different types of MR modalities and image resolution and artifacts can be reproduced.

Most of the abovementioned papers present a validation of the proposed approaches by classifying synthetic data, a phantom or real data. However, as far as we know, few validation studies comparing classification methods from different research groups have been published. For instance, Van Leemput *et al.* [21] presented a new statistical parametrical framework for PV segmentation as well as the validation on two-dimensional (2-D) multispectral simulated data. They performed a fuzzy classification instead of assigning a label to each voxel. Recently, Grau *et al.* [30] proposed an improved watershed method using prior information and they compare their approach for WM and GM segmentation with the methods of Van Leemput *et al.* [24] and Zeng *et al.* [31].

B. Goals of This Study

The goal of this work is to assess the robustness and accuracy of some of the most common tissue models and unsupervised classification methods. The work presented here is the continuation of [32]. Two main assumptions are made in this work. First, we consider that only 3-D T1-weighted MR brain images are available. This hypothesis creates a base line for later comparisons since classification methods will perform better if multispectral data (T1, T2, and PD weighted) is available. Moreover, this assumption is not unrealistic, since often only T1 is available for a concrete study and it undoubtedly a widely used modality. Second, no enhancement of the images is done neither before nor during the classification process.

The methods under study have been selected to cover the range of hypotheses made in the classification paradigm. The first method considers the finite Gaussian mixture model (FGMM). The second one adds to the FGMM a hidden MRF (HMRF) model to account for spatial prior information as in [23]. The third method models pure tissues by a Gaussian distribution but uses a specific PV distribution for mixture tissues. The fourth method adds to the previous one spatial interactions among voxels by means of a HMRF as in [10], [11]. The fifth algorithm does not model the tissue classes by parametric probability densities, but rather by nonparametric models [26]. The resulting algorithm minimizes an information theoretic quantity, called the error probability. The final method is also nonparametric, but again adds to the previous one a HMRF to model spatial prior information.

Various measures of the validity of the classification methods under consideration are presented for the simulated data [28]. We choose to focus primarily on the ability of the methods to correctly classify individual voxels. Later, we investigate how this ability reflects on global and local volumetric measurements. Classification is also performed on real data where a quantitative validation compares the methods' results with an estimated ground truth from manual segmentations by experts as proposed by Warfield *et al.* in [33]. While the scope of real data is limited, it allows us to show that conclusions drawn on simulated data can be extended to real data.

This paper is organized as follows. First, in Section II, the general theory used in this work for both intensity and spatial prior models is presented. Then, in Section III, the methods analyzed in this comparative study are summarized. In Section IV, the data set we use for this assessment study is presented. Next, in Sections V and VI, the validation method, the classification results on both simulated and real data are presented and discussed. Finally, conclusions and our current research are in Section VII.

II. Image Model

A. Intensity Distribution Model

Let us index N data points to be classified with $i \in \mathcal{S} = \{1, 2, \ldots, N\}$. In the case of 3-D images, such as MR images, they index the image voxels. Let us furthermore denote the observed data features by $y_i \in \mathbb{R}$. In the case of classification of single MR images, y_i represents the intensity of voxel i. Y is the random variable associated to the data features y_i, with the set of possible outcomes, $\mathcal{D}$. Any simultaneous configuration of the random variables, Y_i, is denoted by $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\} \in \mathcal{D}^N \subset \mathbb{R}^N$.

The classification process aims to classify the data $\mathcal{S}$ into one of the hidden underlying classes present in the image labeled by one of the symbols $\mathcal{L} = \{\mathrm{CSF}, \mathrm{GM}, \mathrm{WM}, \mathrm{CG}, \mathrm{CW}, \mathrm{GW}, \mathrm{CGW}\}$, where CG, CW, GW, and CGW are the mixtures of CSF+GM, CSF+WM, GM+WM, and CSF+GM+WM, respectively. The family of random variables X represents these classes. $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\} \in \mathcal{L}^N$ denotes a possible configuration of X. $\mathcal{X}$ is the space of all possible configurations.

Let us suppose that all the random variables, Y_i, are identically and independently distributed. Then, the probability density function of the voxel intensity is

$$P(y_i) = \sum_{\forall x_i \in \mathcal{L}} P(x_i) P(y_i \,|\, x_i). \tag{1}$$

where $P(x_i)$ is the prior probability of the tissue class x_i and $P(y_i \,|\, x_i)$ is the conditional probability density function of y_i given the tissue class x_i. The prior probability $P(x_i)$ is used to model the spatial coherence of the images in Section II-B. The transition probability $P(y_i \,|\, x_i)$ models the image intensity formation process for each tissue type. Different models are used for pure tissues and for tissue mixtures.

In what follows, we only consider stationary intensity models, for which we can simplify notations and write $P(y \,|\, x)$ instead of $P(y_i \,|\, x_i)$. The simplest model considers only the three pure tissues of the brain, with $\mathcal{L}_p = \{\mathrm{CSF}, \mathrm{GM}, \mathrm{WM}\}$. The probability density function of the observed intensity y for the pure tissue class $x \in \mathcal{L}_p$ is Gaussian, i.e.,

$$P(y \,|\, x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-(y-\mu_x)^2}{2\sigma_x^2}} \tag{2}$$

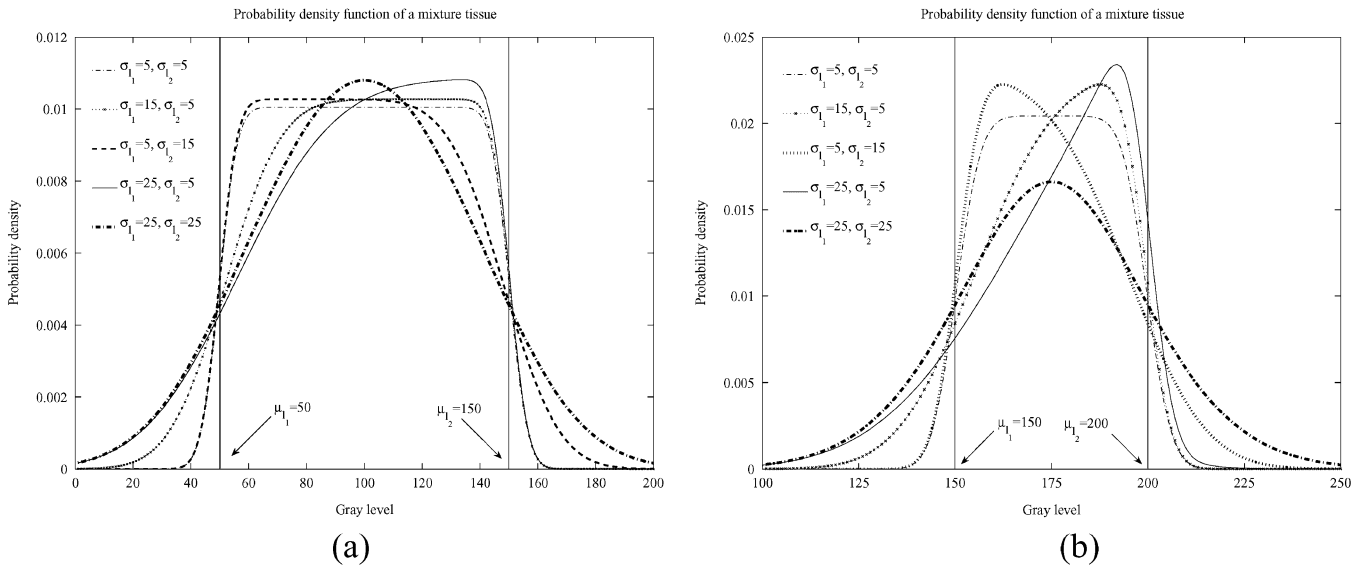


Fig. 1. Probability density function of a mixture tissue as described in (7), varying σ_{l_1} and σ_{l_2} with (a) $\mu_{l_1} = 50$ and $\mu_{l_2} = 150$ and (b) $\mu_{l_1} = 150$ and $\mu_{l_2} = 200$.

where the model parameters $\theta_x = \{\mu_x, \sigma_x\}$ are, respectively, the mean and standard deviation of the Gaussian, which is a good approximation of the Rician acquisition noise present in MR images at high signal-to-noise ratio (SNR). As in [17], [18], and [34], different tissues are assumed to have different noise variances. While this is not motivated by the physics of MR acquisition, it gives more flexibility to the model and allows it to adapt to other types of artifacts.

More evolved intensity models also consider the major tissue mixtures $\mathcal{L}_{\text{pm}} = \{\text{CSF}, \text{CG}, \text{GM}, \text{GW}, \text{WM}\}$. As assumed by most authors [21], the mixtures CW and CGW are not considered because they are so quantitatively insignificant that $P(\text{CW})$ and $P(\text{CGW})$ are not relevant in explaining $P(\boldsymbol{y})$.

Two different models of mixture tissues are considered in this paper. The simplest one assumes that the PV can be modeled by a Gaussian distribution as proposed in [15] and [23]. In this case, (2) is used both for pure and mixture tissues. This Gaussian mixture model is used in the methods described in Sections III-A and III-B.

A more complex probability density model for mixture tissues was proposed by Santago *et al.* [16] and used by [19] and [34]. This improved model is used in the methods described in Sections III-C and III-D. A mixture tissue voxel $x \in \mathcal{L}_{\text{pm}} \backslash \mathcal{L}_p$ has a probability density function

$$P(y \,|\, x, \alpha) = \frac{1}{\sigma_x(\alpha)\sqrt{2\pi}} e^{\frac{-(y-\mu_x(\alpha))^2}{2\sigma_x^2(\alpha)}} \tag{3}$$

where the two pure tissues composing the voxel are denoted by $l_1, l_2 \in \mathcal{L}_p$, and α is the fraction of l_1 present in the mixture voxel. The mean and variance of the mixture are determined by the model parameters of the pure tissues

$$\mu_x(\alpha) = \alpha\mu_{l_1} + (1-\alpha)\mu_{l_2} \tag{4}$$

$$\sigma_x^2(\alpha) = \alpha^2\sigma_{l_1}^2 + (1-\alpha)^2\sigma_{l_2}^2. \tag{5}$$

As discussed before, Santago [17] considers either a physically motivated common noise variance for all tissues or a more flexible model with a different noise level for each tissue. Once again we use the more flexible approach. The probability density function for the whole PV tissue $x \in \{\text{CG}, \text{GW}\}$ is

$$P(y \,|\, x) = \int_0^1 P(y \,|\, x, \alpha) P(\alpha \,|\, x)\, d\alpha. \tag{6}$$

As discussed by Ballester [20], choosing the correct function for $P(\alpha \,|\, x)$ is a complex issue. The true distribution is typically U-shaped, i.e., approximately uniform around $\alpha = 0.5$ with peaks at $\alpha = 0$ and $\alpha = 1$. Unfortunately, choosing this U-shape is not trivial and a wrong choice can lead to poor results. Hence, like most authors, we assume a uniform distribution for α, i.e., $P(\alpha \,|\, x_i) = 1$ for $0 \leq \alpha \leq 1$, this leads to

$$P(y \,|\, x) = \int_0^1 P(y \,|\, x, \alpha)\, d\alpha. \tag{7}$$

This integral has no known closed form and needs to be numerically computed. Its shape varies depending on the parameters $\theta_l = \{\mu_l, \sigma_l\}$, as illustrated in Fig. 1. It approaches a Gaussian when a high and identical variance of noise is assumed for both pure tissues. But in other cases, when noise variances are different, the probability density function of a mixture has an asymmetric bell shape.

Finally, it is also possible not to make any assumption on the shape of the probability density functions of each tissue class. Nonparametric, information theoretic alternatives are also considered in this work. The two such nonparametric approaches assessed in this comparative study were developed and implemented by Butz [26], [27]. Similar ideas can be found in [35].

For nonparametric classification, the posterior probability from parametric classification, $P(y\,|\,x)$, is replaced by an error probability which is defined as follows:

$$P_e = \sum_{x^{\text{est}} \in \mathcal{L}_{\text{pm}}} \sum_{y \in \mathcal{D}} \sum_{x \in \mathcal{L}_{\text{pm}}} P(E=1\,|\,x^{\text{est}}, x) \cdot P(x^{\text{est}}\,|\,y) \cdot P(y\,|\,x) \cdot P(x) \quad (8)$$

where x^{est} is a realization of random variable X^{est} which estimates X from Y. The probabilities $P(x^{\text{est}}\,|\,y)$ and $P(y\,|\,x)$ are estimated by Parzen-window probability density estimation, i.e.,

$$P(y\,|\,x) = \sum_{y_1 \in S_x} \frac{1}{|S_x|} G\left(y - y_1, \sigma_1^2\right) \quad (9)$$

and

$$P(x^{\text{est}}\,|\,y) = \sum_{y_2 \in S_{x^{\text{est}}}} \frac{1}{|y_2|} G\left(y - y_2, \sigma_2^2\right) \quad (10)$$

where $G(y-\mu, \sigma^2)$ is a Gaussian of expectation μ and variance σ^2, S_x denotes the set of voxels being classified into class x, $|S_x|$ is the number of elements of this set, and $|y|$ is the number of samples with intensity y. As in [26], a modified version of (10) is actually used to properly take into account the tails of the Gaussians beyond the range of values in $\mathcal{D}$. The probability $P(E=1\,|\,x^{\text{est}}, x)$ is called the distortion of the nonparametric classification algorithm, and is given by the following equation:

$$P(E=1\,|\,x^{\text{est}}, x) = (1 - \delta_{x^{\text{est}}, x}) \cdot \sum_{k \in \mathcal{L}_{\text{pm}}} \frac{1}{|S_k|}. \quad (11)$$

The final expression used for the estimation of the error probability, P_e, is the class probability, $P(x)$, and is given by $|S_x|/n$, n being the total number of voxels in the image.

B. Spatial Distribution Model

The other term in (1) is $P(x_i)$. It describes the prior knowledge about the spatial distribution of brain tissues in the image volume. The simplest spatial distribution model considers that for a given tissue class, the prior probability if constant over the image, i.e., $P(x_i) = \omega_{x_i}$. This model is used in Sections III-A, III-C, and III-E.

Alternatively, one can consider that the probability of having a given tissue at a given location varies, depending on the tissues found at the neighboring locations. In the methods of Sections III-B, III-D, and III-F, this is done by using a MRF to model spatial interactions among tissue classes [36], [37].

The sites in the image S are related with a neighborhood system $N = \{\mathcal{N}_i, i \in S\}$, where $\mathcal{N}_i$ is the set of sites neighboring i, with $i \notin \mathcal{N}_i$, and $i \in \mathcal{N}_j \Leftrightarrow j \in \mathcal{N}_i$. A random field $\boldsymbol{x}$ is a MRF on S with respect to N if and only if

$$P(\boldsymbol{x}) > 0, \quad \boldsymbol{x} \in \mathcal{X} \quad (12)$$

and

$$P\left(x_i\,|\,x_{S-\{i\}}\right) = P\left(x_i\,|\,x_{\mathcal{N}_i}\right), \quad (13)$$

where x_i denotes the tissue class at location i, and $x_{\mathcal{S}-\{i\}}$ denotes those at all the locations of $\mathcal{S}$ except at i. According to the Hammersley-Clifford theorem [38], [39], a MRF can be equivalently characterized by a Gibbs distribution

$$P(\boldsymbol{x}) = Z^{-1} e^{-U(\boldsymbol{x}, \beta)} \quad (14)$$

where $U(\boldsymbol{x})$ is the energy function, β the spatial parameter and Z a normalization factor. Let us briefly discuss how these parameters are chosen in the particular framework of image segmentation.

First, the choice of the energy function is arbitrary and there are several definitions of $U(x)$ in the framework of image segmentation. A complete summary of those can be found in [40] where a general expression of the energy function for pairwise interactions is denoted by

$$U(\boldsymbol{x}\,|\,\beta) = \sum_{\forall i \in \mathcal{S}} \left(V_i(x_i) + \frac{\beta}{2} \sum_{j \in \mathcal{N}_i} V_{ij}(x_i, x_j) \right). \quad (15)$$

where $V_i(x_i)$ is an external field that weighs the relative importance of the different classes present in the image and $V_{ij}(x_i, x_j)$ models the interactions between neighbors. In image segmentation [41], a simplified model with no external energy, $V_i(x_i) = 0$, is used. Only the local spatial transitions are taken into account and all the classes in the label image are considered equally probable. A typical definition of $V_{ij}(x_i, x_j)$ is the Potts model [23]

$$V_{ij}(x_i, x_j) = \delta(x_i, x_j) = \begin{cases} 1, & \text{if } x_i = x_j \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

which encourages the voxel to be classified like the majority of its neighbors. A more evolved function which takes into account the distance between neighbors and preserves thin structures is used in this paper, as proposed in [19] and [34]

$$V_{ij}(x_i, x_j) = \frac{\delta(x_i, x_j)}{d(i,j)} \quad (17)$$

where

$$\delta(x_i, x_j) = \begin{cases} -2, & \text{if } x_i = x_j \\ -1, & \text{if they share a tissue type} \\ +1, & \text{otherwise} \end{cases} \quad (18)$$

and $d(i,j)$ represents the distance between voxels i and j. With this energy function configurations that are not likely to occur, such as CSF inside WM, are penalized. On the other hand, smooth transitions, such as inserting a GW layer between WM and GM areas, are encouraged. The spatial parameter β controls the relative influence of the spatial prior over the intensity model. $\beta = 0$ corresponds to a uniform distribution over the $\mathcal{L}$ possible states so that only the conditional distribution of the observed data $P(y\,|\,x)$ is considered. On the other hand, with $\beta \to \infty$ the spatial information is dominant over the intensity information and one tends to classify all voxels to a single class [40].

TABLE I
METHODS UNDER STUDY

Method	Intensity model	Spatial model
A-FGMM	5 Gaussian	No spatial dependencies
B-GHMRF	5 Gaussian	MRF
C-GPV	3 Gaussian and 2 PVE	No spatial dependencies
D-GPV-HMRF	3 Gaussian and 2 PVE	MRF
E-EP	Non-parametric	No spatial dependencies
F-NP-HMRF	Non-parametric	MRF

The value of β is sometimes determined by maximum likelihood estimation although the complexity of the MRF model require the use of approximations [39]. β can also be determined empirically as proposed in [42] by gradually increasing its value through the algorithm iterations. In this paper, the value of β is fixed empirically to 1.2 by classifying a training set.

Finally, while the normalization factor of the Gibbs distribution is theoretically well-defined as

$$Z(U) = \sum_{\boldsymbol{x}} e^{-U(\boldsymbol{x},\beta)} \tag{19}$$

this requires a high computational cost. It may even be intractable since the sum among all possible configurations of $\boldsymbol{x}$ is usually not known [43]. Instead of computing Z, the conditional probabilities $P(x \,|\, x_{\mathcal{N}_i})$ are normalized by forcing

$$\sum_{\forall x_i \in \mathcal{L}_{\mathrm{pm}}} P\left(x_i \,|\, x_{\mathcal{N}_i}\right) = 1. \tag{20}$$

III. METHODS

Let us now describe with more details the classification methods considered for this comparative study. These methods, whose main hypotheses regarding the intensity and the spatial model are summarized in Table I, all consider 5 classes of tissues, i.e., $\mathcal{L}_{\mathrm{pm}} = \{\mathrm{CSF}, \mathrm{CG}, \mathrm{GM}, \mathrm{GW}, \mathrm{WM}\}$. The 3-classes methods considered in the later part of this paper are straightforward simplifications of these methods.

A. Finite Gaussian Mixture Model: FGMM

In the FGMM [13], each brain tissue in $\mathcal{L}_{\mathrm{pm}}$ is modeled by a Gaussian distribution and no spatial information is taken into account. The random variables X_i are assumed to be independent of each other, which means that, writing x instead of x_i to simplify notations

$$P\left(x \,|\, x_{\mathcal{N}_i}\right) = P(x) = w_x, \quad \forall x \in \mathcal{L}_{\mathrm{pm}}, \quad \text{and} \quad \forall i \in \mathcal{S}. \tag{21}$$

Then, the probability density function of the image intensity can be written as

$$P(y \,|\, \theta) = \sum_{\forall x \in \mathcal{L}_{\mathrm{pm}}} w_x \cdot P(y \,|\, x) = \sum_{\forall x \in \mathcal{L}_{\mathrm{pm}}} w_x \cdot f_x(y \,|\, \theta_x) \tag{22}$$

where the component densities $f_x(y \,|\, \theta_x)$ are Gaussian distributions defined by the parameters $\theta_x = (\mu_x, \sigma_x)$. The mixing parameters ω_x must also be included among the unknown parameters. The aim is to estimate the parameters $\theta = (\omega_x, \theta_x)$ under the constraint

$$\sum_{x \in \mathcal{L}_{\mathrm{pm}}} \omega_x = 1 \tag{23}$$

that maximize the log-likelihood function

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} P(\boldsymbol{y} \,|\, \theta). \tag{24}$$

One common solution to this optimization problem is the EM algorithm [44]. For Gaussian distributions, it goes as follows:

Initialization Step: Choose the best initialization for $\theta(\hat{0})$.

Expectation Step: Compute the *a posteriori* probabilities $\forall x \in \mathcal{L}_{\mathrm{pm}}$

$$\hat{P}^{(k)}(x \,|\, y_i, \hat{\theta}) = \frac{P\left(y_i \,\middle|\, \hat{\theta}_x^{(k-1)}\right) \hat{P}^{(k-1)}(x)}{\sum_{l, \forall l \in \mathcal{L}_{\mathrm{pm}}} P\left(y_i \,\middle|\, l, \hat{\theta}_l^{(k-1)}\right) \hat{P}^{(k-1)}(l)}. \tag{25}$$

Maximization Step:

$$\hat{\omega}_x^{(k)} = \hat{P}^{(k)}(x) = \frac{1}{N} \sum_{i \in S} \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta}) \tag{26}$$

$$\hat{\mu}_x^{(k)} = \frac{\sum_{i \in S} \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta}) y_i}{\sum_{i \in S} \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta})} \tag{27}$$

$$\left(\hat{\sigma}_x^{(k)}\right)^2 = \frac{\sum_{i \in S} \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta}) \left(y_i - \hat{\mu}_i^{(k)}\right)^2}{\sum_{i \in S} \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta})}. \tag{28}$$

Practically, the sum among all the image voxels of (25) can also be written

$$\sum_{i \in S} \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta}) = \sum_{\forall y_i} h(y_i) \hat{P}^{(k)}(x \,|\, y_i, \hat{\theta}), \tag{29}$$

where h is the image histogram [45]. This decreases significantly the number of computations in (26)–(28). This simplification is also used in the GPV method. Unfortunately, it cannot be adapted to the methods using the HMRF model where each voxel has to be treated with its neighborhood. Finally, once the optimal parameters have been found, classification is performed by choosing for each voxel the class that maximizes the posterior probability. Once again, this is simplified by finding the limits between $\omega_x P(y \,|\, x)$ on the image histogram and thresholding the image with these values.

B. Gaussian Hidden Markov Random Field Model: GHMRF

The second approach adds a Markovian spatial prior to the above method. The image intensity distribution function de-

pends on the parameter set θ and on the voxel neighborhood $x_{\mathcal{N}_i}$

$$P(y\,|\,\theta) = \sum_{x\in\mathcal{L}_{\rm pm}} P\left(x\,|\,x_{\mathcal{N}_i}\right)\cdot f_x(y\,|\,\theta_x) \tag{30}$$

where $f_x(y\,|\,\theta_x)$ is, $\forall x \in \mathcal{L}_{\rm pm}$, a Gaussian distribution parameterized by $\theta_x = \{\mu_x, \sigma_x\}$. $P(x\,|\,x_{\mathcal{N}_i})$ represents the locally dependent probability of the tissue class x_i. The optimal parameters are computed using an adapted version of the EM algorithm called HMRF-EM, as suggested in [23]. The equations for the maximization step are identical to those of A-FGMM, i.e., (26), (27), and (28). The expectation step becomes

$$\hat{P}^{(k)}(x\,|\,y_i,\hat{\theta}) = \frac{P\left(y_i\,\middle|\,\hat{\theta}_x^{(k-1)}\right)\cdot\hat{P}^{(k-1)}\left(x\,|\,x_{\mathcal{N}_i}\right)}{\sum_{l,\forall l\in\mathcal{L}_{\rm pm}} P\left(y_i\,\middle|\,l,\hat{\theta}_l^{(k-1)}\right)\hat{P}^{(k-1)}\left(l\,|\,l_{\mathcal{N}_i}\right)}. \tag{31}$$

In this equation, the term $P^{(k-1)}(x\,|\,x_{\mathcal{N}_i})$ requires a previous estimate of the class labels $\hat{\mathbf{x}}$. Therefore, the classification step needs to be performed at each iteration of the EM algorithm, which becomes

1) Estimate the image labeling $\hat{\mathbf{x}}$ given the current θ, then use it to form the complete data set $\{\hat{\mathbf{x}}, \mathbf{y}\}$.
2) Estimate a new θ by maximizing the expectation of the complete-data log likelihood, $\mathcal{E}[\log P(\mathbf{x},\mathbf{y}\,|\,\theta)]$.

Note that, as detailed in [46], the estimation of $\hat{\mathbf{x}}$ can be simplified by minimizing the energies instead of maximizing the probabilities.

C. Gaussian and Partial Volume Model: GPV

The third approach relies only on the intensity information. Pure tissue intensities are modeled by Gaussian distributions while mixture tissues are modeled as proposed by Santago *et al.* [16], [17] and described by (7). $P(y,\theta)$ is defined by (22) where $P(y_i\,|\,x,\theta_x)$ is either a Gaussian or a PV equation. The optimal parameters are found by minimizing the square difference between observed normalized intensity histogram h_n and the intensity model $p(y_i\,|\,\theta)$ of (22), i.e.,

$$\hat{\theta} = \min_{\theta}\sum_{\forall y_i}(h_n(y_i) - p(y_i\,|\,\theta))^2 \tag{32}$$

where the list of parameters to be optimized is

$$\theta = \{\omega_{\rm CSF},\omega_{\rm CG},\omega_{\rm GM},\omega_{\rm GW},\omega_{\rm WM},\mu_{\rm CSF},\mu_{\rm GM},\mu_{\rm WM},\sigma_{\rm CSF},\sigma_{\rm GM},\sigma_{\rm WM}\}. \tag{33}$$

This model has fewer parameters than A-FGMM since the mean and variance of the PV distributions are determined by the mean and variance of the neighborhood pure tissues composing the mixture. As in [13], a *genetic algorithm* is used to solve the estimation problem (see Section III-G). Finally, the classification is performed by maximazing the MAP criteria, similarly to the A-FGMM approach.

D. GPV and HMRF Model: GPV-HMRF

This method adds a MRF prior to the C-GPV approach. The resulting probabilistic model is the same as (30), with $f_x(y\,|\,\theta_x)$ defined either as a Gaussian for pure tissues or by the PV equation (7).

The parameter optimization is performed similarly to the algorithm for B-GHMRF. The modified EM-algorithm becomes, as in [11]:

$$\hat{P}^{(k)}(x\,|\,y_i,\hat{\theta}) = \frac{\hat{P}\left(y_i\,\middle|\,x,\hat{\theta}_x^{(k-1)}\right)\cdot\hat{P}^{(k-1)}\left(x\,|\,x_{\mathcal{N}_i}\right)}{\sum_{l,\forall l\in\mathcal{L}_{\rm pm}}\hat{P}\left(y_i\,\middle|\,l,\hat{\theta}_l^{(k-1)}\right)\hat{P}^{(k-1)}(l\,|\,l_{\mathcal{N}_i})} \tag{34}$$

$$\mu_x^{(k)} = \frac{\sum_{i\in S}\hat{P}^{(k)}(x\,|\,y_i)\,y_i}{\sum_{i\in S}\hat{P}^{(k)}(x|y_i)} \tag{35}$$

$$\left(\sigma_x^{(k)}\right)^2 = \frac{\sum_{i\in S}\hat{P}^{(k)}(x\,|\,y_i)\left(y_i-\mu_x^{(k)}\right)^2}{\sum_{i\in S}\hat{P}^{(k)}(x\,|\,y_i)}. \tag{36}$$

In this approach, (35) and (36) are only computed for pure tissues $x \in \mathcal{L}_p$. Besides $\hat{P}^{(k)}(y_i\,|\,x,\hat{\theta_x})$ in (34) is either a Gaussian or a PV distribution depending on the tissue type. As for B-GHMRF, the term $P^{(k-1)}(x\,|\,x_{\mathcal{N}_i})$ requires a previous estimate of the classification result $\hat{\mathbf{x}}$. In [11], this is done through

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathcal{X}}\{P(\mathbf{y}\,|\,\mathbf{x})\}. \tag{37}$$

Here, we do

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathcal{X}}\{P(\mathbf{y}\,|\,\mathbf{x})P(\mathbf{x})\}. \tag{38}$$

Contrarily to B-GHMRF, this expression cannot be handled at the energy level [46] and the optimization has to be performed on the probabilities, because $P(\mathbf{y}\,|\,\mathbf{x})$ does not always follow a Gaussian distribution.

E. Error Probability Minimization: EP

The last two approaches do not consider a parametric model for the image intensities, but instead apply an information theoretic framework to both the image formation process and the classification as in [26]. Let us consider a random variable different from X, called $X^{\rm est}$, also over $\mathcal{L}$, which models an estimation of X from the observable data, Y. Naturally, the following stochastic process can be built

$$X \to Y \to X^{\rm est} \to E \tag{39}$$

where E is an error random variable being 1 whenever the estimated class label $x^{\rm est}$ is considered a wrong estimate of the initial class label, x, and 0 otherwise. A key quantity of (39) is the probability of error, $P_{e\,|\,\mathbf{x}}$, of the transmission from X to $X^{\rm est}$, for a given class map $\mathbf{x}$. This probability also equals the expectation of E. Then, the classification objective consists of

determining the class label map $\hat{\mathbf{x}}$ that minimizes an error probability $P_{e|\mathbf{x}}$

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} P_{e|\mathbf{x}}. \tag{40}$$

F. Nonparametric HMRF: NP-HMRF

The probabilistic nature of the above method allows us to add a HMRF spatial prior as before. This results in a nonsupervised nonparametric hidden Markov model (F-NP-HMRF) segmentation

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} P(\mathbf{x}) P_{e|\mathbf{x}}. \tag{41}$$

The optimization objective above is called the minimal error probability principle for F-NP-HMRFs. In complete analogy to parametric HMRFMs, the prior probabilities, $P(\mathbf{x})$, are modeled by a Gibbs distribution (Section II-B). The derived nonparametric framework for classification allows the consideration of voxel features for which no particular parametric model is known. Here, the only feature is the voxel intensity though voxel gradient could be also used as in [27].

G. Practical Implementation

1) Initialization and Settings: Because of the local nature of the EM algorithm, a proper initialization is obviously required, as discussed in [23], [47] for instance. Among the parameters $\theta(0) = \{\omega_t, \mu_t, \sigma_t\}$, with $t \in \mathcal{L}_{\mathrm{pm}}$, the most sensitive appears to be the means μ_t. Those are estimated using a prior k-means classification. The other parameters are set to standard values, i.e., $\omega_t = 1/5$ (because our model has 5 classes) and $\sigma_t = 5$ (a small value different from zero). In addition, the methods using a MRF require an initial estimate of the voxel classification since the MAP is solved using the ICM labeling algorithm that converges locally. Actually, we assume that initial label map is close to the global optimal solution. For this purpose we use the output of FGMM, GPV, and EP to, respectively, initialize GHMRF, GPV-HMRF and NP-HMRF.

A genetic algorithm is used for the parameter optimization of GPV. In this approach, no initial values have to be determined but an optimization space has to be defined: $\omega_t \in [0,1], \mu_t \in [\min(h), \max(h)]$ ranging from the minimum to the maximum intensity value of the image histogram h, and finally $\sigma_t \in [1, 0.4 * (\max(h) - \min(h))]$. The number of chromosomes is set to 11. The evolution strategy is described in [13].

2) Computation Time: All parametric methods are implemented in MATLAB and nonparametric algorithms are in C++. They all run on a Pentium 4, CPU 1.8 Ghz, 764 MB of RAM. The total computing time on an image of $161 \times 187 \times 161$ voxels is around seconds for FGMM and GPV, few minutes for EP and GHMRF, around 20 minutes for GPV-HMRF and between one and two hours for NP-HMRF.

IV. Data Set

A. Simulated Data

The main dataset used in this study comes from the digital brain phantom[1] from McConnell Brain Imaging Center [28]. It consists of a realistic anatomical brain model and of a MRI simulator. The brain model has fuzzy tissue membership volumes where voxel values reflect the proportion of a given tissue within the voxel. It was generated through the semi-manual classification of a very high SNR MRI of a normal subject obtained through repeated acquisitions. The MRI simulator uses this anatomical model and a model of MR acquisition physics to generate images where different RF nonuniformity (bias of 0%, 20%, and 40%) and noise levels (0%, 1%, 3%, 5%, 7%, and 9%) can be added. All the methods have been applied to the whole range of noise and RF levels on the T1-weighted modality. The volume is $217 \times 181 \times 217$ voxels with isotropic 1 mm voxel size. In Fig. 2(a)–(c), three MR images simulated with different levels of noise and inhomogeneities are shown.

For the purpose of this study, a 5-class (CSF, CG, GM, GW, and WM) ground truth classification, Fig. 2(d), was created from the 3-D fuzzy tissue membership volumes. Finally, a ground truth image histogram was computed by splitting each image histogram into the specific pure tissue and their mixture histograms [see Fig. 6(a)].

B. Real Data

While simulated data provides an excellent tool to validate and compare method performance in presence of a variety of artifacts, assessment on real data is ultimately needed since the final purpose of these methods is to classify a real T1w MRI of the human brain for a concrete application. For instance, the study of term and preterm neonates [48]–[50] requires the acquisition of newborn MRIs; the detection, quantification and study of brain disorders is based on MR images of pathological brains [4], [51]; the study of brain aging deals with MRI of aged persons [52]; finally, normal brains are needed to perform statistical studies or create probabilistic atlases [53]. The trouble with real data is that the ground truth is typically not available, or excessively time consuming to generate manually.

In this paper, we consider a single real MR brain image of a normal brain (female adult, no pathology): a three–dimensional (3-D) T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (Siemens Vision®, 1.5 T, Erlangen, Germany) TR 9.7 ms, TE 4 ms, FOV 280×280, matrix 256×256, 146 slices, $0.98 \times 0.98 \times 1.25$ mm^3. Its signal to noise ratio (SNR) and coefficient of joint variations (CJVs) were measured at 18 dB and 0.66, respectively. This corresponds to a digital phantom image with values $\mathrm{N} = 7\%$ and RF between 0% ($\mathrm{SNR} = 16.7$ dB, $\mathrm{CJV} = 0.59$) and 20% ($\mathrm{SNR} = 17.4$ dB, $\mathrm{CJV} = 0.58$).

C. Ground Truth for Validation on Real Data

Manual segmentations were performed for 2 slices: slice 1 contains mostly GM and WM [see Fig. 3(a)], while slice 2 includes the central nuclei and ventricles [see Fig. 3(c)]. These

[1]In this paper, the word phantom stands for a digital synthetic MR brain image where different artifacts can be added. We do not refer to a physical phantom.

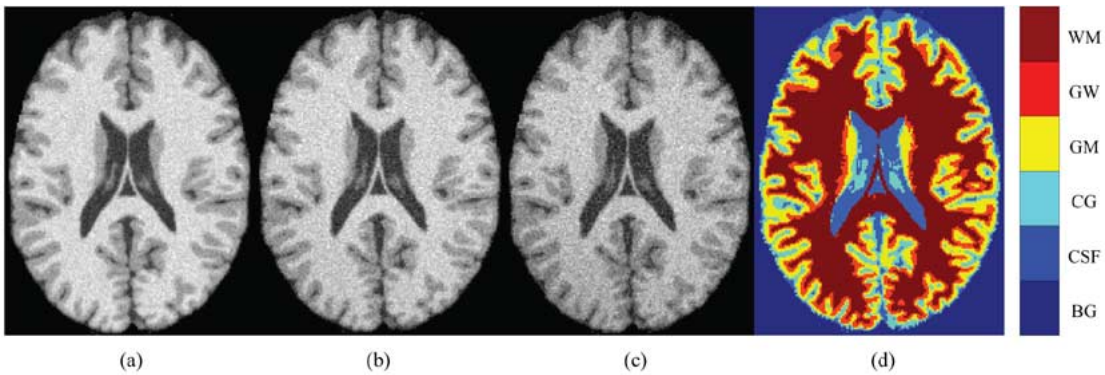


Fig. 2. Simulated brain T1-weighted: (a) 5% noise and 0% RF, (b) 7% noise and 20% RF, (c) 9% noise and 40% RF, and (d) 5 classes ground truth created from Brainweb classification. Colorbar: background (BG) is in dark blue, CSF is in blue, mixture of CSF and GM (CG) is in light blue, GM is in yellow, mixture of GM and WM (GW) is in red, and WM is in dark red.

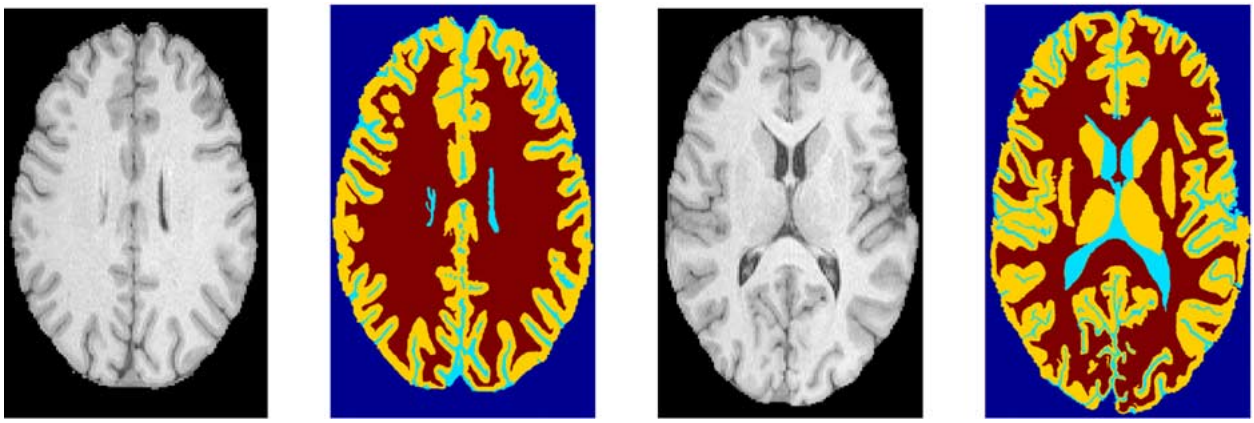


Fig. 3. Real MRI data and manual segmentations at a high image resolution level. CSF is in light blue, GM is in yellow and WM is in dark red (a) Slice 1, (b) Manual segmentation (c) Slice 2 (d) Manual segmentation.

slices were oversampled 8 times, then five experts manually segmented them into pure tissue classes (CSF, GM, WM or background), as illustrated in Fig. 3(b) and (d) for one of the experts.

The ground truth high-resolution CSF, GM, and WM masks were estimated using Warfield's STAPLE algorithm [33], [54]. It generates a probabilistic estimate $W_t(i)$—where t is the tissue $\{\mathrm{CSF}, \mathrm{GM}, \mathrm{WM}\}$ and i is the voxel position—of a ground truth T_i, from a group of expert segmentations D_j, where $j = \{1, \ldots, N_{\mathrm{experts}}\}$. At the same time, two measures of quality for each expert segmentation are also estimated: the sensitivity $\hat{p_j}$ and the specificity $\hat{q_j}$,

$$p_j = \Pr(D_j = 1 \,|\, T_i = 1)$$

and

$$q_j = \Pr(D_j = 0 \,|\, T_i = 0). \tag{42}$$

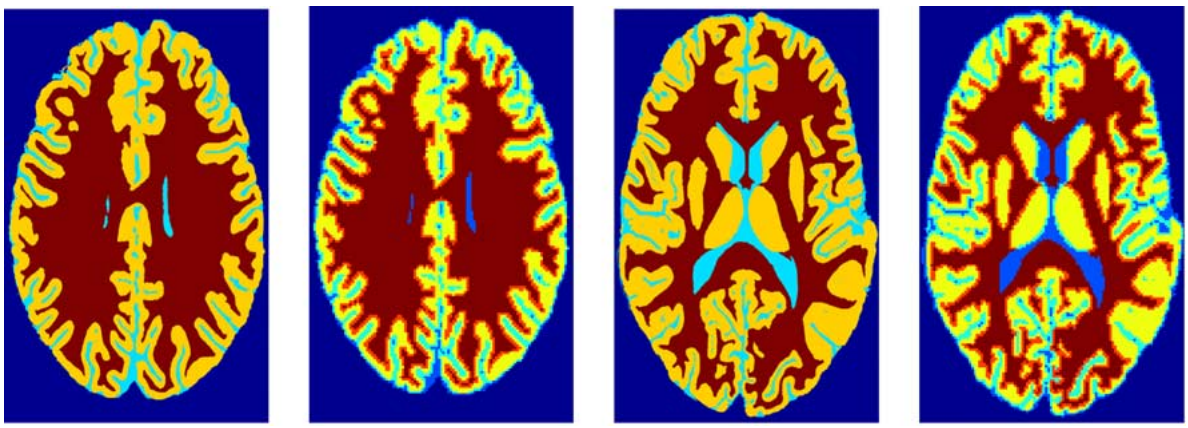
The parameters $p_j, q_j \in [0, 1]$ are characteristic of rater j. Initially they are fixed to $\{p, q\} = \{0.9, 0.9\}$. The ground truth prior probabilities are assumed to be 0.05, 0.25, and 0.3 for CSF, GM, and WM, respectively, as suggested in [33]. The ground truth estimate and rater performances are computed iteratively within an expectation maximization (EM) framework. The algorithm stops at iteration k when $C_k - C_{k-1} = 0$, with $C_k = \sum_{i=1}^{N} W_t(i)$. Convergence is usually reached with less than 15 iterations. The final 3 class ground truth estimated are shown in Fig. 4(a) and (c).

The 3 class high-resolution ground truth is downsampled back to the original resolution, as can be seen at Fig. 4(b) and (d). Each pixel at the lower resolution corresponds to a group of high resolution pixels. If this group consists of a single tissue class, the low resolution pixel is a pure tissue. If this group includes several tissue classes, then the low resolution pixel is PV. Eventual pixels mixing CSF and WM are removed manually. Let us note that this technique to generate the ground truth only creates PV at the interface between pure tissues. In particular, the thalamus or the caudate nuclei at Fig. 4(c) are classified as pure GM.

The estimated quality parameters of each expert segmentation $\{\hat{p_j}, \hat{q_j}\}$ and the Dice Similarity Measure (see its definition in Section V) with respect to the estimated ground truth for slice 2, Fig. 4(c), are shown in Table II. All $\hat{p_j}$ and $\hat{q_j}$ values are high (between 0.85 and 1) except for the CSF where large variability of the experts segmentations is shown (see for instance $\hat{p_3}, \hat{p_4}$, and $\hat{p_5}$).

V. Validation Methods

The data described above allows us to compute many different measures of the validity of the various classification methods under consideration. We choose to focus primarily on



(a) (b) (c) (d)

Fig. 4. Estimated ground truth from manual segmentations: (a) and (c) are, respectively, the 3 classes STAPLE estimates of slices 1 and 2. (b) and (d) are, respectively, the 5 classes ground truth images at the original image resolution level of slices 1 and 2. CSF is in blue, CG is in light blue, GM is in yellow, GW is in red, and WM is in dark red.

TABLE II
QUALITY PARAMETERS AND DICE SIMILARITY MEASURE OF EACH EXPERT SEGMENTATION WITH RESPECT TO THE 3 CLASSES STAPLE GROUND TRUTH OF SLICE 2

Tissue	Expert $\{\hat{p};\hat{q};\text{DSM}\}$				
	1	2	3	4	5
CSF	0.836;0.974;0.68	0.848;0.988;0.73	0.749;0.993;0.73	0.623;0.999;0.76	0.682;0.996;0.74
GM	0.905;0.968;0.87	0.887;0.972;0.87	0.859;0.976;0.85	0.944;0.946;0.88	0.917;0.968;0.90
WM	0.928;0.979;0.93	0.933;0.986;0.94	0.968;0.959;0.93	0.911;0.983;0.92	0.951;0.974;0.94

the ability of the methods to correctly classify individual voxels, both for simulated and real data. Later, we investigate how this ability reflects on global and local volumetric measurements.

A. *Classification of Simulated Data*

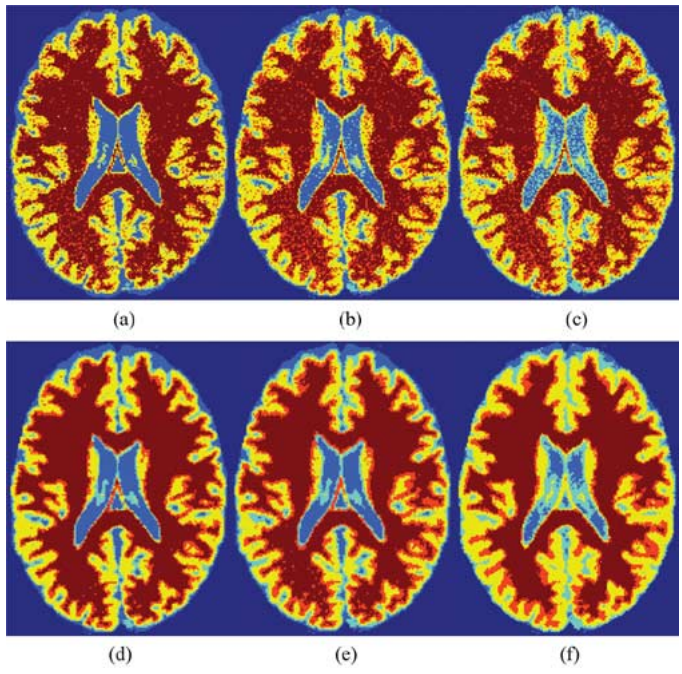
In order to assess the methods presented in Section III, their results are compared to the ground truth classification and to the histograms of the simulated MR brain images. Because of limited space, most results are only shown for images with 7% Noise (N) and 20% of in-homogeneity (RF), noted 7N20RF. The same results for images with 5N0RF and 9N40RF are presented in [46]. The comparison is performed in 5 different ways.

First, each of the volumes classified by each of the algorithms is visually assessed. A comparison of a representative slide of the resulting classified images where all brain tissues are present with the corresponding slide of the ground truth classification volume is presented for 7N20RF in Fig. 5.

Second, in Fig. 6, the intensity image model is assessed by comparing the histogram fitting to the ground truth histogram of 7N20RF.

Third, global measures of quality are represented by the percentage of voxels correctly classified (called *pergood*). This value is computed with respect to the ground truth volume and background voxels are not considered. This synthetic quality measure allows us in Fig. 7 to compare all methods in terms of robustness with respect to the full range of noise and inhomogeneities.

Fourth, in Table III, a more detailed tissue dependent quantitative analysis is performed by computing the confusion tables between the ground truth and the classification results for 7N20RF. This table also includes false positive (FP) and false negative (FN) percentages for all tissue classes.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Nonsupervised classification of the brain digital phantom with 7% noise and 20% RF. First row, methods using intensity information only: (a) A-FGMM, (b) C-GPV, and (c) E-EP. Second row, methods that add to the intensity the spatial prior information: (d) B-GHMRF, (e) D-GPV-HMRF, and (f) F-NP-HMRF. Background is in dark blue, CSF is in blue, CG is in light blue, GM is in yellow, GW is in red, and WM is in dark red.

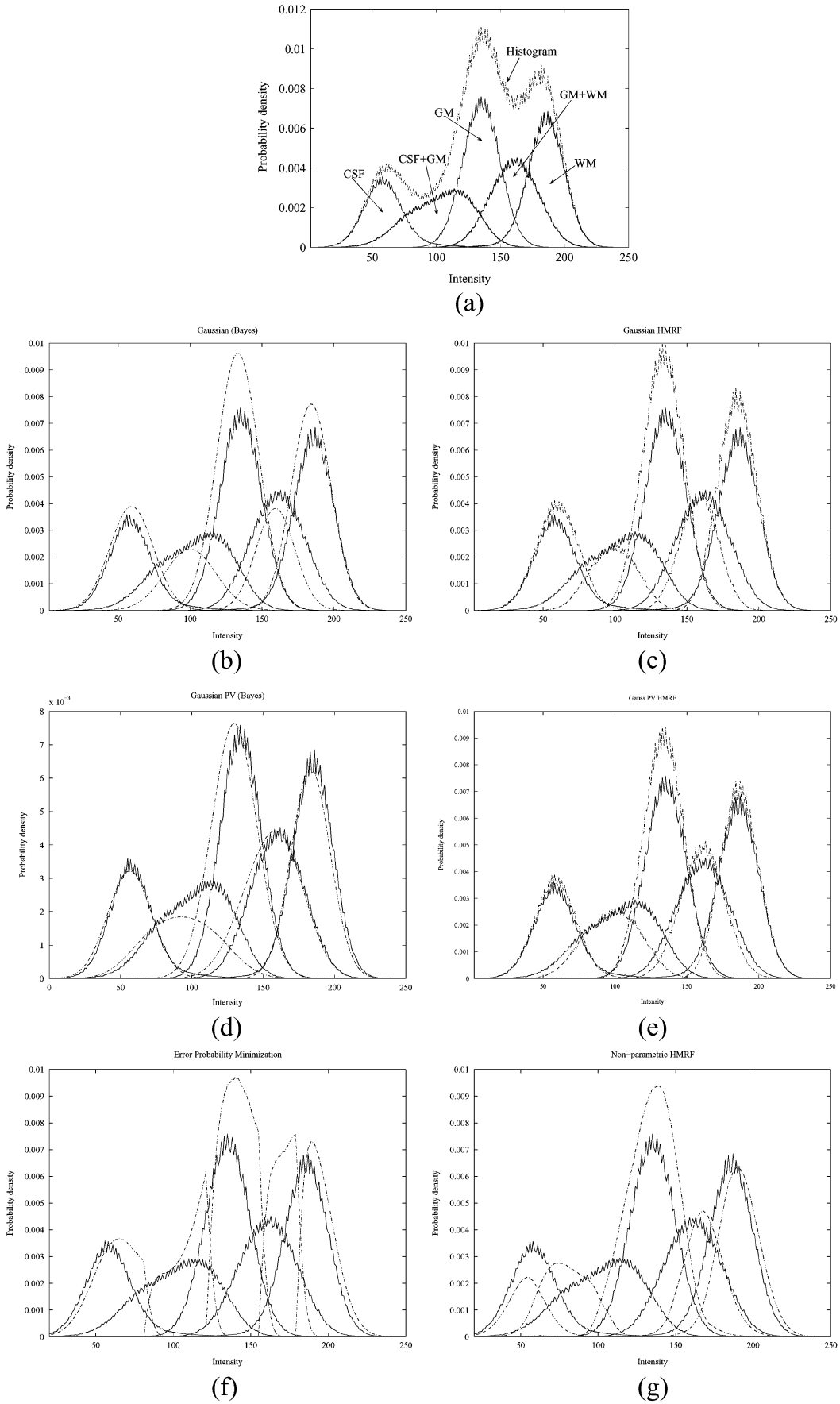


Fig. 6. Analysis of the probability density functions for the phantom 7N20RF: (a) 5 classes ground truth histogram and tissue distributions, and from (b) to (g) Histogram fitting (ground truth is in solid line and estimated probability density functions are in dotted line). (a) 5 classes ground truth probability density functions. (b) Method A: FGMM. (c) Method B: GHMRF. (d) Method C: GPV. (e) Method D: GPV-HMRF. (f) Method E: EP. (g) Method F: NP-HMRF.

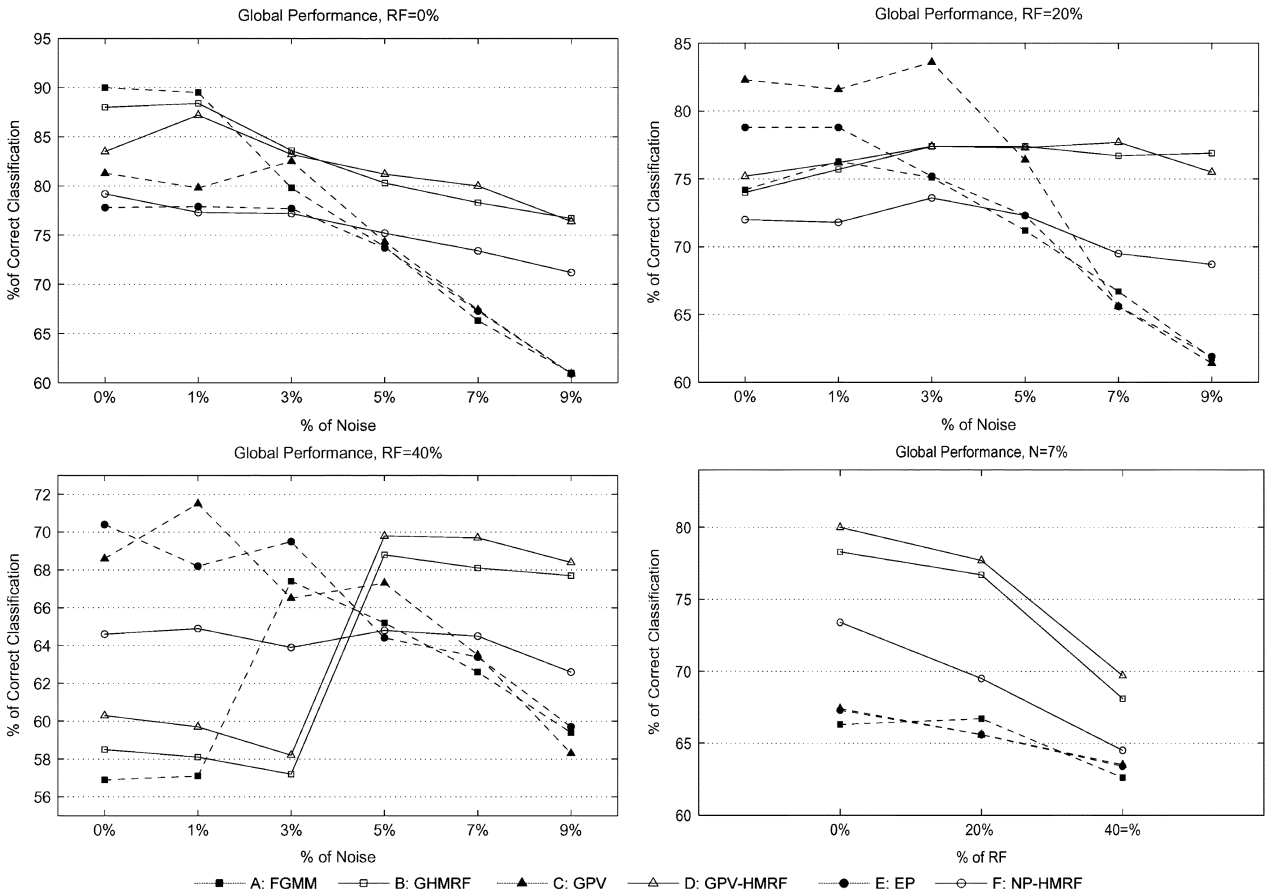


Fig. 7. Percentage of voxels correctly classified: all noise and inhomogeneity levels are considered.

TABLE III
CONFUSION TABLE OF PHANTOM 7N20RF. VALUES ARE IN PERCENTAGE COMPUTED OVER ALL VOXELS: A-FGMM, B-GHMRF, C-GPV, D-GPV-HMRF, E-EP, F-NP-HMRF. FALSE POSITIVES (FP) AND FALSE NEGATIVES (FN) ARE COMPUTED IN PERCENTAGE WITH RESPECT TO THE TOTAL TISSUE VOLUME OF THE REFERENCE AND TO THE OWN CLASSIFICATION, RESPECTIVELY

		Reference					FP
		CSF	CG	GM	GW	WM	
A	CSF	92.3	20.9	0.1	0.0	0.0	24.6
	CG	6.5	33.6	3.3	0.1	0.0	22.1
	GM	1.2	44.9	87.8	36.3	2.2	40.9
	GW	0.0	0.6	7.7	26.9	8.4	42.5
	WM	0.0	0.0	1.2	36.6	89.4	27.5
FN		7.7	66.4	12.2	73.0	10.6	

		Reference					FP
		CSF	CG	GM	GW	WM	
B	CSF	89.9	11.5	0.1	0.0	0.0	14.3
	CG	9.5	47.2	2.3	0.0	0.1	12.8
	GM	0.6	41.2	88.9	19.3	0.6	35.9
	GW	0.0	0.1	8.8	59.4	5.6	24.2
	WM	0.0	0.0	0.0	21.2	93.7	10.6
FN		8.3	48.7	2.8	47.5	14.2	

		Reference					FP
		CSF	CG	GM	GW	WM	
C	CSF	67.0	9.4	0.1	0.0	0.0	13.9
	CG	31.7	48.3	7.0	0.6	0.1	34.6
	GM	1.3	39.0	69.8	26.3	2.6	40.5
	GW	0.0	3.2	21.8	45.8	26.9	42.3
	WM	0.0	0.1	1.3	27.3	70.4	32.4
FN		28.4	48.9	13.0	85.4	5.9	

		Reference					FP
		CSF	CG	GM	GW	WM	
D	CSF	90.5	10.8	0.1	0.0	0.0	14.7
	CG	9.1	57.2	4.7	0.1	0.2	19.2
	GM	0.3	31.9	84.4	18.8	0.3	30.2
	GW	0.0	0.1	10.8	66.3	10.6	28.2
	WM	0.0	0.0	0.0	14.8	88.9	12.9
FN		9.5	42.8	15.6	33.7	11.1	

		Reference					FP
		CSF	CG	GM	GW	WM	
E	CSF	91.5	19.8	0.1	0.0	0.0	23.8
	CG	8.0	55.0	17.3	1.4	0.1	37.5
	GM	0.5	24.7	74.8	36.9	2.4	38.8
	GW	0.0	0.5	7.7	45.7	30.6	49.6
	WM	0.0	0.0	0.1	16.0	66.9	17.6
FN		8.5	45.0	25.2	54.3	33.1	

		Reference					FP
		CSF	CG	GM	GW	WM	
F	CSF	54.1	2.0	0.1	0.0	0.0	5.1
	CG	44.5	39.9	0.3	0.0	0.0	44.1
	GM	1.5	58.1	93.7	30.8	0.9	40.8
	GW	0.0	0.1	6.0	55.5	13.6	29.4
	WM	0.0	0.0	0.0	13.7	85.5	12.4
FN		45.9	60.1	6.3	44.4	14.5	

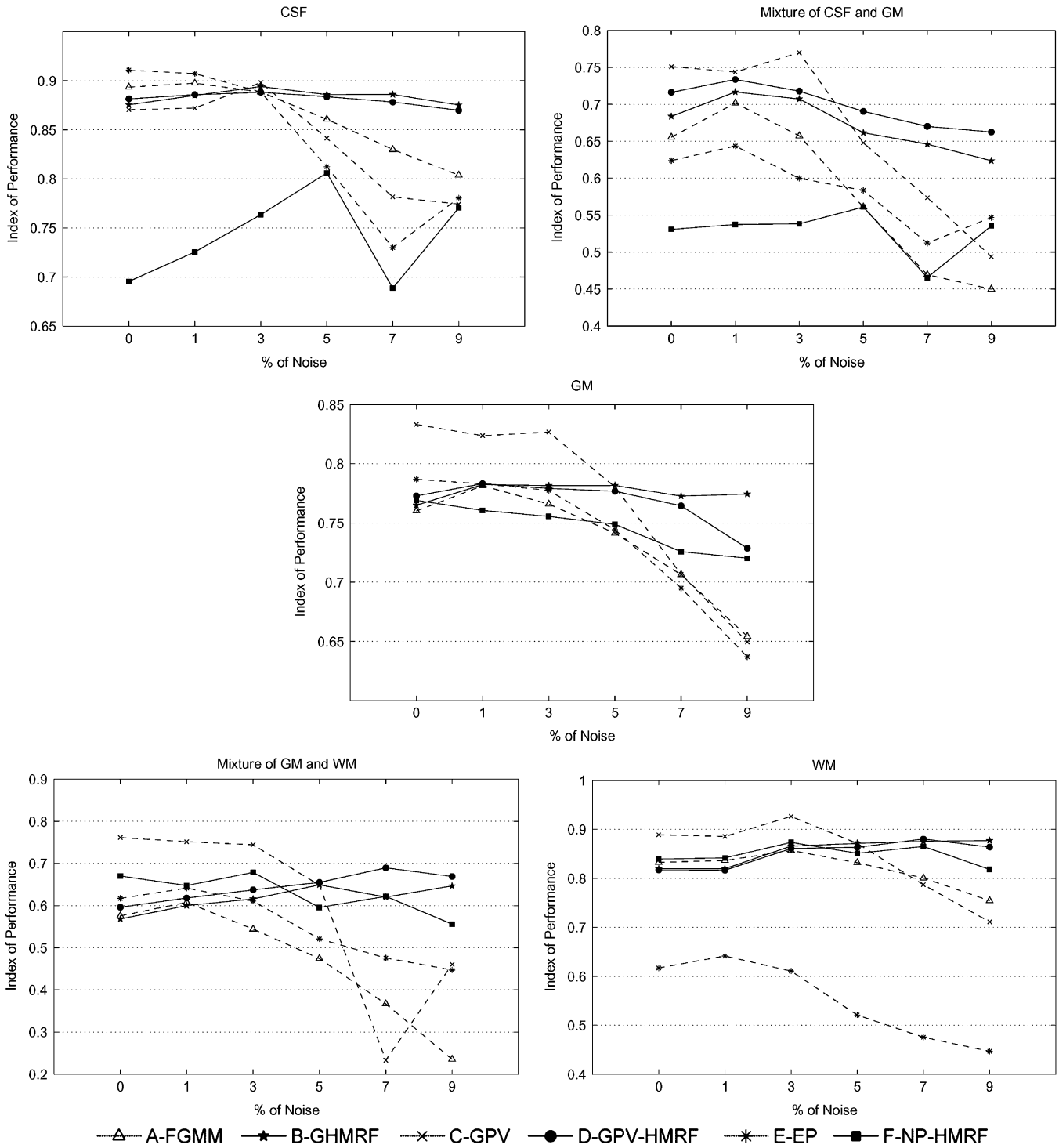


Fig. 8. DSM: all methods and all levels of noise are considered, $\mathrm{RF} = 20\%$.

Finally, we compute the Dice Similarity Measure (DSM) [55] for each tissue as a relative index of similarity. DSM is defined as

$$\mathrm{DSM}_{a,b}^{t} = \frac{2 \cdot N_{a \cap b}^{t}}{N_a^t + N_b^t} \tag{43}$$

where N_a^t and N_b^t are the voxels classified as tissue t with the methods a and b, respectively, and $N_{a \cap b}^t$ is the number of voxels classified as tissue t by both methods. This measure is sensitive to both differences in size and location. Although $\mathrm{DSM} > 0.7$ is considered as an excellent agreement between the two segmentations, DSM is hardly interpreted as an absolute value but as a value to compare the similarities between pairs of methods. In Fig. 8, the DSM with respect to the ground truth is presented for all methods and all levels of noise. In Table IV, DSM is shown for all levels of bias and $\mathrm{N} = 7\%$.

B. Volumetric Measures

Because volumetry is a major application of tissue classification, we investigate how the above results affect volume mea-

TABLE IV
DICE SIMILARITY MEASURE WITH RESPECT TO THE 5 CLASSES GROUND TRUTH: PHANTOM WITH 7% NOISE AND ALL RF LEVELS

	DSM {0%; 20%; 40%}					
	A-FGMM	B-GHMRF	C-GPV	D-GPV-HMRF	E-EP	F-NP-HMRF
CSF	0.83;0.82;0.80	0.88;0.88;0.85	0.81;0.78;0.73	0.88;0.87;0.84	0.77;0.73;0.77	0.75;0.68;0.69
CG	0.49;0.46;0.41	0.62;0.64;0.51	0.57;0.57;0.54	0.68;0.67;0.57	0.57;0.51;0.51	0.52;0.46;0.45
GM	0.71;0.70;0.62	0.79;0.77;0.65	0.70;0.70;0.66	0.79;0.76;0.67	0.70;0.69;0.63	0.74;0.72;0.66
GW	0.31;0.36;0.43	0.65;0.62;0.55	0.48;0.23;0.39	0.73;0.68;0.58	0.46;0.47;0.44	0.66;0.62;0.53
WM	0.78;0.80;0.78	0.88;0.87;0.83	0.77;0.78;0.78	0.89;0.88;0.82	0.78;0.79;0.78	0.89;0.86;0.81

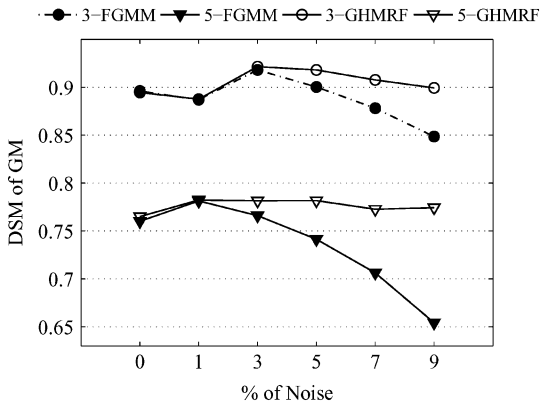


Fig. 9. Comparison of DSM values for GM between 3 and 5 tissues classification. Only methods A-FGMM and B-GHMRF are shown. Notice that DSM values obtained with 5 classes are lower than the ones obtained with 3 classes.

surements. In this section, we include both 3 and 5 class classification methods.

The true volume for each tissue of the synthetic data is obtained by computing the integral of the fuzzy tissue volumes. For the 3 class classification methods, the volume of CSF, GM, and WM are estimated by counting the voxels of each class. For the 5 class classification methods, PV voxels also contribute to the volume of each tissue. Their gray level is used to estimate the percentage of each pure tissue in the PV voxel. For instance, the total volume of GM is computed as

$$V_{\mathrm{GM}} = \sum_{\forall i \in \mathrm{GM}} 1 + \sum_{\forall i \in \mathrm{GW}} \left(\frac{y_i - \mu_{\mathrm{WM}}}{\mu_{\mathrm{GM}} - \mu_{\mathrm{WM}}} \right)^* + \sum_{\forall i \in \mathrm{CG}} \left(\frac{y_i - \mu_{\mathrm{CSF}}}{\mu_{\mathrm{GM}} - \mu_{\mathrm{CSF}}} \right)^* \quad (44)$$

where x^* means $\max(0, \min(x, 1))$. The other pure tissues have similar expressions but with a single PV class contribution. Tissue volume is computed on six brain digital phantoms (5N0RF, 7N0RF, 9N0RF, 5N20RF, 7N20RF, 9N20RF) and the root-mean-square (RMS) error over all phantoms for each tissue was computed with respect to the reference tissue volume. In Fig. 10, this error is shown as percentage with respect to the real volume of each tissue.

Finally, volumes are also measured locally, using the above formulae for cubes of (15 mm)^3. In Fig. 11(a) and (b) a slice of the simulated phantom 7N0RF and its corresponding local volume of GM (V_{ref}) are shown. Then, the difference between local volume computation for all methods using both 3 and 5 classes and the reference local volume image is shown in Fig. 11.

C. Classification of Real Data

Because of its more limited scope, real data is analyzed less exhaustively. The validation relies only on two of the above tests. First, visual inspection of the results for the 2 selected slices is performed in Fig. 12. Second, quantitative validation is presented in Table V where the DSM is computed for each method on both slices.

VI. DISCUSSION

A. Global Performance

There is no global winner as the most suitable tissue classification technique for T1-MR brain image. In fact, if we define the best classification as the one with the highest percentage of correct classified voxels, as in Fig. 7, the optimal method varies depending on the noise (N) and in-homogeneity (RF) levels present in the images. For low noise levels ($N = 3\%$), no method clearly outperforms the others. However, for higher noise levels ($N = 5\%$), D-GPV-HMRF almost always performs the best classification, closely followed by B-GHMRF, whose performance differs by less than 2%. In [46], methods are also compared by allowing small errors such as confusing a pure tissue with a PV containing it or confusing a PV voxel with one of its pure tissues. In this case, C-GPV and D-GPV-HMRF, both methods using the PV equation, have the lowest error rates for low and high noise levels, respectively. However, differences are less than 1%.

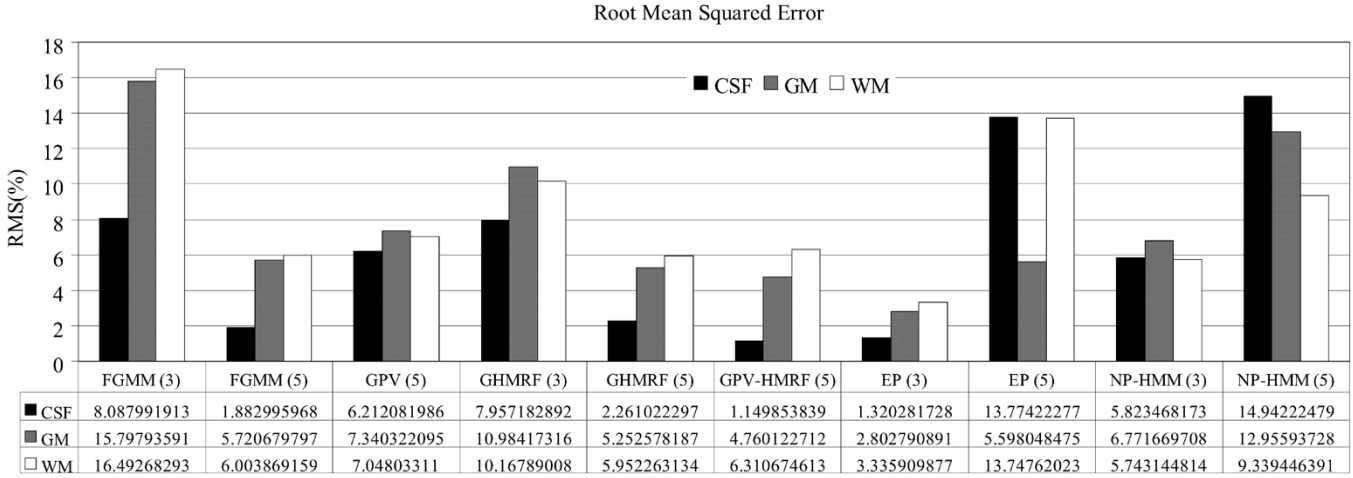
B. Robustness to Noise and Inhomogeneities

In order to evaluate their intrinsic robustness, none of the methods under study includes a preprocessing step to compensate for image artifacts such as noise or bias. In Fig. 7, all possible levels of noise and inhomogeneities present in the MRI simulator are considered. Robustness depends primarily on whether the methods use voxel intensity only or include a spatial prior.

Methods that consider intensity only are represented with dotted lines. In general, classification accuracy decreases with increasing noise and nonuniformities. A-FGMM is very sensitive to both noise and inhomogeneities. However, for low levels of noise, methods C-GPV and E-EP are equally performant in $\mathrm{RF} = 0$ and in $\mathrm{RF} = 20$. For very high noise levels ($\mathrm{N} \geq 7\%$), all methods perform a classification that converges toward a range of *pergood* equal to [60–65]% for any value of RF.

TABLE V
DICE SIMILARITY MEASURE OF EACH CLASSIFICATION METHOD WITH RESPECT TO THE ESTIMATED 5 CLASSES GROUND TRUTH

	DSM {Slice 1; Slice 2}					
Tissue/Method	A-FGMM	B-GHMRF	C-GPV	D-GPV-HMRF	E-EP	F-NP-HMRF
CSF	0.40; 0.54	0.21; 0.46	0.60; 0.64	0.30; 0.55	0.45; 0.53	0.56; 0.54
CSF+GM	0.47; 0.56	0.47; 0.56	0.62; 0.57	0.48; 0.57	0.40; 0.38	0.44; 0.40
GM	0.73; 0.72	0.71; 0.71	0.71; 0.64	0.73; 0.72	0.55; 0.50	0.62; 0.53
GM+WM	0.41; 0.41	0.57; 0.52	0.45; 0.45	0.57; 0.51	0.34; 0.37	0.43; 0.42
WM	0.92; 0.85	0.94; 0.85	0.92; 0.85	0.94; 0.85	0.89; 0.81	0.90; 0.82



	FGMM (3)	FGMM (5)	GPV (5)	GHMRF (3)	GHMRF (5)	GPV-HMRF (5)	EP (3)	EP (5)	NP-HMM (3)	NP-HMM (5)
CSF	8.087991913	1.882995968	6.212081986	7.957182892	2.261022297	1.149853839	1.320281728	13.77422277	5.823468173	14.94222479
GM	15.79793591	5.720679797	7.340322095	10.98417316	5.252578187	4.760122712	2.802790891	5.598048475	6.771669708	12.95593728
WM	16.49268293	6.003869159	7.04803311	10.16789008	5.952263134	6.310674613	3.335909877	13.74762023	5.743144814	9.339446391

Fig. 10. Study of tissue volume estimation: RMS error is computed over six phantoms (5N0RF, 7N0RF, 9N0RF, 5N20RF, 7N20RF, 9N20RF) and RMS is in percentage with respect to reference volume of every tissue.

Solid lines represent all methods using local spatial priors, which present similar behaviors with noise and bias. With no bias field, $\mathrm{RF} = 0$, *pergood* decreases proportionally to the increase of noise. For $\mathrm{RF} = 20$, there is no decrease of quality but almost a constant *pergood*. Finally, for $\mathrm{RF} = 40$, the *pergood* actually increases for high noise levels. The reason for this unexpected behavior is that—in the presence of a strong bias field—low noise levels $(\mathrm{N} = 3\%)$ are not realistically modeled by Gaussian distributions.

C. Pure Tissues and Partial Volume

Using confusion tables such as Table III, the global conclusions can be refined on a tissue per tissue basis. Considering such confusion tables for noises $N = 5\%$ and bias fields $\mathrm{RF} = 0\%, \mathrm{RF} = 20\%$, and $\mathrm{RF} = 40\%$, we observe that the best classifier for CSF is B-GHMRF (70% of the cases), the best classifier for GM is F-NP-HMRF (70% of the cases) and the best classification of WM tissue is performed by B-GHMRF in more than 50% of the cases. D-GPV-HMRF almost always achieves the best classification score for both PV tissues: 78% of the cases for CG and 100% for GW.

These results show that PV distributions are not properly modeled by a Gaussian function. This is also clear when looking at the histogram fitting of Fig. 6 where CG and GW mixtures are always better fitted by methods C-GPV and D-GPV-HMRF using the PV equation. However, the percentage of voxels correctly classified for a mixture tissue never reaches more than 73% while the best scores for pure tissues usually reach 90% of correctly classified voxels. This poor result indicates that PV distribution is not properly modeled yet, and may require additional study and modeling. For instance, different types of GW mixtures could be considered as recently suggested in [27]. This anatomical model splits the GW mixture into a geometrical GW mixture corresponding to the brain cortico-subcortical interface and a mosaic GW mixture corresponding to the deep cerebral nuclei structures such as the thalamus.

Tissue per tissue robustness to noise is analyzed with the DSM for bias field $\mathrm{RF} = 20\%$ (see Fig. 8). DSM for pure tissues is almost always above 0.7, which is considered as an excellent similarity. The best methods to classify the CSF are B-GHMRF and D-GPV-HMRF, whose $\mathrm{DSM}^{\mathrm{CSF}}$ stands between 0.85 and 0.9. They also show excellent noise robustness. The other methods also have a good $\mathrm{DSM}^{\mathrm{CSF}}$, but they prove to be more sensitive to noise. Similar conclusions can be obtained in the case of WM classification. All methods present an excellent similarity to the ground truth classification $(\mathrm{DSM}^{\mathrm{WM}} > 0.8)$ and a small noise sensitivity. The largest variabilities were obtained for GM classification. All methods using spatial prior have $\mathrm{DSM}^{\mathrm{GM}} > 0.7$, while the methods using only the image intensity are more sensitive to noise, decreasing their performance down to $\mathrm{DSM}^{\mathrm{GM}} = 0.64$. DSM measures are much less satisfactory for mixture tissues, for

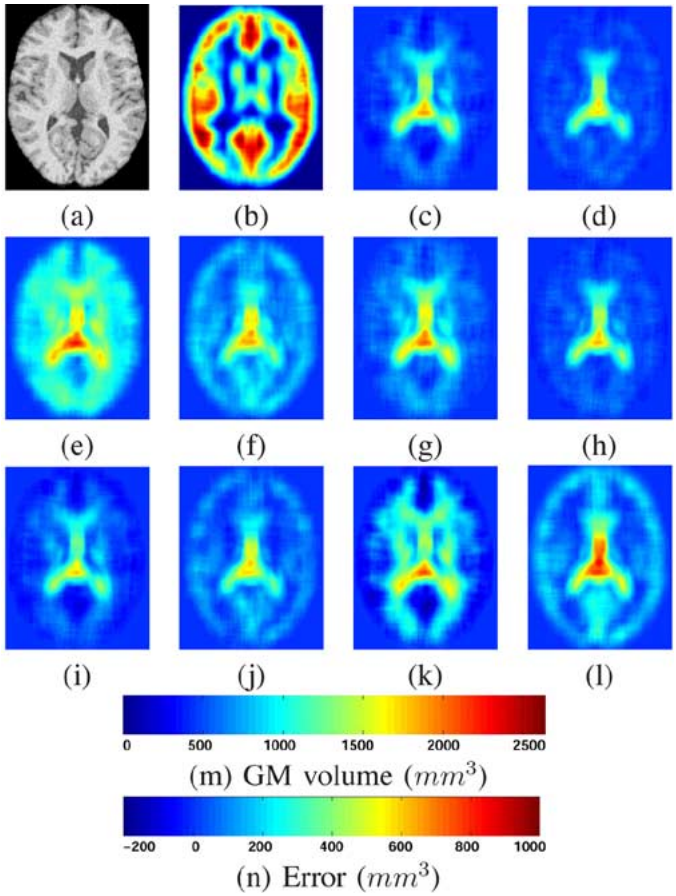


Fig. 11. (a) A Brainweb simulated MRI (7N0RF). (b) Local volume of GM in $mm^3(V_{\text{ref}})$. Difference of local volume computation with respect to V_{ref}: (c) FGMM in 5 classes, (d) GHMRF in 5 classes, (e) FGMM in 3 classes, (f) GHMRF in 3 classes, (g) GPV in 5 classes, (h) GPV-HMRF in 5 classes, (i) EP in 3 classes, (j) NP-HMRF in 3 classes, (k) EP in 5 classes, and (l) NP-HMRF in 5 classes. Negative and positive values, respectively, mean an underestimation and overestimation of the computed local tissue volume. (m) Local volume of GM in mm^3, and (n) Volume estimation error in mm^3.

which they range between 0.2 and 0.8. For low noise levels, C-GPV obtains an excellent agreement but it decreases for high noise levels. The most robust classification is obtained by B-GHMRF and D-GPV-HMRF whose DSM values are within 0.6 and 0.7.

Notice that DSM values obtained with a 5 class classification are automatically lower than the ones obtained with a 3 class classification. Hence, we cannot directly compare DSM values to the ones published by other groups [56]. In Fig. 9, DSM^{GM} values for 3 tissue and 5 tissue classification using A-FGMM and B-GHMRF are compared.

D. Computation of Tissue Volume

The results in Fig. 10 allow us to assess how the above classification accuracies affect a practical problem such as tissue volumetry. In this section, we consider both the 6 classification methods into 5 classes of this paper, but also 4 methods that generate 3 pure tissue classes and no PV. The 3-classes methods include a parametric model made of mixture of 3 Gaussian distributions and a nonparametric approach, both with or without a MRF to ensure spatial coherence.

For the parametric methods, 5-classes classification gives better estimates of the volume of each tissue than the 3-classes approaches. Tissue volume computation is globally improved by the parametric methods that use a MRF in comparison to the ones that only consider the intensity information.

However, for nonparametric approaches we have a completely opposite behavior. Among those classification methods, the lowest error is obtained by the 3-classes classification using EP. Hence, in the case of nonparametric approaches, considering 5-classes does not improve the results. Indeed, nonparametric approaches do not estimate correctly the PV classes, usually overestimating them. In the case of nonparametric approaches, using a MRF does not improve the total tissue volume computation.

This analysis can be refined by looking at the local volume measures (see Section V-B) of GM in Fig. 11. In Fig. 11(a) and (b) a MR image and the local volume of $\text{GM}(V_{\text{ref}})$ are shown. The intensity value of every voxel in Fig. 11(b) represents the volume of GM in a region of interest of $(15\text{ mm})^3$ centered on this voxel, see colorbar (m). The rest of the images show the difference between the local volume computed by a classification method and V_{ref}, see colorbar (n). In the first two rows, parametric methods using 5 classes and a MRF (d, h) are clearly better at estimating local tissue volume than methods that do not consider MRF (c, g) or that only consider 3 classes (e, f). For the most efficient methods (d, h), cortical grey matter volume is accurately measured while significant errors remain for deep brain structures such as the putamen, thalamus and caudate nucleus.

In the bottom row, nonparametric approaches display a different behavior. 3-classes EP (i) has both under and over estimation of GM in the cortex, so that when the total tissue volume is computed there is an error compensation effect. This effect is removed by adding the MRF (j), which always overestimates cortical GM. In the 5–class classification (k) compensation errors increase around the cortex as well as errors in deep brain structures. When adding a MRF (l), the errors in deep brain structures are smaller, but as previously, an overestimation of cortical GM appears, which removes the error compensation entirely. This explains why MRF do not improve global tissue volume computation in the case of nonparametric approaches.

E. Real Data

Even though simulated data provides an excellent tool to validate and compare the performance and robustness of algorithms, assessment on real data is ultimately needed. The results in Fig. 12 and Tables IV and V show that conclusions drawn on simulated data can be extended to real data. By analogy, the classification methods can directly process patient data with degenerative brain tissue diseases only, that is, without other pathological processes involved such as early stages of Alzheimer's disease or Schizophrenia.

In Fig. 12, visual validation on real data shows that—similarly to the simulated data study—the methods using local spatial prior are less noisy than the ones using only intensity information. Quantitative validation is shown in Table V with DSM for each method on the selected slices with respect to the 5-classes estimated ground truth. DSM values are similar on

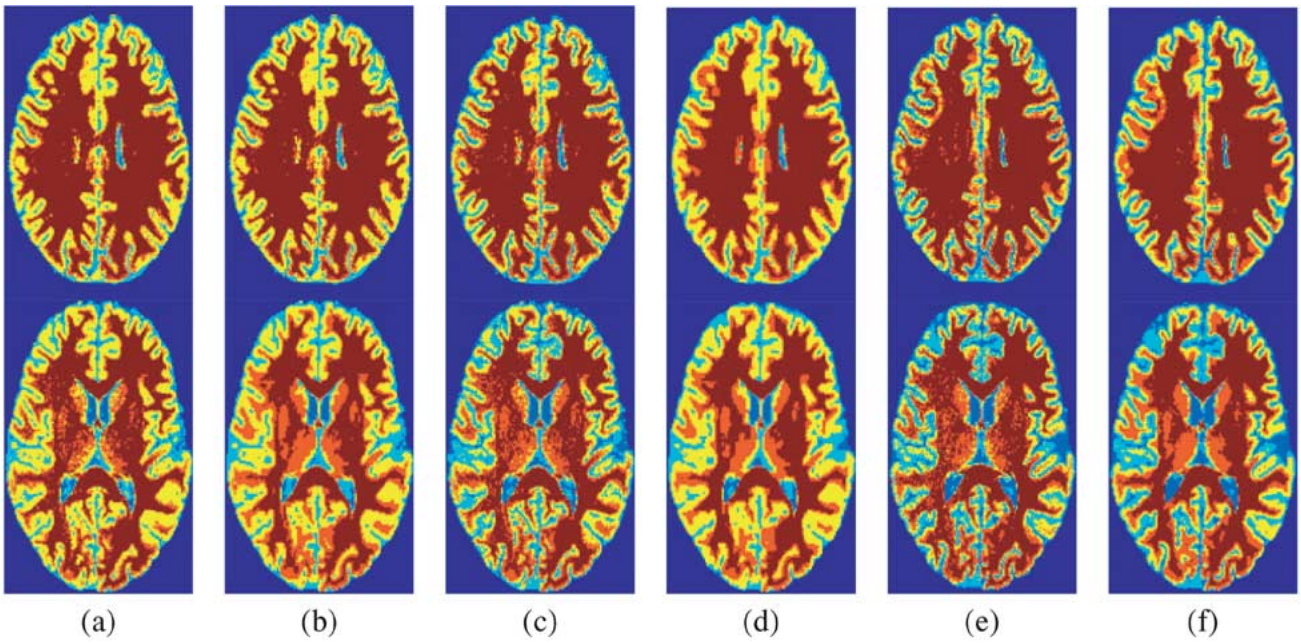


Fig. 12. Statistical nonsupervised classification on a real MR brain image. Columns are: (a) FGMM, (b) GHMRF, (c) GPV, (d) GPV-HMRF, (e) EP, and (f) NP-HMRF. Slice 1 and 2 are in top and bottom row, respectively.

TABLE VI
CONFUSION TABLE OF METHOD D-GPV-HMRF WITH RESPECT TO THE ESTIMATED 5 CLASSES GROUND TRUTH. VALUES ARE PERCENTAGE WITH RESPECT TO THE ESTIMATED GROUND TRUTH. GREY MATTER HAVE BEEN SEPARATED INTO CORTICAL AND DEEP BRAIN CLASSES

Reference	Method				
	CSF	CSF+GM	GM	GM+WM	WM
CSF	46.2295	45.0273	5.1366	3.6066	0
CSF+GM	2.8022	52.9121	41.8681	2.3077	0.1099
GM cortex	0	4.1111	85.7337	9.9374	0.2178
GM central	0	0	15.9871	61.6953	22.3176
GM+WM	0.0474	0.1422	18.7293	61.3561	19.725
WM	0	0.0432	0.367	15.0475	84.5423

both slices and they are comparable to the ones obtained with simulated data at Table IV, except for CSF because of its large variability. Satisfactory results ($\mathrm{DSM} > 0.7$) are obtained for GM, except for nonparametric methods. All methods give excellent results for WM ($\mathrm{DSM} > 0.8$).

Again in agreement with the simulation results, mixture tissues have lower rates of correct classification. In the case of GW, this arises primarily because of the choice of the experts to classify the central nuclei—part of the thalamus and caudate nuclei—as pure GM instead of GW mixture tissue. Nonsupervised classification always select these structures as GW or WM. This disagreement has been quantified for D-GPV-HMRF by creating a confusion table at Table VI, where cortical GM and central GM were split into two different classes. Almost 85% of cortical GM is correctly classified as GM. On the contrary, only 16% of central GM is classified as GM while 60% is classified as GW and 22% as WM. The confusion table also shows that most of the CSF is actually classified as CG.

In summary, conclusions for real data are the same as for the simulated data: B-GHMRF and D-GPV-HMRF perform, in general, much better than the others while C-GPV shows the best performance for CSF and CG.

VII. CONCLUSION

This paper presents a validation study on statistical classification of brain tissue in MR images. Several image models have been assessed assuming different hypotheses regarding the intensity distribution model, the spatial model and the number of classes. Both qualitative and quantitative validation on simulated data allows us to obtain the following conclusions.

The percentage of correct classification never reaches 100% and, even if pure tissues are in general correctly classified, PVs are not. Nonparametric models have performed in some cases equally or even better than parametric approaches. Actually, 3-classes EP algorithm has proved very well-suited to perform volume computation. This is because the misclassification made with nonparametric approaches is mainly due to an overestimation of both mixture classes. However, in the case of parametric approaches, results show that, even if the assumptions regarding the mixture tissues are imperfect, it is necessary to take them into account. Actually, 5-class models not only better estimate the image histogram but they also reduce considerably the errors in the estimation of tissue volume. Our study has also revealed that techniques considering spatial information increase in average the accuracy of the classification by 7%. Finally, we have shown that the results obtained with simulated data can also be representative of real conditions of *normal* brains.

Emerging classification methods add atlas information to the intensity and local spatial priors [11], [24], [56], [57]. One main line of our current research is to quantify the importance of this kind of information. Some preliminary results [46] have shown that the performance of such methods is very sensitive to registration errors and to the precision of the atlas prior. Actually, mixture tissues are particularly affected by prior class template errors while pure tissue classification has almost always been improved.

Our current research also aims to quantify the sensitivity of the algorithm to their parameters such as β in HMRF methods as well as the effect of preprocessing the images by an anisotropic filter or a bias corrector or adding a bias field estimation model. We expect both the preprocessing and bias model (as in [2], [10], and [19]) to make the classification more robust faced with noise and inhomogeneities. However, we suspect the preprocessing could displace PV voxels, so that errors might be added in mixture tissue classification.

ACKNOWLEDGMENT

The authors would like to thank Dr. D. Viceic and Dr. P. Hagmann for their precious contributions to this work. They would also like to thank Dr. S. Warfield and Dr. C. Gaser for their helpful discussions.

REFERENCES

[1] C. Guttmann, R. Benson, S. K. Warfield, X. Wei, M. Anderson, C. Hall, K. Abu-Hasaballah, J. Mugler, and L. Wolfson, "White matter abnormalities in mobility-impaired older persons," *Neurology*, vol. 54, no. 6, pp. 1277–1283, 2000.

[2] W. Wells, R. Kikinis, W. Grimson, and F. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, no. 4, pp. 429–442, Aug. 1996.

[3] S. Smith, Y. Zhang, M. Jenkinson, J. Chen, P. Matthews, A. Federico, and N. D. Stefano, "Accurate, robust, and automated longitudinal and cross-sectional brain change analysis," *NeuroImage*, vol. 17, no. 1, pp. 479–489, 2002.

[4] F. Maes, K. Van Leemput, L. E. DeLisi, D. Vandermeulen, and P. Suetens, "Quantification of cerebral grey and white matter asymmetry from MRI," *Med. Image Computing Comput.-Assist. Intervention*, pp. 348–357, 1999.

[5] J. Ashburner and K. Friston, "Voxel-based morphometry—The methods," *NeuroImage*, vol. 11, pp. 805–821, 2000.

[6] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, template moderated, spatially varying statistical classification," *Med. Image Anal.*, vol. 4, no. 1, pp. 43–55, Mar. 2000.

[7] L. P. Clarke, R. P. Velthuizen, M. A. Camacho, J. Heine, M. Vaidyanathan, L. O. Hall, R. W. Thatcher, and M. L. Silbiger, "Mri segmentation: Methods and applications," *Magn. Reson. Imag.*, vol. 13, pp. 343–368, 1995.

[8] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognit.*, vol. 26, no. 9, pp. 1277–1294, 1993.

[9] J. S. Suri, S. Singh, and L. Reden, "Computer vision and pattern recognition techniques for 2-D and 3-D MR cerebral cortical segmentation (part I): A state-of-the-art review," *Pattern Anal. Applicat.*, vol. 5, pp. 46–76, 2002.

[10] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 885–896, Oct 1999.

[11] A. Noe, S. Kovacic, and J. C. Gee. (2001) Segmentation of cerebral mri scans using a partial volume model, shading correction, and an anatomical prior. *Proc. SPIE (Medical Image Processing)* [Online]. Available: . Available: http://vision.fe.uni-lj.si/docs/AljazN/noe-SPIE-01.pdf

[12] L. Nocera and J. C. Gee, "Robust partial volume tissue classification of cerebral MRI scans," *Proc. SPIE (Medical Imaging 1997: Image Processing)*, vol. 3034, pp. 312–322, 1997.

[13] P. Schroeter *et al.*, "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 17, no. 2, pp. 172–186, Apr. 1998.

[14] R. Guillemaud and M. Brady, "Estimating the bias field of MR images," *IEEE Trans. Med. Imag.*, vol. 16, no. 3, pp. 238–251, Jun. 1997.

[15] S. Ruan, C. Jaggi, J. Xue, and J. Bloyet, "Brain tissue classification of magnetic resonance images using partial volume modeling," *IEEE Trans. Med. Imag.*, vol. 19, no. 12, pp. 172–186, Dec. 2000.

[16] P. Santago and H. D. Gage, "Quantification of MR brain images by mixture density and partial volume modeling," *IEEE Trans. Med. Imag.*, vol. 12, no. 3, pp. 566–574, Sep. 1993.

[17] P. Santago and H. Gage, "Statistical models of partial volume effect," *IEEE Trans. Image Process.*, vol. 4, no. 11, pp. 1531–1540, Nov. 1995.

[18] D. H. Laidlaw, K. W. Fleischer, and A. H. Barr, "Partial-volume Bayesian classification of material mixtures in MR volume data using voxel histograms," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 74–86, Feb. 1998.

[19] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, pp. 856–876, 2001.

[20] M. A. Gonzalez Ballester, A. P. Zisserman, and M. Brady, "Estimation of the partial volume effect in MRI," *Med. Image Anal.*, vol. 6, no. 4, pp. 389–405, Dec. 2002.

[21] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "A unifying framework for partial volume segmentation of brain MR images," *IEEE Trans. Med. Imag.*, vol. 22, no. 1, pp. 105–119, Jan. 2003.

[22] K. Held, E. Rota Kops, B. J. Krause, W. M. Wells, R. Kikinis, and H.-W. Muller-Gartner, "Markov random field segmentation of brain MR images," *IEEE Trans. Med. Imag.*, vol. 16, no. 6, pp. 878–886, Dec. 1997.

[23] Y. Zhang *et al.*, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[24] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 897–908, Oct. 1999.

[25] S. Ruan, B. Moretti, J. Fadili, and D. Bloyet, "Fuzzy Markovian segmentation in application of magnetic resonance images," *Comput. Vis. Image Understanding*, vol. 85, no. 1, pp. 54–69, 2002.

[26] T. Butz, "From error probability to information theoretic signal and image processing," Ph.D. dissertation, Signal Process. Inst., Swiss Federal Inst. Technol., Zurich, Switzerland, Jun 2003.

[27] T. Butz, P. Hagmann, E. Tardif, R. Meuli, and J.-P. Thiran, "A new brain segmentation framework," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2003, vol. , Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 586–593.

[28] D. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, and A. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–468, 1998.

[29] R. K.-S. Kwan, A. C. Evans, and G. B. Pike, "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Trans. Med. Imag.*, vol. 18, no. 11, pp. 1085–1097, Nov. 1999.

[30] V. Grau, A. U. J. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 447–458, Apr. 2004.

[31] X. Zeng, L. H. Staib, R. T. Schultz, and J. S. Duncan, "Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 927–937, Oct. 1999.

[32] M. Bach Cuadra, B. Platel, E. Solanas, T. Butz, and J.-Ph. Thiran, "Validation of tissue modelization and classification techniques in T1-weighted MR brain images," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, Oct. 2002, vol. , Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 290–297.

[33] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[34] A. Noe and J. C. Gee. Partial volume segmentation of cerebral mri scans with mixture model clustering. presented at Information Processing in Medical Imaging: 17th International Conference (IPMI). [Online]. Available: http://vision.fe.uni-lj.si/docs/AljazN/noe-IPMI-01.pdf

[35] E. Gokcay and J. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 158–171, Feb. 2002.

[36] J. Zhang, "The mean field theory in EM procedures for Markov random fields," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2570–2583, Oct. 1992.

[37] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, I. Karatzas and M. Yor, Eds. New York: Springer, 1995.

[38] J. M. Hammersely and P. Clifford, "Markov fields on finite graphs and lattices," unpublished, 1968.

[39] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Statist. Soc.*, vol. 36, pp. 192–326, 1974.

[40] N. Peyrard, "Approximations de type champ moyen des modles de champ de Markov pour la segmentation de donnes spatiales," Ph.D. Dissertation, University J. Fourier, Grenoble, France, Oct. 2001.

[41] G. Celeux, F. Forbes, and N. Peyrard, "EM-based image segmentation using potts models with external field," INRIA, Paris, France, Tech. Rep. 4456, Apr. 2002.
[42] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statist. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.
[43] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, pp. 721–741, 1984.
[44] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, S. R. Institute, Ed. New York: Wiley, 1973.
[45] P. N. Jones and G. J. McLachlan, Algorithm AS 254. Maximum likelihood estimation from grouped and truncated data with finite normal mixture models, in Appl. Statist., vol. 39, pp. 273–312, 1990.
[46] M. B. Cuadra, "Atlas-based segmentation and classification of magnetic resonance brain images," Ph.D. dissertation, Signal Process. Inst., Swiss Federal Inst. Technol., Zurich, Switzerland, Nov. 2003. thesis 2875.
[47] G. J. McLachlam and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.
[48] J. Matsuzawa, M. Matsui, T. Konishi, K. Noguchi, R. Gur, W. Bilder, and T. Miyawaki, "Age-related volumetric changes of brain gray and white matter in helathy infants and children," *Cereb. Cortex*, vol. 11, pp. 335–342, Apr. 2001.
[49] P. Huppi, S. Warfield, R. Kikinis, P. Barnes, G. Zientara, F. Jolesz, M. Tsuji, and J. Volpe, "Quantitative magnetic resonance imaging of brain development in premature and mature newborns," *Ann. Neurol.*, vol. 43, pp. 224–235, 1998.
[50] G. Gerig, M. Prastawa, W. Lin, and J. Gilmore, "Assessing early brain development in neonates by segmentation of high-resolution 3T MRI," *Lecture Notes in Computer Science* , vol. 2879, pp. 979–980, 2003.
[51] K. Van Leemput, "Quantitative Analysis of Signal Abnormalities in MRI Imaging for Multiple Sclerosis and Creutzfeldt-Jacob Disease," Ph.D. Dissertation, Faculteit Toegepaste Wetenschappen, Faculteit Geneeskunde, Katholieke Universiteit Leuven, Belgium, May 2001.
[52] C. Guttmann, F. Jolesz, R. Kikinis, R. Killiany, M. Moss, T. Sandor, and M. Albert, "White matter changes with normal aging," *Neurology*, vol. 50, pp. 972–978, 1998.
[53] A. Evans, D. Collins, P. Neelin, M. Kamber, and T. S. Marrett, "Three-dimensional correlative imaging: Applications in human brain mapping," *Functional Imaging: Technical Foundations*, pp. 145–162, 1994.
[54] S. K. Warfield, K. H. Zou, and W. Wells, "Validation of image segmentation and exper quality with an expectation-maximization algorithm," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2002, Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 298–306.
[55] A. P. Zijdenbos, B. M. Dawant, R. A. MArgolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 716–724, Dec. 1994.
[56] J. Marroquin, B. C. Vemuri, S. Botello, F. Calderon, and A. Fernandez-Bouzas, "An accurate and efficient Bayesian method for automatic segmentation of brain MRI," *IEEE Trans. Med. Imag.*, vol. 21, no. 8, pp. 934–945, Aug. 2002.
[57] K. Pohl, W. Wells, A. Guimond, K. Kasai, M. Shenton, R. Kikinis, W. E. L. Grimson, and S. K. Warfield, "Incorporating nonrigid registration into expectation maximization algorithm to segment MR images," in *Lecture Notes in Computer ScienceMedical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer-Verlag, 2002, Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 564–571.