Validation of Tissue Modelization and Classification techniques in T1-weighted MR Brain Images

Meritxell Bach Cuadra, Leila Cammoun, Torsten Butz,
Olivier Cuisenaire and J.-P. Thiran

Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland

Email:{Meritxell.Bach,Leila.Cammoun,Olivier.Cuisenaire,JP.Thiran}@epfl.ch
http://ltswww.epfl.ch/~brain

Rapport ITS 05.04

Abstract

We present a deep study of the performance of tissue modelization and classification techniques for T1-weighted MR images. It is assumed that only T1-weighted MR image modality is available. The methods presented here were selected to represent the whole range of prior information that can be used in the classification, i.e. intensity, spatial and anatomical priors. First, we consider the finite Gaussian mixture model (A-FGMM) with a Bayes classification. The second method is closely related to A-FGMM but also considers a hidden Markov random field (HMRF) model to account for spatial prior information. For this model, the maximum a posteriori (MAP) criterion is used as the classification decision rule. Third, we study a tissue model that assumes the mixture tissues to be probablistically modeled by the linear combination of their correspondent pure Gaussian tissue densities (C-GPV). Here again, Bayes classification is used for the final classification. The fourth method, D-GPV-HMRF, uses the same image model as method C-GPV, but it also encodes spatial information by a hidden Markov random field as done in method B-HMRF. The fifth algorithm does not model the tissue classes by parametric probability densities, but rather by non-parametric models. As a result, the probabilistic tissue model and the classification criterion can not be distinguished anymore, but are directly interdependent. The resulting algorithm minimizes an information theortic quantity, called the error probability (E-EP). The final method is also non-parametric, but again adds a HMRF to model spatial prior information (F-NPHMRF). All methods have been tested on Digital Brain Phantom images for which the classification ground truths were known. Noise and intensity non-uniformities were added to simulate real imaging conditions. No enhancement of the image quality is considered either before or during the classification process. This way robustness and accuracy of the methods is tested in front of the image artifacts. Results demonstrate clearly that methods relying on both, intensity and spatial information, are in more robust to noise and field inhomogeneities. We demonstrate also that partial volume (PV) is still not perfectly modeled, even though methods that account for mixture classes outperform methods that just consider pure classes.

I. Introduction

Accurate and robust brain tissue segmentation from MR images is a key issue in many applications of medical image analysis [1] and, particularly, in the study of many brain disorders [2], [3]. Manual tracing of the three brain tissue types, white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF), in MR images by an expert is far too time consuming as the data involved in most studies is large. On the other hand, automated and reliable tissue classification is a demanding task as the intensity representation of the data normally does not allow a clear delimitation of the different tissue types present in a natural MRI. This is due to the partial volume (PV) effect (presence of more than one brain tissue type in a voxel), image noise and intensity non-uniformities caused by the inhomogeneities in the magnetic field of the MR scanner.

Two main groups can be distinguished in statistical classification: *supervised* and *non-supervised* methods. Supervised classification techniques, also called semi-automatic methods, explicitly need user interaction while non-supervised classification is completely automatic. Actually, a large number of approaches have been proposed to deal with the MR brain image classification problem but a complete review of all these classification methods is beyond the purpose of this chapter. However, let us briefly present a state of the art of the automatic segmentation methods.

Statistical parametric approaches are widely used for brain MR image segmentation. These approaches usually solve the estimation problem of assigning a class label to a voxel with a suitable assumption about the intensity distribution but the classification can be made also by the estimation of the relative amounts of the various tissue types within a voxel rather than assigning each voxel to one single tissue type [4], [5]. Actually, parametric classification methods try to solve a twofold problem: on one hand, the *classification* which is an easy task if the tissue type model is good while, on the other hand, the *parameter estimation* of the tissue class which is an easy task if all the voxels within this class are known. *Finite Gaussian Mixture (FGM)* models assume a Gaussian distribution for the intensities of the main brain tissues [6]: GM, WM, and CSF. Other algorithms [7] add separate classes to take into account the PV voxels and model them also by independent Gaussian densities. In more elaborate methods [8] mixing proportions are taken into account to build a more realistic model for PV voxels which differs from a Gaussian distribution. However, some of the *finite mixture (FM)* models have the limitation of not considering the spatial information. That is the reason why increasing attention has been paid recently to methods that model the mixing proportions of the PV voxels by a Markov Random Field (MRF) [9]. Finally, non-parametric classification techniques can be considered when no well justified parametric model is known [10], as for instance the intra-class statistics.

As in the case of registration and atlas-based segmentation, validation of brain tissue classification is a complex issue in medical image processing. Visual inspection and comparison with manual segmentation are labor intensive and almost not reliable since the amount of data to deal with is usually large. Tissue classification methods can also be assessed by using synthetic data even if these kind 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 *phantom* [11]. This phantom is very well-suited for this purpose since a ground-truth classification is known while different types of T1w MR modalities and image artifacts can be reproduced.

The goal of this work is to assess the robustness and accuracy of some of the most used unsupervised classification methods. In this comparative analysis and validation only T1w MR brain image are considered. The goal is to be able to specify the most suitable tissue classification technique depending on the different conditions that could be encountered in T1w MR brain image. The work presented here is the continuation of [12].

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. Section III and Section IV briefly introduce the basic theoretical concepts of the classification criteria and parameter estimation. Then, in Section V and Section VI, the methods analyzed in this comparative study are summarized. Next, Section VII and Section VIII, the classification results and their validation are discussed. Finally, important conclusions resulting from the presented work are given.

II. Image model

A. Intensity distribution model

In this thesis, the theory behind the intensities in T1w MR brain images is similar to the one introduced by Noe et al. [5]. Its main concepts are recalled in what follows.

Let us index N data points to be classified with $i \in \mathcal{S} = \{1, 2, ..., N\}$. In the case of 3D images, such as MR images, they index the image's voxels. Let us furthermore denote the data feature vectors

by $y_i \in \mathbb{R}$. In the case of classification of single MR images, y_i represent the *ith*-voxel intensity. 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 $\boldsymbol{y} = \{y_1, y_2, ..., y_N\} \in \mathcal{D}^\mathcal{N} \subset \mathbb{R}^\mathbb{N}$.

The classification process aims to classify the data $\mathcal{S}$ into one of (hidden) underlying classes present in the image labelled by one of the symbols $\mathcal{L} = \{CSF, GM, WM, CG, CW, GW, CGW\}$[1]. The family of random variables X represents these classes, $\boldsymbol{x} = \{x_1, x_2, ..., x_N\} \in \mathcal{L}^N$ denotes a possible configuration of X, and $\mathcal{X}$ is the space of all possible configurations.

Now, let us suppose that all the random variables, Y_i, are identically and independently distributed. Then, the probability density function of the intensity voxel can be defined by:

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

where $i \in \mathcal{S}$, $P(x_i)$ is the probability of the tissue class x_i and $P(y_i|x_i)$ is the probability density function of y_i given the tissue class x_i.

The simplest intensity model that could be used considers only the three pure tissues of the brain, that is, $\mathcal{L}_p = \{CSF, GM, WM\}$ and the probability density function for the observing intensity y_i given the pure tissue class x_i is given by the Gaussian function:

$$P(y_i | x_i) = \frac{1}{\sigma_{x_i} \sqrt{2\pi}} \, Exp \left[\frac{-(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} \right], x_i \in \mathcal{L}_p. \tag{2}$$

where the model parameters $\theta_{x_i} = \{\mu_{x_i}, \sigma_{x_i}\}$ are respectively the mean and variance of the Gaussian function. This is a good approximation since the noise present in a MRI follows a Rician distribution that, at high signal-to-noise ratio (SNR), can be modelled by a Gaussian distribution[2].

In this thesis, a more evolved intensity model that adds to the main brain tissues their most important mixtures is used, i.e., $\mathcal{L}_{pm} = \{CSF, GM, WM, CG, GM\}$[3]. A voxel containing only a pure tissue is still modelled by a Gaussian distribution while a mixture voxel is modelled as suggested in [8] by

$$P(y_i | x_i, \alpha) = \frac{1}{\sigma_{x_i}(\alpha) \sqrt{2\pi}} \, Exp \left[\frac{-(y_i - \mu_{x_i}(\alpha))^2}{2\sigma_{x_i}^2(\alpha)} \right], \; x_i \in \mathcal{L}_{pm} \backslash \mathcal{L}_p, \tag{3}$$

where the two pure tissues composing the mixture voxel are denoted by $l_1, l_2 \in \mathcal{L}_p$, and α is a uniformly distributed random variable that represents the fraction of l_1 present in the mixture voxel (then, tissue l_2 is present in a fraction of $1 - \alpha$). The mean and variance of the mixture are determined by the model parameters of the pure tissues:

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

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

Finally, the probability density function of a partial volume tissue is computed by

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

The integral in Eq. (6) is numerically computed and its form can largely vary depending on the parameters $\theta_l = \{\mu_l, \sigma_l\}$. Some particular cases of Eq. 6 are plotted in Fig. 1. It can be observed that

---

[1]CG, CW, GW and CGW are the mixtures of CSF+GM, CSF+WM, GM+WM, and CSF+GM+WM, respectively.

[2]Note that for low SNR, i.e. the background image, the Rician noise can be modelled by a Rayleigh distribution.

[3]CW and CGW are not considered because these mixtures are uncommon, and thus $P(CW)$ and $P(CGW)$ are not relevant in explaining $P(\boldsymbol{y})$.
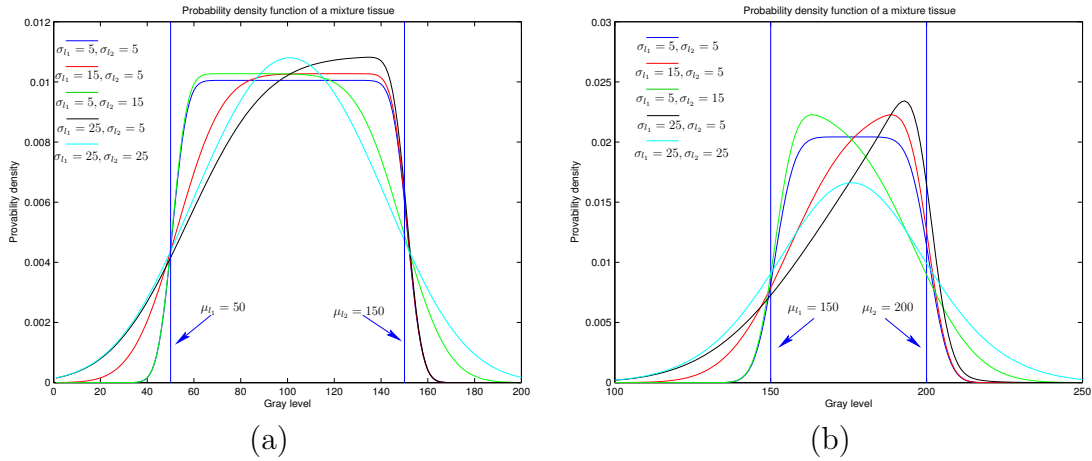
Fig. 1. Plot of Equation 6 varying σ_{l_1} and σ_{l_2} with (a) $\mu_{l_1} = 50$ and $\mu_{l_2} = 150$ and (b) $\mu_{l_1} = 150$ and $\mu_{l_2} = 200$.

the probability density function of the mixture between two pure tissues, l_1 and l_2, varies depending on how much different σ_{l_1} and σ_{l_2} are and also depending on the difference between μ_{l_1} and μ_{l_2}.

Note that this more evolved intensity model that includes some mixture tissues does not actually add any additional parameter to the 3-class model. Only the *weight* of each new tissue type ($P(CG)$ and $P(GW)$) should be also considered.

B. Spatial distribution model

B.1 Markov Random Fields

The spatial information can be encoded in terms of correlated pixels using the theory of Markov Random Fields (MRF) to characterize relationships between spatial features [9]. The MRF theory, as in the case of Markov chains, considers that the dependence of one voxel state on the whole image information can be reduced to the information contained in a local neighborhood.

Then, all 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 said to be a MRF on S with respect to a neighborhood system N if and only if

$$P(\boldsymbol{x}) > 0, \boldsymbol{x} \in \mathcal{X}, \text{ and,} \tag{7}$$

$$P(x_i|x_{S-\{i\}}) = P(x_i|x_{\mathcal{N}_i}), \tag{8}$$

where x_i denotes the current estimate at location i, and $x_{S-\{i\}}$ denotes all the locations at $\mathcal{S}$ except i. According to the Hammersley-Clifford theorem, an MRF can be equivalently characterized by a Gibbs distribution,

$$P(\boldsymbol{x}) = Z^{-1}e^{-U(\boldsymbol{x},\boldsymbol{\beta})}, \tag{9}$$

that has several *free* parameters to be determined: the *normalization factor* Z, the *spatial* parameter β, and the *energy function* $U(\boldsymbol{x})$. Let us briefly discuss in what follows how these parameters can be determined in the particular framework of image segmentation.

B.2 The energy function $U(\boldsymbol{x})$

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 them is done in [13] where a general expression for the energy function is denoted by

$$U(\boldsymbol{x}|\beta) = \sum_{\forall i \in \mathcal{S}} \left(V_i(x_i) + \frac{\beta}{2} \sum_{j \in \mathcal{N}_i} V_{ij}(x_i, x_j) \right). \tag{10}$$

This is known as Potts model with an *external field*, $V_i(x_i)$, that weighs the relative importance of the different classes present in the image. Eq. (10) can be for instance modelled by an Ising model at 2 states [4]. However, the use of an external field includes additional parameter estimation, thus this model is less used in image segmentation [14]. Instead, a simplified Potts model with no external energy, $V_i(x_i) = 0$, is used. Then, only the local spatial transitions are taken into account and all the classes in the label image are considered equally probable. The key point is how to model $V_{ij}(x_i, x_j)$ to guide the final segmentation, $\hat{\mathbf{x}}$, as near as possible to the real image $\mathbf{x}^*$. They can be defined for instance as in [9]:

$$V_{ij}(x_i, x_j) = \delta(x_i, x_j) = \begin{cases} -1 & \text{if } x_i = x_j \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Intuitively, the equation above encourages one voxel to be classified as the tissue that the most of its neighbors belongs to. However, this model does not take into account the distance between neighbors but the class they belong to. It is not moreover well suited to model partial volume since it tends to eliminate it. A more evolved function is used in this work as proposed in [15], [16]:

$$V_{ij}(x_i, x_j) = \frac{\delta(x_i, x_j)}{d(i, j)}, \tag{12}$$

where,

$$\delta(x_i, x_j) = \begin{cases} -2 & \text{if } x_i = x_j \\ -1 & \text{if they share a tissue type} \\ +1 & \text{otherwise,} \end{cases} \tag{13}$$

and $d(i, j)$ represents the distance between voxels i and j. With this energy function configurations that are not likely to occur (e.g. CSF inside WM) are penalized while smooth transitions, more likely to occur in a brain (e.g. WM next to the partial volume GW), are encouraged.

B.3 The spatial parameter β

The *spatial* value of β controls the influence of the spatial prior over the intensity. Note that its influence on the final segmentation[4] is important. $\beta = 0$ corresponds to a uniform distribution over the $\mathcal{L}$ possible states, that is, the maximization is done only on the conditional distribution of the observed data $P(y|x)$ (Eq. (61)). On the contrary, if the spatial information is dominant over the intensity information, that is $\beta \to \infty$, MAP tends to classify all voxels to a single class [13].

The value of β can be estimated by ML estimation. However, many problems arise due to the complexity of MRF models and alternative approximations have to be done (for instance, Monte-Carlo Simulations or by maximum pseudo-likelihood [17]). The β parameter can be also determined *arbitrarily* as proposed in [18] by gradually increasing its value over the algorithm iterations. Here, the value of β has been fixed empirically by choosing the one that results in a better classification on a training set. In this work, β is fixed to 1.2.

B.4 The normalization factor Z

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

$$Z(U) = \sum_{\boldsymbol{x}} e^{-U(\boldsymbol{x}, \beta)}, \tag{14}$$

---

[4]Note that classification is done here by MAP estimation and this requires the application of the ICM algorithm. We refer to Appendix B for more details.

but it requires a high computational cost or it is even intractable since the sum among all the possible configurations of $\boldsymbol{x}$ is usually not known [19]. Note also its dependence on the definition of the energy function U. Instead, the conditional probabilities $P(x|x_{\mathcal{N}_i})$ can be easily normalized by forcing:

$$\sum_{\forall x_i \in \mathcal{L}_{pm}} P(x_i|x_{\mathcal{N}_i}) = 1. \tag{15}$$

B.5 Hidden Markov Random Fields

The theory of a *Hidden Markov Random Field (HMRF)* model is derived from *Hidden Markov Models* (HMM), which are defined as stochastic processes generated by a Markov chain whose state sequence cannot be observed directly (X), only through a sequence of observations (Y). Here we consider the special case since, instead of a Markov chain, a MRF will be used as the underlying stochastical process. The concept of a *hidden* MRF is different from that of an MRF, in the sense that HMRF is defined with respect to a pair of random variable families (X, Y) while MRF is only defined with respect to X.

In summary, a HMRF model is characterized by the following:

1. **Hidden Random Field (MRF)**: $X = \{X_i, i \in S\}$ is an underlying MRF assuming values in a finite state space $\mathcal{L}$ with probability distribution as defined in Eq. 9. The state of X is unobservable.

2. **Observable Random Field**: $Y = \{Y_i, i \in S\}$ is a random field with a finite state space $\mathcal{D}$. Given any particular configuration $\mathbf{x} \in \mathcal{L}^N$, every Y_i follows a known conditional probability distribution $P(y_i|x_i)$ of the same functional form $f(y_i; \theta_{x_i})$, where θ_{x_i} are the involved parameters. This distribution is called the *emission probability function* and Y is also referred to as the *emitted random field*.

3. **Conditional Independence**. For any $\boldsymbol{x} \in \mathcal{L}^N$, the random variables Y_i are supposed to be independent, which means that

$$P(\mathbf{y}|\boldsymbol{x}) = \prod_{i \in S} P(y_i|x_i). \tag{16}$$

Based on this, the joint probability of (X, Y) can be written as

$$P(\mathbf{y}, \mathbf{x}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x}) = P(\mathbf{x}) \prod_{i \in S} P(y_i|x_i). \tag{17}$$

According to the local characteristics of MRF's, the joint probability of any pair of (X_i, Y_i), given X_i's neighborhood configuration $X_{\mathcal{N}_i}$, is

$$P(y_i, x_i|x_{\mathcal{N}_i}) = P(y_i|x_i)P(x_i|x_{\mathcal{N}_i}). \tag{18}$$

So now it is possible to compute the marginal probability distribution of Y_i dependent on the parameter set θ (in this case θ is treated as a random variable) and $X_{\mathcal{N}_i}$,

$$\begin{aligned} P(y_i|x_{\mathcal{N}_i}, \theta) &= \sum_{\forall x_i \in \mathcal{L}} P(y_i, x_i|x_{\mathcal{N}_i}, \theta) \\ &= \sum_{\forall x_i \in \mathcal{L}} P(x_i|x_{\mathcal{N}_i})P(y_i|\theta_x), \end{aligned} \tag{19}$$

where $\theta = \{\theta_{x_i}, x_i \in \mathcal{L}\}$. A density function of this form is called *finite mixture (FM)* density. The conditional densities $P(y_i|\theta_{x_i})$ are called *component densities* and, in this case, they encode the intensity information. The *a priori* probabilities $P(x_i|x_{\mathcal{N}_i})$ are the *mixing parameters* and they encode the spatial information.

C. Anatomical prior model

In the previous section, the mixing parameters of the FM model encode the local spatial information. Other additional information could be used to define the energy function (Eq. (10)). For instance, for each tissue class, the probability of a voxel belonging to the class can be obtained after the registration with a probabilistic atlas. Intuitively, in this case, the accuracy of including the prior probability information depends on the errors of the registration process. It is not obvious where the anatomical prior probability should be introduced. It could be for instance included in the classification process as:

$$\hat{\mathbf{x}} = arg \max_{\mathbf{x} \in \mathcal{X}} \{ P(\mathbf{y}|\mathbf{x}) P(\mathbf{x}) P_A(\mathbf{x}) \}, \tag{20}$$

where $P_A(x)$ is the anatomical prior probability, according the chosen template. However, as proposed in [5], the anatomical probability influence can be better controlled if it is included in the energy function $U(\boldsymbol{x})$ by

$$V_{ij}(x_i, x_j) = \frac{\delta(x_i, x_j) - \gamma P_A(x_i)}{d(i, j)}, \tag{21}$$

where γ is a constant defined to control the influence of the probability maps over the local spatial information. Note that no external energy is considered.

No prior anatomical information of mixture tissues is usually considered in a class template. Then, anatomical prior of partial volume voxels can be computed from the pure tissue probability composing the mixture as proposed in [5]:

$$P_A(x) = 2\sqrt{P_A(l_1)P_A(l_2)}, \ x \in \mathcal{L}_{pm}, \ \text{and} \ l_1, l_2 \in \mathcal{L}_p \tag{22}$$

That is, PV voxels are assumed to be most likely at locations where the anatomical prior probability of both pure tissues within the voxels are high. Of course, Eq. (22) is arbitrary and its validity depends, in this case, on how the reference image used to assess the results is constructed. For instance, the reference image used here considers a most relaxed assumption of having a mixture tissue voxel when both pure tissues probabilities are different from zero. So, pure tissue and new mixture anatomical priors are arbitrarily raised to the power of ε_p and ε_m respectively in order to widen ($\varepsilon < 1$) or shrink ($\varepsilon > 1$) the tissue borders:

$$P'_A(x_i) = P_A^{\varepsilon_p}(x_i), \ x_i \in \mathcal{L}_p, \tag{23}$$

and,

$$P'_A(x_i) = P_A^{\varepsilon_m}(x_i), \ x_i \in \mathcal{L}_{pm} \setminus \mathcal{L}_p, \tag{24}$$

where $P'_A(\mathbf{x})$ denotes the new anatomical probability maps. Finally, all the probability maps are normalized so that they sum up to unity over all tissue classes.

III. CLASSIFICATION CRITERIA

A. Cost function

The notion of *cost function* should be recalled before introducing the Bayesian criterion. The classification process can be seen as an estimation problem: using the available data, the real value of the unknown labelling configuration, denoted by x^*, is estimated by $\widetilde{x}$, where both are a particular realization of random field X. The elementary cost function is defined as [20]:

$$L : \mathcal{L} \times \mathcal{L} \to \mathcal{R}^+ \tag{25}$$

$$L(\widetilde{x}, x^*) \begin{cases} = 0 & \Leftrightarrow x^* = \widetilde{x} \\ > 0 & \text{otherwise} \end{cases} \tag{26}$$

B. Bayesian criterion

The Bayesian estimation assumes that a cost function L is defined and that *a posteriori* distribution $p(x|y)$ and an observation y of Y are known. Then, the objective is to find an estimator $\widetilde{x}$ that minimizes the *Bayes risk*, that is, the expected cost. Mathematically,

$$\overline{L}(\widetilde{x}) = E[L(\widetilde{x}, X)|Y = y], \tag{27}$$

is the expected cost and,

$$\hat{\mathbf{x}} = arg \min_{x \in \mathcal{L}} \overline{L}(x) \tag{28}$$

is the Bayesian estimation of x^*.

The Bayesian strategy is optimal in the sense of the minimization of error probability. In fact, among all the other strategies it is the one for which the average cost is minimal.

Different cost functions can be defined. For instance, if a quadratic cost function is defined as

$$L(\widetilde{x}, x^*) = (x^* - \widetilde{x})^2, \tag{29}$$

the Bayes estimator is called Minimum Mean Squared Error (MMSE) estimator and it corresponds to the conditional mean of the posterior probability density function $p(x|y)$. If the cost function is defined by the absolute error

$$L(\widetilde{x}, x^*) = |x^* - \widetilde{x}|, \tag{30}$$

the Bayesian estimate is called Minimum Mean Absolute Error (MMAE) estimator and it corresponds to the median of $p(x|y)$.

C. Maximum a posteriori (MAP)

If the cost function is uniform,

$$L(\widetilde{x}, x^*) = \begin{cases} 0, & \Leftrightarrow x^* = \widetilde{x}, \\ 1, & \text{otherwise}, \end{cases} \tag{31}$$

the Bayes estimator is reduced to a Maximum *a posteriori* estimator (MAP). That is,

$$\overline{L}(x) = 1 - P(x|y), \tag{32}$$

and,

$$\hat{\mathbf{x}} = arg \min \overline{L}(x) = arg \max_{\mathbf{x} \in \mathcal{X}} \{P(\mathbf{x}|\mathbf{y})\}. \tag{33}$$

Note that MMSE, MMAE and MAP are the same if the posterior probability density function is a Gaussian.

IV. Parameter estimation of a stochastic process

It has been seen in the previous section that an optimal Bayesian classifier can be applied if the a posteriori probability density function is known. However, a complete knowledge of the probabilistic structure of the problem is rarely available [21] but it can be simplified if some assumptions on the available data can be made:

• The conditional density function, $P(y|x, \theta_x)$, has a known parametric form and it is uniquely defined by the value of the parameter vector θ_x.

• The set of unlabelled samples $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ are independent.

In what follows, the concept of Maximum Likelihood estimate and the Expectation Maximization algorithm used to find this estimate are briefly presented.

A. Maximum Likelihood (ML)

Let $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ be a set of unlabelled data with a marginal probability function (Eq. (19)) that can be written as

$$P(y, \theta) = \sum_{\forall x \in \mathcal{L}_{pm}} P(y, x, \theta_x) = \sum_{\forall x \in \mathcal{L}_{pm}} P(x)P(y|x, \theta_x), \qquad (34)$$

The likelihood of the observed sample is by definition the joint conditional probability :

$$L(\theta) = P(\mathcal{Y}|\theta) \qquad (35)$$

The maximum likelihood estimate $\widehat{\theta}$ is that value of θ that maximizes $L(\theta)$:

$$\widehat{\theta} = \arg \max_{\theta} L(\theta) \qquad (36)$$

Maximizing the likelihood is equivalent to making the derivative of the log-likelihood zero. The derivative of the log-likelihood can be expressed in terms of the expectation of the gradient with respect to the probability $P_{\theta}(x|\mathcal{Y}, \theta)$,

$$\frac{d}{d\theta} \log(L(\theta)) = E[\frac{d}{d\theta} \log P(\mathcal{Y}, x, \theta)] = 0. \qquad (37)$$

B. Expectation Maximization (EM)

Expectation Maximization (EM) is an iterative algorithm that estimates the maximum of the log-likelihood by solving:

$$\frac{d}{d\theta} \log(L(\theta)) = 0. \qquad (38)$$

Another way to solve the above equation is to determine θ that verifies:

$$E_{\theta}[\frac{d}{d\theta} \log P(\mathcal{Y}, x, \theta)] = 0; \qquad (39)$$

We can note that the unknown parameter θ appears in the expectation and in the derivative. The basic idea of the EM algorithm is to give a current $\theta(k)$ related to the expectation to make the solution easier (Expectation step). The algorithm is then reduced to give an initial solution $\theta(0)$ and to calculate at the $(k + 1)^{th}$ step the current estimation $\theta(k + 1)$ solution of :

$$E_{\theta_{(k)}}[\frac{d}{d\theta} \log P(\mathcal{Y}, x, \theta)] = 0; \qquad (40)$$

For any k step, this Expectation can be written as:

$$\frac{d}{d\theta} E_{\theta_{(k)}}[\log P(\mathcal{Y}, x, \theta)] = 0; \qquad (41)$$

So, the ML can be estimated by the maximization of $E_{\theta_{(k)}}[\log P(\mathcal{Y}, x, \theta)]$ instead of solving the annulling of the derivative equation (Maximisation Step). Then, the steps of the algorithm are:

Step 0: Choose the best initialization for $\theta(0)$.

Step (k+1): Calculate $\theta(k + 1)$ solution of $\max_{\theta} E_{\theta_{(k)}}[\log P(\mathcal{Y}, x, \theta)]$.

V. Parametric methods

Here the different parametric classification methods that participate in the comparative study are defined in detail. The methods that model all the brain tissues having a Gaussian distribution are described first. Then, the methods that consider a different intensity distribution for the partial volume voxels are presented. Finally, the method that consider *prior* tissue templates is presented.

A. Finite Gaussian Mixture Model: FGMM (A)

The finite Gaussian mixture model (FGMM) is one of the most commonly used approaches to solve the classification problem for MR images of the brain in its main tissues [6]. This model considers only the intensity information: each of the brain tissues is modelled by a Gaussian distribution. No spatial information is taken into account. Moreover the random variables X_i are assumed to be independent of each other, which means that,

$$P(x|x_{\mathcal{N}_i}) = P(x) = w_x \ , \forall x \in \mathcal{L}_{pm}, \text{ and } \forall i \in \mathcal{S}. \tag{42}$$

Then, Eq. 19 is reduced to

$$P(y|\theta) = \sum_{\forall x \in \mathcal{L}_{pm}} w_x \cdot P(y|x) = \sum_{\forall x \in \mathcal{L}_{pm}} w_x \cdot f_x(y|\theta_x), \tag{43}$$

where the component densities $f_x(y|\theta_x)$ are a Gaussian distribution defined by the parameters $\theta_x = (\mu_x, \sigma_x)$. The mixing parameters ω_x can also be included among the unknown parameters. Thus, the mixture density parameter estimation tries to estimate the parameters $\theta = (\omega_x, \theta_x)$ such that,

$$\sum_{x \in \mathcal{L}_{pm}} \omega_x = 1. \tag{44}$$

As presented in Section IV, a possible approach to solve the parameter estimation problem is to find the maximum of the log-likelihood function. One of the most used methods to solve the maximization problem is the EM algorithm (Section IV-B). For the particular case of Gaussian distributions, the resulting equations of the EM algorithm that numerically approximate the parameters of the mixture are:

Initialization Step. Choose the best initialization for $\theta(0)$.

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

$$\widehat{P}^{(k)}(x|y_i, \widehat{\theta}) = \frac{P(y_i|\widehat{\theta}_x^{(k-1)})\widehat{P}^{(k-1)}(x)}{\sum_{l, \forall l \in \mathcal{L}_{pm}} P(y_i|l, \widehat{\theta}_l^{(k-1)})\widehat{P}^{(k-1)}(l)} \tag{45}$$

Maximization Step:

$$\widehat{\omega}_x^{(k)} = \widehat{P}^{(k)}(x) = \frac{1}{N}\sum_{i \in S}\widehat{P}^{(k)}(x|y_i, \widehat{\theta}) \tag{46}$$

$$\widehat{\mu}_x^{(k)} = \frac{\sum_{i \in S}\widehat{P}^{(k)}(x|y_i, \widehat{\theta})y_i}{\sum_{i \in S}\widehat{P}^{(k)}(x|y_i, \widehat{\theta})} \tag{47}$$

$$(\widehat{\sigma}_x^{(k)})^2 = \frac{\sum_{i \in S}\widehat{P}^{(k)}(x|y_i, \widehat{\theta})(y_i - \widehat{\mu}_i^{(k)})^2}{\sum_{i \in S}\widehat{P}^{(k)}(x|y_i, \widehat{\theta})} \tag{48}$$

Note that, in this case, and also for GPV as it will be seen later, the sum among all the image voxels of Eq. (45) is equivalent to

$$\sum_{i \in S}\widehat{P}^{(k)}(x|y_i, \widehat{\theta}) \Longleftrightarrow \sum_{\forall y_i} h(y_i)\widehat{P}^{(k)}(x|y_i, \widehat{\theta}), \tag{49}$$

where h is the image histogram. This decreases significantly the number of computations to be made in Eq. (46), Eq. (47), and Eq. (48). Unfortunately, the methods using the HMRF model cannot use Eq. (49). Finally, once the estimation parameter problem is solved, the classification is performed by Bayesian rule (Section III).

B. Gaussian Hidden Markov Random Field model: GHMRF (B)

The theoretical concepts of this approach are the same as presented in Section II-B. As defined in Eq. 19, the intensity image distribution function, dependent on the parameter set θ and on the voxel neighborhood $x_{\mathcal{N}_i}$, is:

$$P(y|\theta) = \sum_{x \in \mathcal{L}_{pm}} P(x|x_{\mathcal{N}_i}) \cdot f_x(y|\theta_x), \tag{50}$$

where, $f_x(y|\theta_x)$, is, $\forall x \in \mathcal{L}_{pm}$, a Gaussian distribution (see Eq. (2)) defined by $\theta_x = \{\mu_x, \sigma_x\}$, and $P(x|x_{\mathcal{N}_i})$ represents the locally dependent probability of the tissue class x_i. Actually, if this equation is compared with Eq. (43), it can be seen that the FGMM model is a special case of an HMRF model.

To solve the parameter estimation problem, an adapted version of the EM algorithm, called the HMRF-EM, is used as suggested in [9]. The update equations for the θ parameters are actually the same update equations as for the FGMM (see Eq. (46), Eq. (47), and Eq. (48)), except that

$$\widehat{P}^{(k)}(x|y_i, \widehat{\theta}) = \frac{P(y_i|\widehat{\theta}_x^{(k-1)}) \cdot \widehat{P}^{(k-1)}(x|x_{\mathcal{N}_i})}{\sum_{l, \forall l \in \mathcal{L}_{pm}} P(y_i|l, \widehat{\theta}_l^{(k-1)}) \widehat{P}^{(k-1)}(l|l_{\mathcal{N}_i})}. \tag{51}$$

The calculation of $P^{(k-1)}(x|x_{\mathcal{N}_i})$ involves a previous estimation of the class labels, $\hat{\mathbf{x}}$, that is, the classification step. In fact, the strategy underlying the EM algorithm consists of applying iteratively the following steps:
1. Estimate the image labelling, $\hat{\mathbf{x}}$, 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, $\mathcal{E}[\log P(\mathbf{x}, \mathbf{y}|\theta)]$. The classification step is actually obtained through a MRF-MAP estimation (refer to Appendix B for more details).

C. Gaussian and Partial Volume model: GPV (C)

The approach described here only uses the intensity information as in the FGMM. It exactly follows the image model defined in Section II-A that considers a density function for the mixture brain tissues different from a Gaussian distribution. Then, the same probabilistic model as in Eq. (43) is used but, in this case, $P(y_i|x, \theta_x)$ is defined either by a Gaussian or by a PV equation Eq. (6). Finally, the following minimization problem is defined:

$$\widehat{\theta} = \min_{\theta} \sum_{\forall y_i} (h_n(y_i) - p(y_i|\theta))^2, \tag{52}$$

where h_n denotes the normalized intensity histogram. The *genetic algorithm* presented in [6] is used to solve the estimation problem. Fewer parameters have to be estimated 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 FGMM, once the distribution parameters are found, the classification is done following the Bayesian rule.

D. GPV and HMRF model: GPV-HMRF (D)

This method adds to the GPV approach the spatial information that is encoded following the HMRF theory. As usual, the same probabilistic model as in Eq. (50) is defined and, as in method GPV, $P(y_i|x, \theta_x)$ is defined either by a Gaussian or by a PV eqution Eq. (6).

The estimation parameter problem is solved almost identically as for method GHMRF. An adapted version of the EM-algorithm is used as proposed in [5]:

$$\widehat{P}^{(k)}(x|y_i, \widehat{\theta}) = \frac{\widehat{P}(y_i|x, \widehat{\theta}_x^{(k-1)}) \cdot \widehat{P}^{(k-1)}(x|x_{\mathcal{N}_i})}{\sum_{l, \forall l \in \mathcal{L}_{pm}} \widehat{P}(y_i|l, \widehat{\theta}_l^{(k-1)}) \widehat{P}^{(k-1)}(l|l_{\mathcal{N}_i})}, \tag{53}$$

$$\mu_x^{(k)} = \frac{\sum_{i \in S} \widehat{P}^{(k)}(x|y_i) y_i}{\sum_{i \in S} \widehat{P}^{(k)}(x|y_i)}, \tag{54}$$

$$\left(\sigma_x^{(k)}\right)^2 = \frac{\sum_{i \in S} \widehat{P}^{(k)}(x|y_i)(y_i - \mu_x^{(k)})^2}{\sum_{i \in S} \widehat{P}^{(k)}(x|y_i)}. \tag{55}$$

Note that, in this approach, the updating equations Eq. (54) and Eq. (55) are only computed for pure tissues ($x \in \mathcal{L}_p$), and that $\widehat{P}^{(k)}(y_i|x, \widehat{\theta_x})$ is now either a Gaussian or a PV distribution. The strategy underlying the EM algorithm is similar to that of the GHMRF method. However, in this case, the calculation of $P^{(k)}(x|x_{\mathcal{N}_i})$ does not involve a previous estimation of the class labels since $\hat{\mathbf{x}}$ since spatial prior is retrieved from:

$$\hat{\mathbf{x}} = arg \max_{\mathbf{x} \in \mathcal{X}} \{P(\mathbf{y}|\mathbf{x})\}, \tag{56}$$

Finally, the classification is done by the MRF-MAP step:

$$\hat{\mathbf{x}} = arg \max_{\mathbf{x} \in \mathcal{X}} \{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})\}, \tag{57}$$

and no minimization of the energy can be computed instead of Eq. (57) (as it is done in Appendix B to solve the MAP estimation in GHMRF) since $P(\mathbf{y}|\mathbf{x})$ does not always follow a Gaussian distribution.

E. GPV-HMRF model and Anatomical prior: GPV-HMRF-AP

The GPV-HMRF-AP method segments the brain tissues according to the image model presented in the previous section. Moreover, several anatomical prior models (see Eq. (21)) are considered:

1. **GT.** The ground truth class priors are considered first. However, it is noticed that adding such *perfect* prior class templates is not realistic since they are not available. In practice, it is used here only as a basis for comparison with other templates. The construction of mixture tissue probability maps is done as presented in Section II-C using $\gamma = 2$. Then, both PV and pure tissue prior probabilities have been raised to $\epsilon_m = \frac{1}{6}$ and $\epsilon_p = 6$ respectively. The resulting normalized class templates are shown in Figure 2.

2. **GTC.** As proposed in [5], the ground truth class templates are slightly corrupted by rotation (1 degree in the axial plane) and translation (2 mm in direction of the axial plane normal vector) in order to simulate registration errors. Here, γ is equal to 1 in order to make the prior class information less important than the local priors since some errors have been introduced.

3. **SPM.** The probability maps of CSF, GM and WM used in SPM [22] are also considered as class priors (see Section **??**). These templates are almost in the same reference as the Brain Web phantom, thus a rigid transformation would be enough to globally register both the phantom and the SPM class templates. However, a non-rigid registration between the phantom image and the T1 average image of SPM is done in order to make the SPM maps less smooth and more similar to the phantom anatomy. Then, mixture maps have been created as done for the ground truth priors. No change on the border tissues is made (ε's are equal to 1) in order not to introduce many errors since the SPM probability maps are very smooth. γ is, as for GTC, arbitrarily fixed to 1. The resulting templates are shown in Figure 3.

VI. Non parametric methods

In the previous sections, parametric segmentation algorithms were introduced, which means that the intra-class probability densities $P(y|x)$ are modelled by a family of parametric functions $f_x(y|\theta_x)$, such as Gaussian densities. The success of the resulting algorithms is therefore reliant upon the choice of an appropriate family of parametric functions. However, if no well justified parametric model of the data is known, parametric approaches could dramatically fail. Thus, non-parametric, information
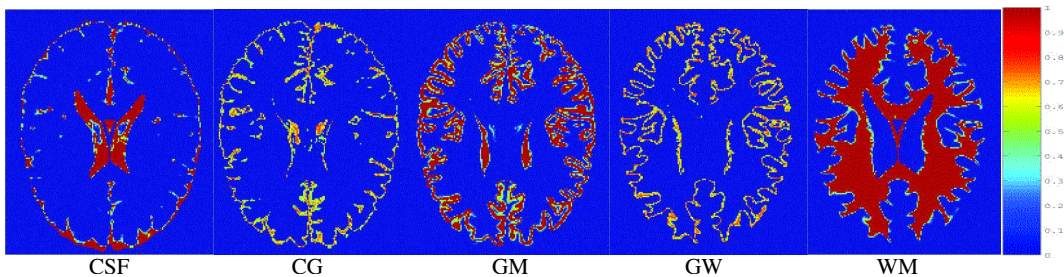
Fig. 2.   Probability maps for the 5 brain tissues constructed from the ground truth.
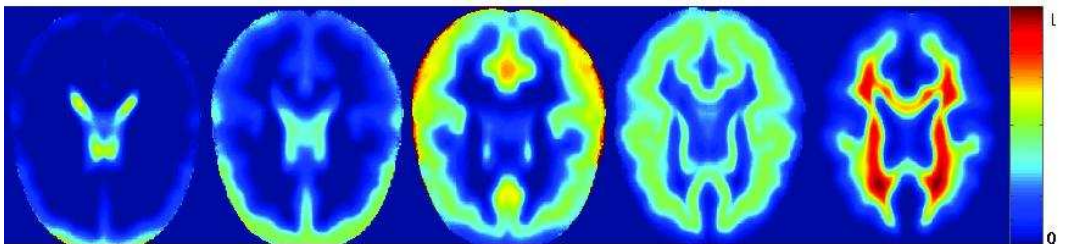

Fig. 3.   From left to right: CSF, CG, GM, GW and WM probability class maps constructed from SPM maps.

theoretic alternatives are introduced in what follows. The two non-parametric approaches assessed in this comparative study have been developed and implemented by Butz [10]. It is beyond the scope of this section to present in detail the framework he proposed. Let's however summarize the main concepts of his approach.

A. Error probability minimization: EP (E)

Let us consider a random variable different from X, called X^{est}, also over $\mathcal{L}$, which models an estimation of X from the observable data, Y. Naturally, the following stochastic process can be built:

$$X \rightarrow Y \rightarrow X^{est} \rightarrow E, \tag{58}$$

where E is an error random variable being 1 whenever the estimated class label, x^{est}, is considered a wrong estimate of the initial class label, x, and 0 otherwise. A key quantity of Eq. (58) is the probability of error, $P_{e|\mathbf{x}}$, of the transmission from X to X^{est}, for a given class map, $\mathbf{x}$. This probability also equals the expectation of E.

Considering the introduced formalism (see [10]), the information theoretic classification objective consists of determining the class label map $\hat{\mathbf{x}}$ that minimizes an error probability $P_{e|\mathbf{x}}$:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} P_{e|\mathbf{x}}. \tag{59}$$

B. Non-parametric HMRF: NP-HMRF (F)

The approach proposed above does not consider any spatial priors on the class label map. However, the probabilistic nature of the formalism allows the addition of a HMRF, just as for the parametric approaches introduced in the previous sections, resulting in a non-supervised non-parametric hidden markov model (NP-HMRF) segmentation:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathcal{L}} P(\mathbf{x}) \cdot \mathbf{P_{e|x}}. \tag{60}$$

The optimization objective above is called the minimal error probability principle for NP-HMRFs. In complete analogy to parametric HMRFMs, the prior probabilities, $P(\mathbf{x})$, are modeled by a Gibbs

distribution (Section II-B.5). The derived non-parametric framework for classification allows the consideration of voxel features for which any particular parametric model is known, as it is the case for e.g. voxel gradients.

VII. Results and Validation

A. Data set

All the methods have been validated using the *digital brain phantom* from the McConell Brain Imaging Center [11]. They provide an MRI simulator where different RF non-uniformity (bias of 0%, 20%, and 40%) and noise levels (0%, 1%, 3%, 5%, 7%, and 9%) can be added to the MR brain phantom image.

Then, a 5-class (CSF, CG, GM, GW and WM) ground truth classification image, Figure 6(b), has been created from the 3-dimensional 'fuzzy' tissue membership volumes provided by [11] where voxel values reflect the proportion of tissue present within the voxel. This makes these images suitable for segmentation assessment. Finally, a ground truth image histogram is created by splitting each image histogram into the specific pure tissue and their mixture histograms (see Figure 7(a)).

B. Results

Validation is made by comparing the results obtained with the classification methods presented in Section V and Section VI to both the 5-class ground truth classification image and the brain phantom image histograms. Because of limited space, only a complete study of these results for brain phantom image of 7% Noise (N) and 20% of in-homogeneity (RF) is shown here[5]. Also, note that the analysis of the methods including anatomical prior is done separately in comparison with GPV-HMRF and that GPV-HMRF-AP are tested in few brain phantom images.

First, each of the resulting volumes classified by each of the algorithms is qualitative validated visually. 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 (see Figure 6).

Second, the intensity image model is assessed by comparing the histogram fitting to the ground truth brain phantom image histogram (see Figure 7).

Third, quantitative analysis is performed by computing the *confusion tables* with respect to the 5-class reference classification (Table I). These values assess the quality of the classification for each tissue class.

Fourth, global measures of quality (*Pergood, Perhalf* and *Perfault*) are presented in Table **??**. Percentages are always computed with respect to the 5-class ground truth classification and voxels belonging to the brain phantom background are not considered. *Pergood* is the percentage of voxels correctly classified (confusion table diagonal). *Perghalf+* and *Perghalf-* represent the percentage of voxels that has not been correctly classified but misclassified into a neighbor tissue, e.g. a WM voxel classified as WG, ('+' and '-' refer to superior and inferior of the confusion table diagonal, respectively). *Perfalse* is the percentage of voxels that has been completely misclassified.

Fifth, the robustness in front on the noise and in-homogeneities is analyzed separately for each method in Figure 5.

Sixth, a global assessment is done for all possible noise and inhomogeneity levels of the digital brain phantom. Both percentage of the correct and false classification are showed in Figure 10 and Figure 11 for all methods.

Finally, GPV-HMRF-AP is only applied to the 5N0RF, 7N20RF, and 9N40RF phantoms. All the measures presented before (classified images, histogram fitting and confusion table) are considered here to study the influence of the different class templates on the final classification (see Figure 8, Figure 9

---

[5]In order to simplify the notation, the phantom will be denoted by 7N20RF.

and Table II). For each phantom the ground truth class priors (GT), the corrupted ground truth class priors (GTC), and the probabilistic class maps of SPM are compared with respect to GPV-HMRF method.

VIII. Discussion

A. General performance

One of the goals of this comparative study is to be able to specify the most suitable tissue classification technique for T1-MR brain image. Unfortunately, there is not a single winner. Actually, the answer depends on the noise (N) and the in-homogeneity (RF) level present in the images. It is considered that the best classification corresponds to the highest percentage of correct classified voxels (*pergood*). For low levels of noise ($N = \{0, 1, 3\}\%$), it is not evident to determine a method that better classifies than others (as we can see in In Figure 10). However, for higher levels of noise ($N = \{5, 7, 9\}\%$), method GPV-HMRF has almost always performed the best classification closely followed by method GHMRF (their *pergood* differs from less than 2%). Now, we can also determine the methods that perform smaller errors (lowest *perfault*). In this case, method GPV and GPV-HMRF (both using PV equation) always have the lowest *perfault* for low and high noise levels respectively. However, differences between all *perfault* values are not more than 1%.

B. Real MRI conditions

A wide range of noise level exists in the brain phantom simulator but actually not all of these values are realistic to represent the noise present in a *typical* T1-weighted MR brain image. The signal to noise ratio (SNR) in a *normal* T1-MR image has been computed and, then, it has been compared with the SNR present in the brain phantoms. The conclusion is that a *normal* noisy image corresponds to the mean of the phantom 5N0RF and 7N0RF. Thus, from now on conclusions are based on the classification results of phantoms with $N = \{5, 7, 9\}\%$ and $RF = \{0, 20, 40\}\%$. For these ranges of noise and inhomogenities, method GPV-HMRF has almost always the highest *pergood* and the lowest *perfault*. It is always closely followed by method GHMRF that usually differs from less than 2% from the *pergood* and less than 0.1% from the *perfault*.

C. Pure tissues and partial volume

In this work, a T1-MR brain image is modelled by three main tissues (CSF, GM and WM) and two mixtures (CG and GW). As is done by most of the methods described in the literature, the two other possible mixtures, CW and CGW, have been ignored. Actually, the importance of CW and CGW has been measured from the digital brain phantom: 12.8% of the image voxels are CSF, 18% CG, 26% GM, 20% GW and 23% WM while only 0.18% of the images voxels belong to the CW and only 0.02% to the CGW. Visually, the probability density function of the CSF and WM mixture has been drawn in Figure 4. Thus, it is justified to affirm that P(CW) and P(CGW) do not significantly contribute in the total probability density function of the MR intensity image.

Thanks to the confusion tables, the study of the classification score for each tissue class becomes an easy task. The best classification for CSF is performed (the 70% of the cases) by method EP, for GM it is method NPHMM (also the 70% of the cases) and method GHMRF performs for more than 50% of the cases the best classification of tissue WM. Method GPV-HMRF almost always achieves the best classification score for both partial volume tissues: 78% of the cases for CG and 100% for GW.

Results show that partial volume distributions are hardly well represented by a Gaussian function. This is obvious when looking at the histogram fitting where CG and GW mixtures are always better fitted by methods C and D using the partial volume equation (Figure 7). In fact, even if the mixtures may look like a Gaussian for high levels of noise and inhomogeneities, the assumption of using a normal distribution for a PV is false. 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 voxels correctly
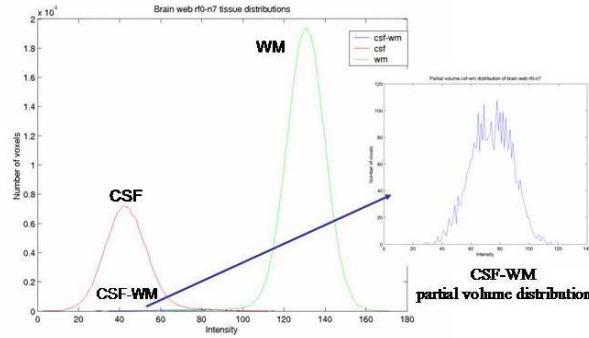
Fig. 4.   Intensity probability distribution of CSF and WM mixture.

classified. This poor result indicates that partial volume distribution does not seem to be completely well modelled yet. Thus, more work has to be done in the study and modelling partial volume intensity distribution. For instance, two different mixtures between GM and WM can be considered as recently suggested in [23]. They propose a pioneer anatomical model that splits the GM and WM mixture into a *geometrical* mixture corresponding to the brain cortico-subcortical interface and a *mosaic* GW mixture corresponding to the deep cerebral nuclei structures such as the thalamus.

D. Robustness in front of noise and inhomogeneities

None of the classification methods under study tries to compensate for image artifacts such as noise or bias. No pre-processing is applied for image quality enhancement: neither an anisotropic filtering nor a bias correction are considered. This way the robustness of the methods can be analyzed with these artifacts. This is clearly shown in Figure 5 where all possible levels of noise and inhomogeneities present in the brainweb simulator are considered again. Methods that consider only intensity information are represented in the left column. In general, the quality of the classification decreases with increasing noise and non-uniformities. Method A is very sensitive to both noise and inhomogeneities. However, for low levels of noise, methods C and E are equally performant in RF=0 than in RF=20. For very high noise levels (N={7,9}), all methods perform a classification that converges towards a range of *pergood* equal to [60-65]% for any value of RF. The right-hand column represents all the methods using HMRF. All of these methods present exactly the same behavior with noise and bias. If we consider RF=0, *pergood* decreases proportionally to the increment of noise. For RF=20, there is not a decrease of quality but almost a constant value of *pergood*. And, for RF=40, the *pergood* even increases for high noise levels. That is due to the fact that the phantoms considering low noise levels (N={0,1,3}) are actually not realistic to model T1-weighted MR brain images. Then, given a constant level of noise, RF=40 always makes *pergood* decrease about 12% for low and about 7% for high noise levels.

E. Intensity versus spatial prior

It can be seen in Figure 6 that the classification based only on intensity information (methods FGMM, GPV, and EP) is much more noisy than classification that also encodes spatial information. Errors are due to the overlap between tissue distributions and this overlap is larger for higher values of N and RF. On the contrary, when spatial information is also used in the classification process results are much less noisy: methods B, D and F improve the *pergood* percentage, with respect to methods A, C and E, by a 7% on average. However, they still make some errors mostly in the mixtures classification because the partial volume distribution model is probably not well-suited but also because of the MRF. In fact, results show that MRF considerably increases the classification quality and that makes the algorithms more robust when faced with noise than the intensity-based approaches. More evolved MRF are needed though in the particular case of T1W MR image segmentation. Recently, it has been
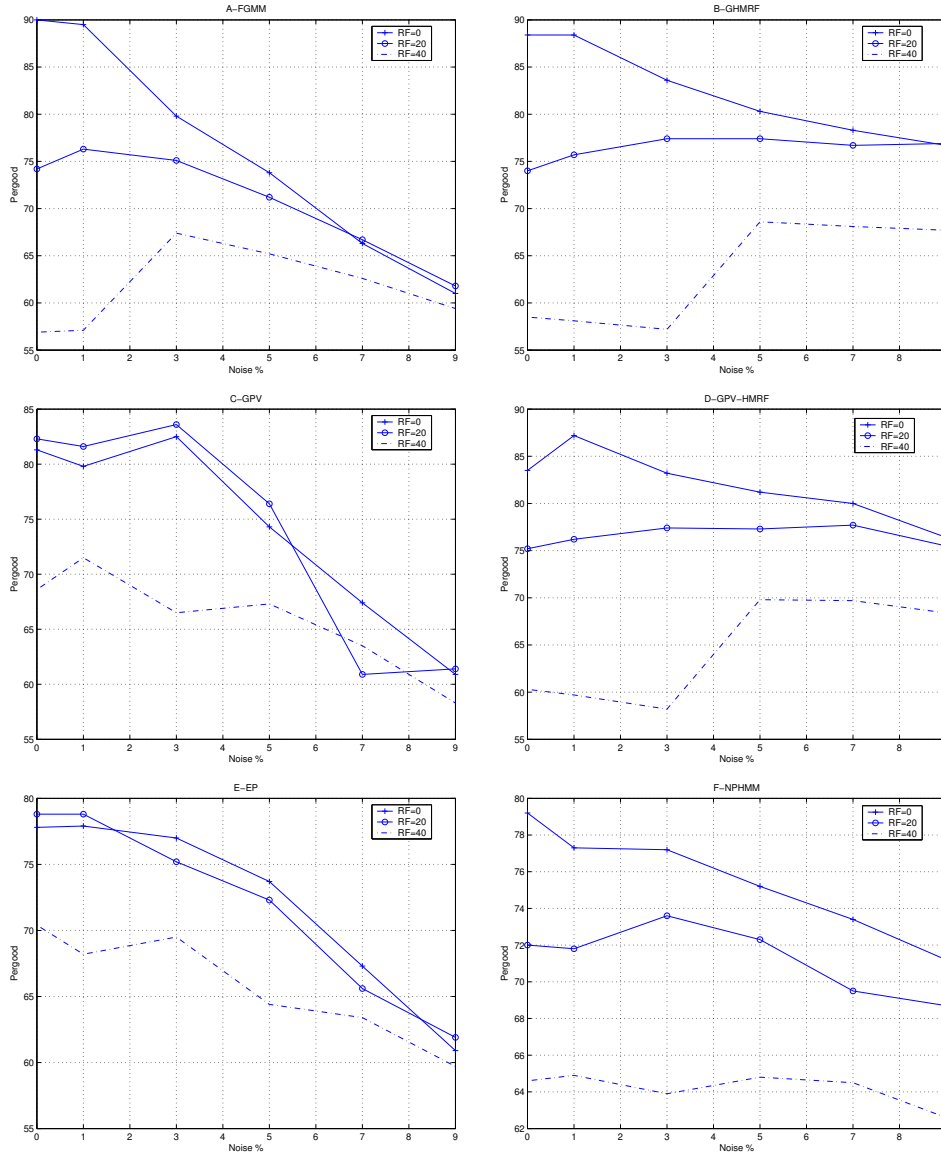
Fig. 5. Robustness of the classification methods in front of different levels of noise and inhomogeneities.

suggested to model either pure or mixture brain tissues with different MRF parameters [4]. Also, the addition of atlas information in the energy function could better guide the MRF model as proposed in Section II-C and in [5].

F. Parametric vs Non-parametric

Non-parametric models have performed in many cases equally or even better than parametric approaches. EP has slightly lower *Pergood* than FGMM using a Bayes classification or GPV for low levels of noise and non-uniformities. But almost the same quality of classification or even better than parametric models has been obtained by EP for high levels of noise N={5,7,9}.

When spatial information is also included, parametric models (GHMRF and GPV-HMRF) have almost always better *Pergood* than non-parametric approaches (there is usually a difference of 6% between them). Also, parametric methods commit fewer errors, they have a lower *Perfault*, than NP-HMRF. The misclassification made with both non-parametric approaches is mainly due to an overestimation of both mixture classes.

In conclusion, a non-parametric approach is more performant if no well justified assumption about

the data model can be made. However, logically, a good data model is better performant than no model.

G. Using prior class atlases

The results of the methods that include atlas information are discussed here in comparison with the method GPV-HMRF. Three different class priors have been added to the spatial model used in D and they are denoted by GT, GTC, and SPM (see Section V-E). The global performance of these four approaches is presented in Figure 12. The GT prior leads logically to the best results: the highest percentage of voxels correctly classified, around 88%, and the lowest percentage of fatal errors, around 0.17%, for any level of noise or inhomogeneity. However, as has been said before, the use of such a *perfect* class prior as GT is not possible in practice. The atlas information introduced by GTC and SPM are more realistic but results show that they do not always improve the results performed by D. Actually, significant changes have only been obtained for the 9N40RF phantom: the *perfault* is reduced from 0.79% to 0.57% for both GTC and SPM and the *pergood* is improved by a 5% with GTC. All resulting classified images look similar (see Figure 8). Almost no noise is visible either for GT or for GTC methods. Methods D and SPM lead to slightly noisy classifications for 7N20RF and 9N40RF phantoms.

Pure tissues are always better classified by GT and GTC than D while SPM only improves GM (in all three phantoms) and WM (5% and 7% of noise) classification. Significant errors are though introduced by GTC and SPM in the classification of partial volume voxels (see the histogram fitting in Figure 9). This effect is quantified in the confusion tables by a percentage of mixture voxels correctly classified much lower in GTC and SPM than in D or GT (from 2 to 12 % of degradation).

Finally, notice that rotation and translation of GT have a significant influence on PV classification while pure tissue classification remain robust with these *simulated* registration errors. Also, SPM probabilistic atlas has not demonstrated important improvements with respect to D. That is probably because SPM maps are too smooth and no anatomical variability is present in the prior class templates, thus, the information added by SPM is not precise.

IX. Conclusions

A validation study on MR brain tissue classification techniques has been proposed in this chapter. Both parametric and non-parametric approaches have been assessed in this work. Intensity-based classification methods are compared to the techniques that add spatial prior. The effect of considering prior class templates is also studied.

All tests have been done in several phantoms considering different noise and intensity non-uniformity levels. Then, the assessment is done by comparing to a 5 class ground truth image.

Results have shown that the techniques considering spatial information lead in better classification when high noisy images are considered while for low level of noise and in-homogeneities (that is not necessarily near real MR images) histogram-based techniques lead to comparable results. However, it has been demonstrated that percentage of correct classification never reaches the 100% and, even if pure tissues are in general well-classified, partial volumes are still not.

Methods including atlas information have not considerably improved the final classification with respect to the techniques that model local spatial priors. On the contrary, classification has shown to be highly sensitive to the registration errors or to the use of a wrong template. Actually, mixture tissues are particularly affected by prior class template errors while pure tissue classification has been almost always improved by these methods. This is because the initial pure class templates are not precise enough (too smooth or errors because of registration are present) but probably also because PV prior class maps are not optimally defined. In conclusion, no atlas class prior should be included if its quality cannot be assessed before.

Finally, we plan to measure the effect of pre-processing the images (by an anisotropic filter or a bias

corrector) or adding a bias field estimation model (as proposed by [24], [25] for instance). We expect both the pre-processing and bias model to make the classification more robust at high levels of noise and inhomogeneities. However, we suspect the pre-processing to *displace* partial volume voxels, thus some errors would probably be added in mixture tissue classification.

References

[1] 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," *Magnetic Resonance Imaging*, vol. 13, pp. 343–368, 1995.

[2] K. V. Leemput, *Quantitative Analysis of Signal Abnormalities in MRI imaging for Multiple Sclerosis and Creutzfeldt-Jacob Disease.* PhD thesis, Faculteit Toegepaste Wetenschappen, Faculteit Geneeskunde, Katholieke Universiteit Leuven, Belgium, May 2001.

[3] 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. 1277–83, 2000.

[4] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "A unifying framework for partial volume segmentation of brain mr images," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 105–119, 2003.

[5] A. Noe, S. Kovacic, and J. C. Gee, "Segmentation of cerebral mri scans using a partial volume model, shading correction, and an anatomical prior," in *SPIE Medical Image Processing*, 2001.

[6] P. Schroeter and et al., "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Transactions on Medical Imageing*, vol. 17, no. 2, pp. 172–186, 1998.

[7] S. Ruan, C. Jaggi, J. Xue, and J. Bloyet, "Brain tissue classification of magnetic resonance images using partial volume modeling," *IEEE Transactions on Medical Imaging*, vol. 19(12), pp. 172–186, 2000.

[8] P. Santago and H. D. Gage, "Quantification of mr brain images by mixture density and partial volume modeling," *IEEE Trans. Medical Imaging*, vol. 12, pp. 566–574, 1993.

[9] Y. Zhang and et al., "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE Trans. Medical Imaging*, vol. 20, pp. 45–57, 2001.

[10] T. Butz, *From error probability to information theoretic signal and image processing.* PhD thesis, Signal Processing Institut, Swiss Federal Institut of Technology, Switzerland, June 2003.

[11] 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 Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463–468, 1998. http://www.bic.mni.mcgill.ca/brainweb/.

[12] M. Bach Cuadra, B. Platel, E. Solanas, T. Butz, J.-Ph. Thiran, "Validation of Tissue Modelization and Classification Techniques in T1-Weighted MR Brain Images," in *In Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 290–297, October 2002.

[13] N. Peyrard, *Approximations de type champ moyen des modles de champ de Markov pour la segmentation de donnes spatiales.* PhD thesis, Universit J. Fourier, Grenoble, October 2001.

[14] G. Celeux, F. Forbes, and N. Peyrard, "EM-based image segmentation using Potts models with external field," Tech. Rep. 4456, INRIA, April 2002.

[15] 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.

[16] A. Noe and J. C. Gee, "Partial volume segmentation of cerebral mri scans with mixture model clustering," in *Information Processing in Medical Imaging: 17th International Conference (IPMI)*, 2001.

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

[18] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, vol. 48, no. 3, pp. 259–302, 1986.

[19] 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.

[20] B. Chalmond, *Elements de modelization pour l'analyse d'images.* Springer-Verlag, 2000.

[21] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[22] K. Friston, J. Ashburner, J. H. C. D. Frith, J.-B. Poline, and R. Frackowiak, "Spatial registration and normalization of images," *Human Brain Mapping*, vol. 2, pp. 165–189, 1995.

[23] T. Butz, P. Hagmann, E. Tardif, R. Meuli, and J.-P. Thiran, "A new brain segmentation framework," *In Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2003.

[24] W. Wells, R. Kikinis, W. Grimson, and F. Jolesz, "Adaptive segmentation of mri data," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 429–442, 1996.

[25] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of mr images of the brain," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 885–896, 1999.

X. Figures
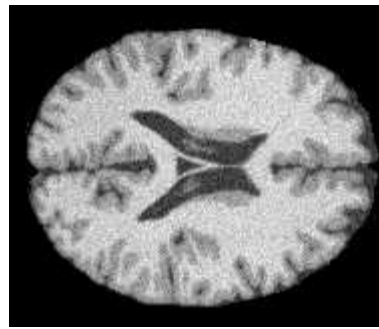
(a)Magnetic Resonance Image
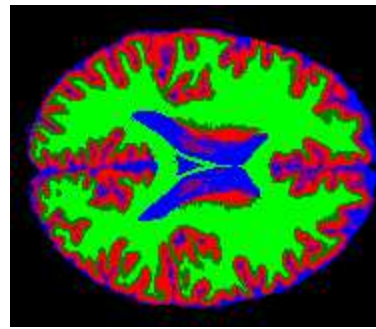
(b) 5 classes classification.
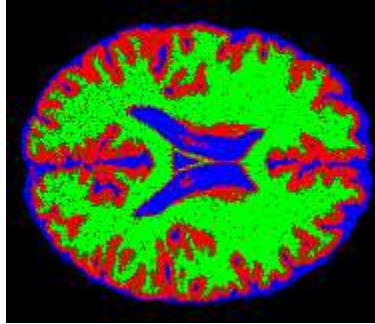
(c) Method A: FGMM.
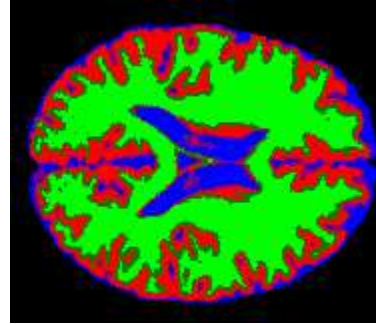
(d) Method B: GHMRF.
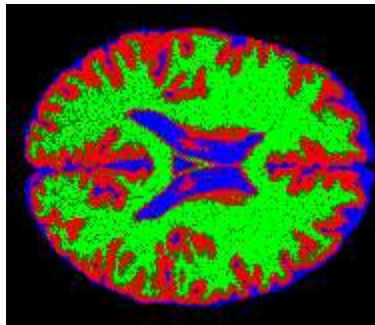
(e) Method C: GPV.
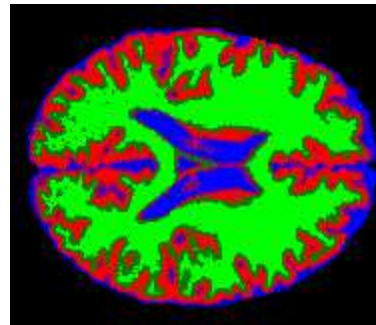
(f) Method D: GPV-HMRF.
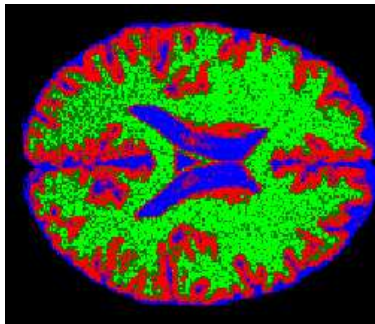
(g) Method E: EP.
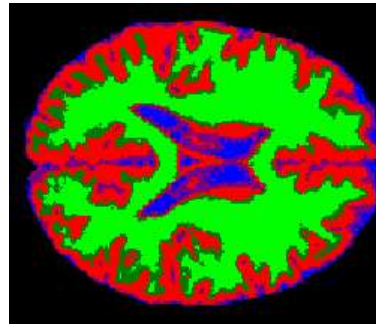
(h) Method F: NP-HMRF.

Fig. 6. Digital brain phantom T1-MRI with 7% noise and 20% RF and its classification. (a)Brainweb phantom simulated T1-MRI with 7% noise and 20% RF. (b) 5 classes ground truth created from Brainweb classification.
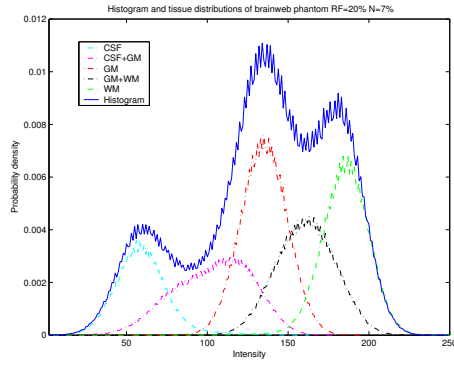
(a) 5 classes ground truth histogram.
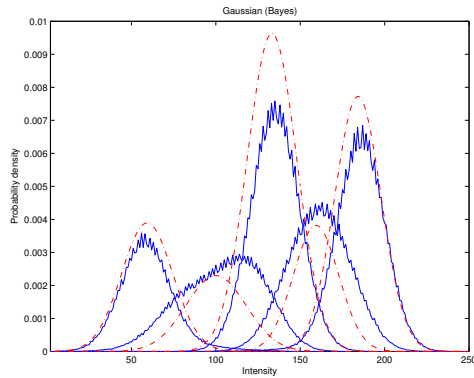
(a) Method A: FGMM.
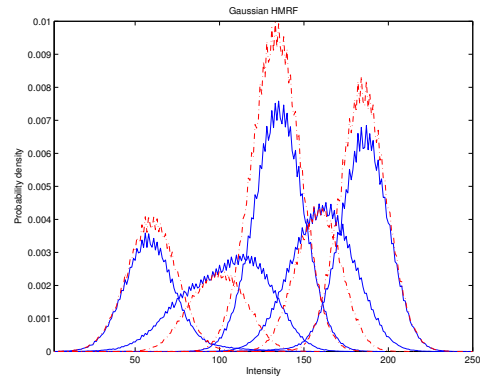
(b) Method B: GHMRF.
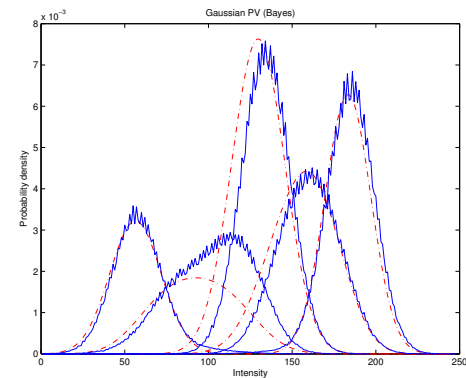
(c) Method C: GPV.
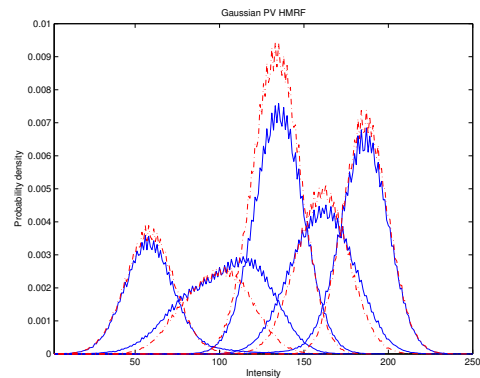
(d) Method D: GPV-HMRF.
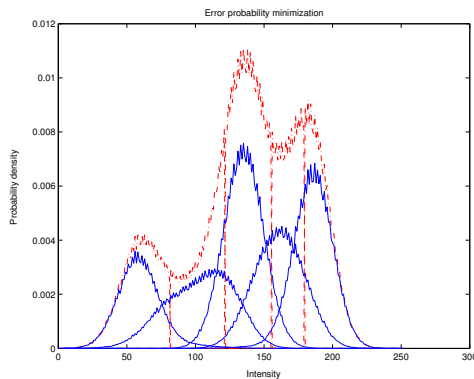
(e) Method E: EP.
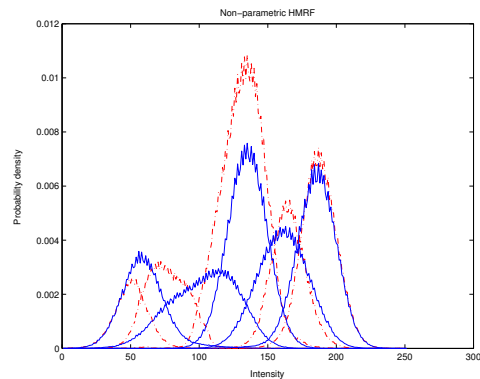
(f) Method F: NP-HMRF.

Fig. 7.   Histogram fitting of the Brainweb phantom 7N20RF. Results are in dotted line.
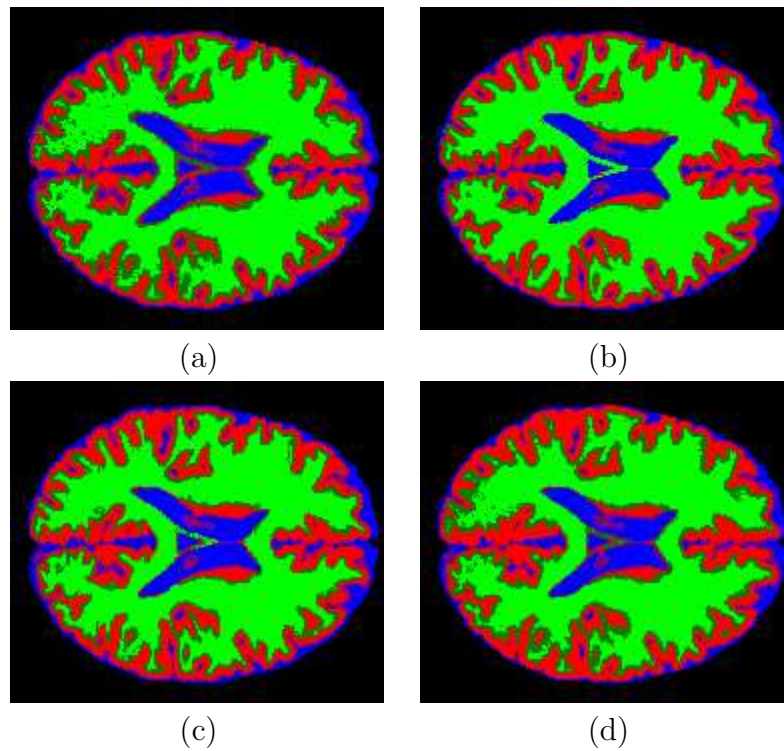
Fig. 8.   Classification image results of 7NRF using atlas prior. (a) Method D: GPV-HMRF. (b) Method D with GT. (c) Method D GTC. (d) Method D with SPM.
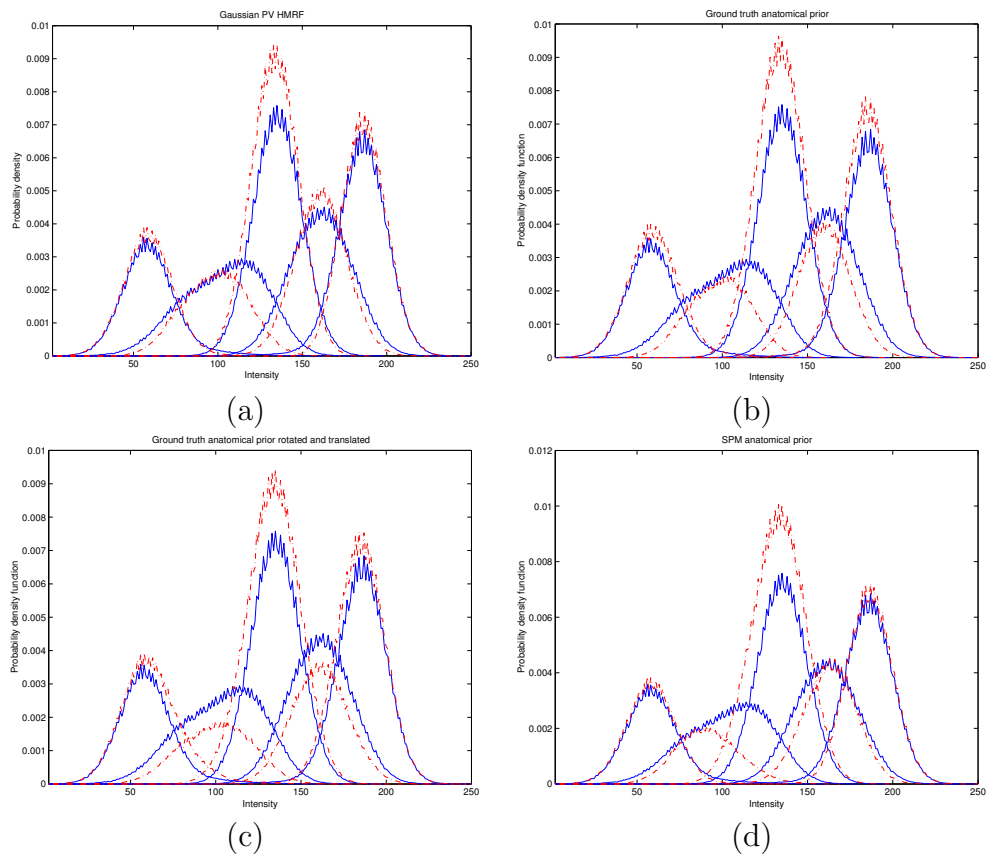
Fig. 9.   Histogram fitting of the phantom 7N20RF using atlas prior. Results are in dotted line. (a) Method D: GPV-HMRF. (b) Method D with GT. (c) Method D GTC. (d) Method D with SPM.

		Reference				
		CSF	CG	GM	GW	WM
A	CSF	**92.3**	20.9	0.1	0.0	0.0
	CG	6.5	33.6	3.3	0.1	0.0
	GM	1.2	44.9	87.8	36.3	2.2
	GW	0.0	0.6	7.7	26.9	8.4
	WM	0.0	0.0	1.2	36.6	89.4

		Reference				
		CSF	CG	GM	GW	WM
B	CSF	89.9	11.5	0.1	0.0	0.0
	CG	9.5	47.2	2.3	0.0	0.1
	GM	0.6	41.2	88.9	19.3	0.6
	GW	0.0	0.1	8.8	59.4	5.6
	WM	0.0	0.0	0.0	21.2	**93.7**

		Reference				
		CSF	CG	GM	GW	WM
C	CSF	67.0	9.4	0.1	0.0	0.0
	CG	31.7	48.3	7.0	0.6	0.1
	GM	1.3	39.0	69.8	26.3	2.6
	GW	0.0	3.2	21.8	45.8	26.9
	WM	0.0	0.1	1.3	27.3	70.4

		Reference				
		CSF	CG	GM	GW	WM
D	CSF	90.5	10.8	0.1	0.0	0.0
	CG	9.1	**57.2**	4.7	0.1	0.2
	GM	0.3	31.9	84.4	18.8	0.3
	GW	0.0	0.1	10.8	**66.3**	10.6
	WM	0.0	0.0	0.0	14.8	88.9

		Reference				
		CSF	CG	GM	GW	WM
E	CSF	91.5	19.8	0.1	0.0	0.0
	CG	8.0	55.0	17.3	1.4	0.1
	GM	0.5	24.7	74.8	36.9	2.4
	GW	0.0	0.5	7.7	45.7	30.6
	WM	0.0	0.0	0.1	16.0	66.9

		Reference				
		CSF	CG	GM	GW	WM
F	CSF	54.1	2.0	0.1	0.0	0.0
	CG	44.5	39.9	0.3	0.0	0.0
	GM	1.5	58.1	**93.7**	30.8	0.9
	GW	0.0	0.1	6.0	55.5	13.6
	WM	0.0	0.0	0.0	13.7	85.5

TABLE I

CONFUSION TABLE OF THE PHANTOM 7N20RF. PERCENTAGES ARE COMPUTED OVERALL VOXELS FOR EACH TISSUE TYPE.

		Reference				
		CSF	CG	GM	GW	WM
GT	CSF	**99.1**	10	0.1	0.0	0.0
	CG	0.4	**61.3**	0	0	0.2
	GM	0.2	28.7	**99.9**	10	0
	GW	0	0	0	**74.1**	0
	WM	0.2	0	0	15.9	**99.8**

		Reference				
		CSF	CG	GM	GW	WM
GTC	CSF	94.4	17.7	0.5	0	0
	CG	5.3	44.5	4	0	0.1
	GM	0.2	37.6	89	22.9	0.2
	GW	0.1	0.1	6.4	57.3	5.5
	WM	0	0	0	19.8	94.1

		Reference				
		CSF	CG	GM	GW	WM
SPM	CSF	89.8	10.1	0.1	0.0	0.0
	CG	9.9	56	4.2	0	0.2
	GM	0.2	33.8	85.6	20.2	0.3
	GW	0.	0.1	10.1	67.5	12.7
	WM	0.0	0.0	0.0	12.2	86.6

TABLE II

CONFUSION TABLE OF THE PHANTOM 7N20RF USING ATLAS PRIOR.

Intensity Classification
Percentage of Correct Classification

	0/0	0/20	0/40	1/0	1/20	1/40	3/0	3/20	3/40	5/0	5/20	5/40	7/0	7/20	7/40	9/0	9/20	9/40
Finite Gaussian Mixture Model (A)	90.0	74.2	56.9	89.5	76.3	57.1	79.8	75.1	67.4	73.8	71.2	65.2	66.3	66.7	62.6	61.0	61.8	59.4
Gaussian and Partial Volume (C)	81.3	82.3	68.6	79.8	81.6	71.5	82.5	83.6	66.5	74.3	76.4	67.3	67.4	65.6	63.5	60.9	61.4	58.3
Error Probability (E)	77.8	78.8	70.4	77.9	78.8	68.2	77.0	75.2	69.5	73.7	72.3	64.4	67.3	65.6	63.4	60.9	61.9	59.7

HMRF Classification
Percentage of Correct Classification

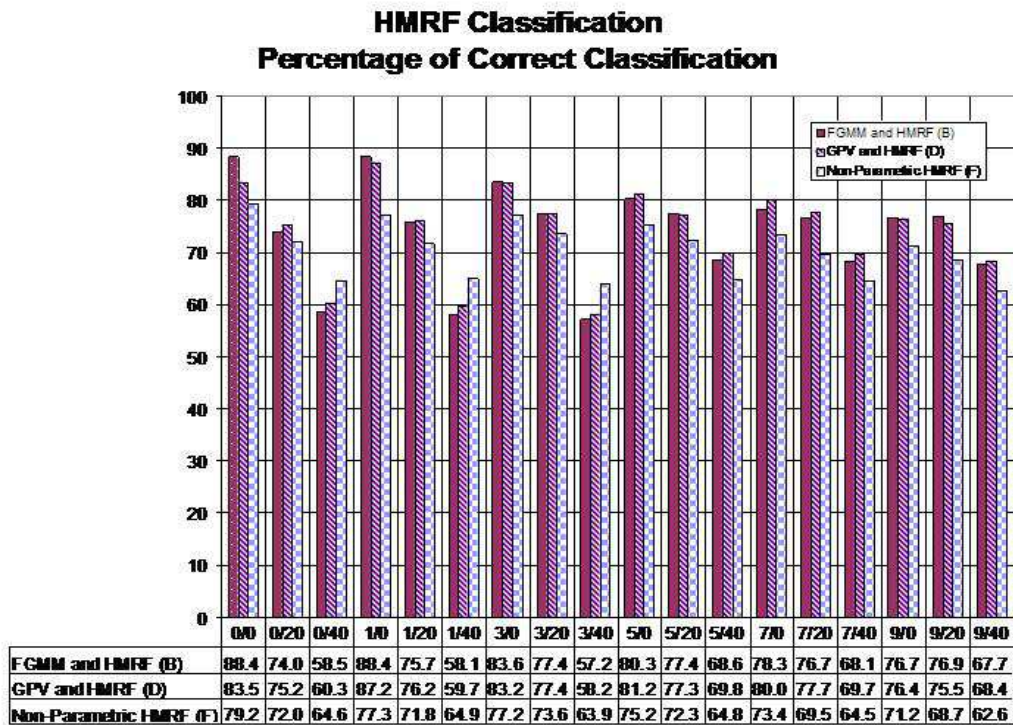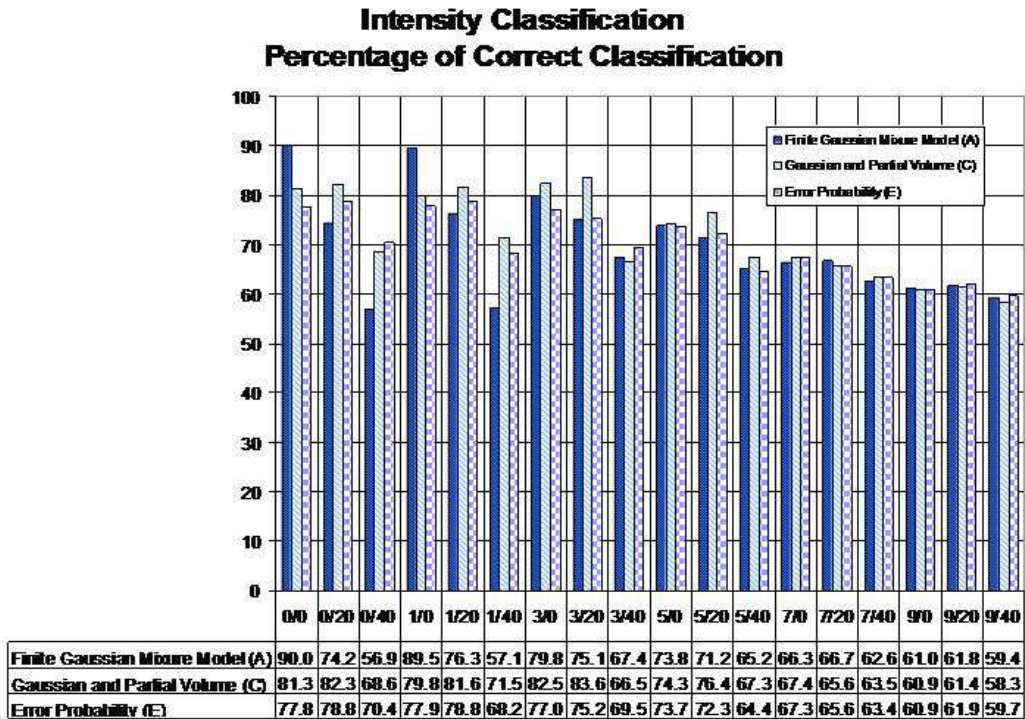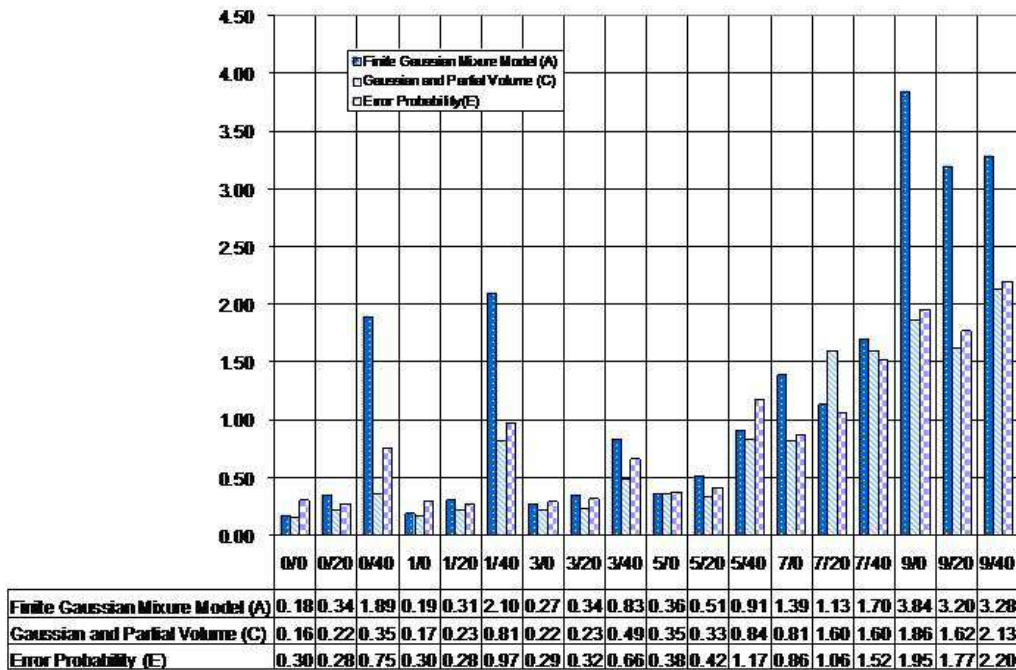|	0/0	0/20	0/40	1/0	1/20	1/40	3/0	3/20	3/40	5/0	5/20	5/40	7/0	7/20	7/40	9/0	9/20	9/40
FGMM and HMRF (B)	88.4	74.0	58.5	88.4	75.7	58.1	83.6	77.4	57.2	80.3	77.4	68.6	78.3	76.7	68.1	76.7	76.9	67.7
GPV and HMRF (D)	83.5	75.2	60.3	87.2	76.2	59.7	83.2	77.4	58.2	81.2	77.3	69.8	80.0	77.7	69.7	76.4	75.5	68.4
Non-Parametric HMRF (F)	79.2	72.0	64.6	77.3	71.8	64.9	77.2	73.6	63.9	75.2	72.3	64.8	73.4	69.5	64.5	71.2	68.7	62.6

Fig. 10. Percentages of correct classified voxels.

Intensity-based Classification
Percentage of Completely False Classification

	0/0	0/20	0/40	1/0	1/20	1/40	3/0	3/20	3/40	5/0	5/20	5/40	7/0	7/20	7/40	9/0	9/20	9/40
Finite Gaussian Mixure Model (A)	0.18	0.34	1.89	0.19	0.31	2.10	0.27	0.34	0.83	0.36	0.51	0.91	1.39	1.13	1.70	3.84	3.20	3.28
Gaussian and Partial Volume (C)	0.16	0.22	0.35	0.17	0.23	0.81	0.22	0.23	0.49	0.35	0.33	0.84	0.81	1.60	1.60	1.86	1.62	2.13
Error Probability (E)	0.30	0.28	0.75	0.30	0.28	0.97	0.29	0.32	0.66	0.38	0.42	1.17	0.86	1.06	1.52	1.95	1.77	2.20

HMRF Classification
Percentage of Completely False Classification

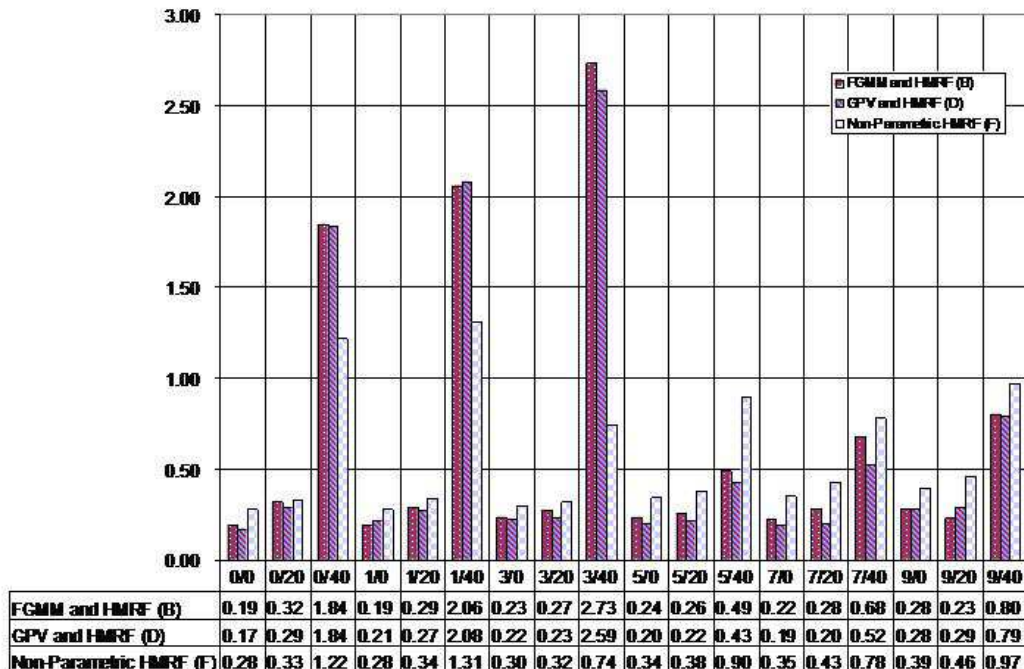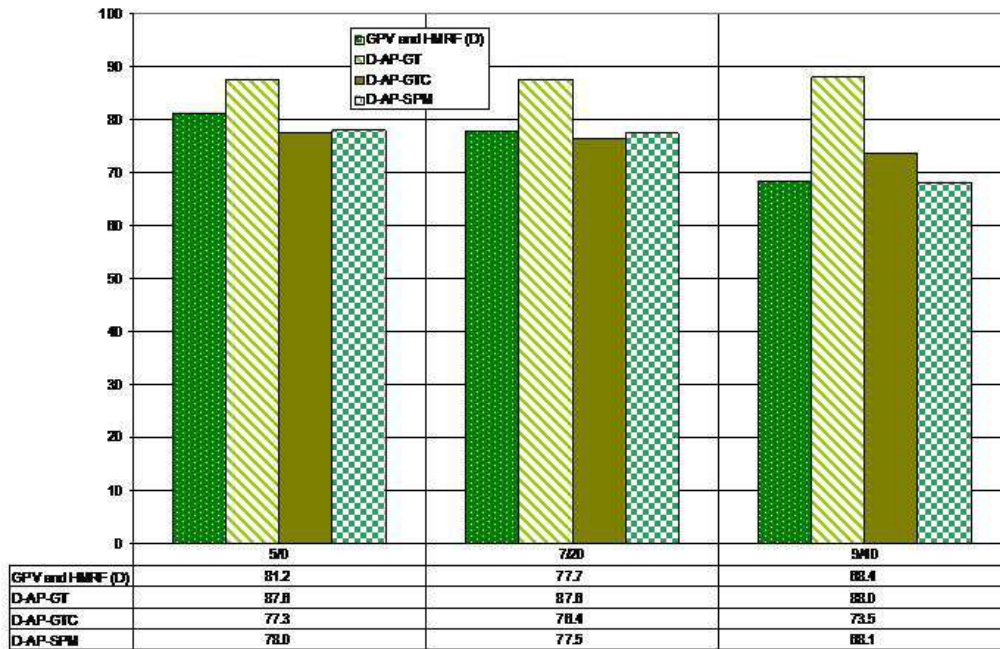|	0/0	0/20	0/40	1/0	1/20	1/40	3/0	3/20	3/40	5/0	5/20	5/40	7/0	7/20	7/40	9/0	9/20	9/40
FGMM and HMRF (B)	0.19	0.32	1.84	0.19	0.29	2.06	0.23	0.27	2.73	0.24	0.26	0.49	0.22	0.28	0.68	0.28	0.23	0.80
GPV and HMRF (D)	0.17	0.29	1.84	0.21	0.27	2.08	0.22	0.23	2.59	0.20	0.22	0.43	0.19	0.20	0.52	0.28	0.29	0.79
Non-Parametric HMRF (F)	0.28	0.33	1.22	0.28	0.34	1.31	0.30	0.32	0.74	0.34	0.38	0.90	0.35	0.43	0.78	0.39	0.46	0.97

Fig. 11. Percentages of completely false classified voxels.

Atlas Classification
Percentage of Correct Classification

	5/0	7/20	9/40
GPV and HMRF (D)	81.2	77.7	68.4
D-AP-GT	87.8	87.8	88.0
D-AP-GTC	77.3	78.4	73.5
D-AP-SPM	78.0	77.5	68.1

Atlas Classification
Percentage of False Classification

	5/0	7/20	9/40
GPV and HMRF (D)	0.20	0.20	0.79
D-AP-GT	0.17	0.12	0.21
D-AP-GTC	0.33	0.32	0.56
D-AP-SPM	0.19	0.19	0.57
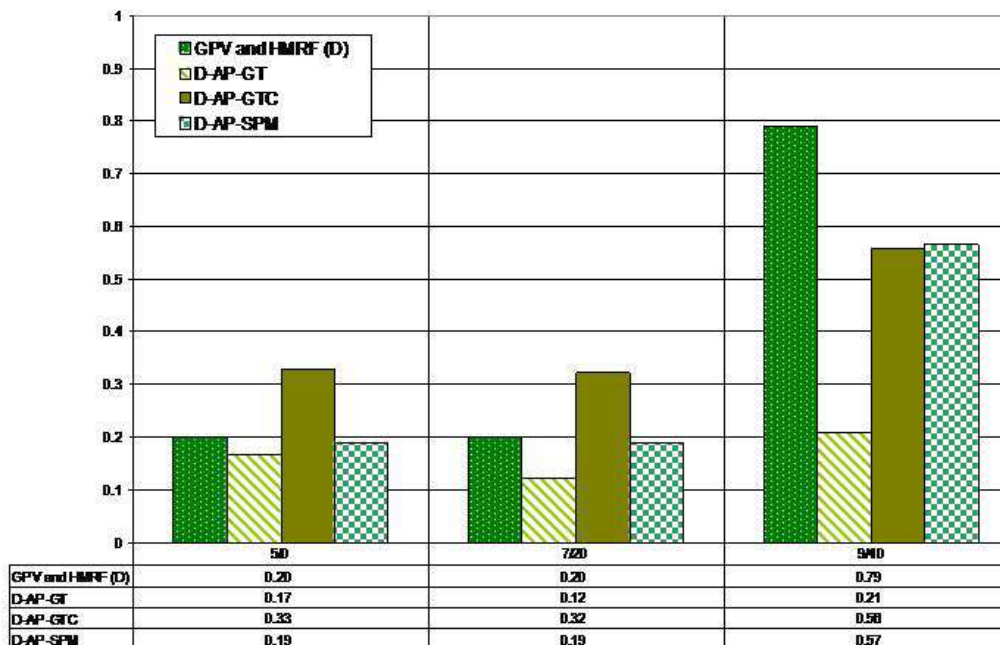
Fig. 12. Percentages of classification of methods using atlas information.

	PerGood	PerFault	PerHalf+	PerHalf-
FGMM	66.67	1.13	13.91	18.29
GHMRF	76.68	0.28	12.65	**10.45**
GPV	65.58	1.6	11.3	21.53
GPV-HMRF	**77.7**	**0.20**	**9.42**	12.69
EP	65.61	1.06	22.63	10.7
NP-HMRF	69.51	0.43	9.81	20.26
GT	**87.65**	**0.12**	**3.81**	**8.42**
GTC	76.44	0.32	10.13	13.10
SPM	77.45	0.19	9.93	12.43

TABLE III

Global performance of all methods for the phantom 7N20RF.

APPENDIX

I. NOTATION

- $\mathcal{S} = \{1, 2, ...N\}$ is the set of indexes of the image, and $\mathcal{N}_i$, $i \in \mathcal{S}$ is the set of sites neighboring the site i.
- $\boldsymbol{y} = \{y_1, y_2, ..., y_N\} \in \mathcal{D}^N \subset \mathbb{R}^{\mathbb{N}}$ is a configuration of Y, where $\mathcal{D} \subset \mathbb{R}$ is the state space representing the intensity.
- $\mathcal{Y} = \{\boldsymbol{y}|y_i \in \mathcal{D}, \rangle \in \mathcal{S}\}$ is the space of possible configurations. $Y = \{Y_i, i \in \mathcal{S}\}$ is the family of random variables.
- State space $\mathcal{L} = \{csf, cg, gm, gw, wm\}$ of all brain tissues, and state space $\mathcal{L}_p = \{csf, gm, wm\}$ of pure tissues.
- $X = \{X_i, i \in \mathcal{S}\}$ is the family of random variables representing the underlying class labels indexed by $\mathcal{S}$.
- $\boldsymbol{x} = \{x_1, x_2, ..., x_N\} \in \mathcal{L}^N$ denotes a configuration of X. $\mathcal{X} = \{\boldsymbol{x}|x_i \in \mathcal{L}, \rangle \in \mathcal{S}\}$ is the space of possible configurations.
- $\theta = \{\mu, \sigma\}$ are the parameters, mean and variance, defining a Gaussian distribution that is denoted by $N(\mu, \sigma)$.
- $P(x)$ is the probability of x, $P(y|x)$ is the conditional probability, and $P(y, x)$ is the joint probability.
- **k** indexes the iterations in time.
- $U(\boldsymbol{x}, \beta)$ is the energy function of a Gibbs distribution and β is called *spatial* factor.
- Z is the normalization factor of the Gibbs distribution.
- $P_A(\boldsymbol{x})$ is the anatomical prior probability map.

II. MAP FOR GHMRF

The objective is to assign a tissue type label $x \in \mathcal{L}$ to each voxel in the image. A labelling of S is denoted by $\mathbf{x}$ where $x_i, i \in S$ is the corresponding class label of voxel i. The true but unknown labelling configuration is denoted by $\mathbf{x}^*$, which is a particular realization of a random field X, which is an MRF with a specified distribution $P(\mathbf{x})$. The observable image itself is denoted by $\mathbf{y}$, which is a realization of a GHMRF as described in section II-B.5. According to the MAP criterion (see Eq. 33), we can define the problem as:

$$\hat{\mathbf{x}} = arg \max_{\mathbf{x} \in \mathcal{X}}\{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})\}. \tag{61}$$

The prior probability of the class and the likelihood probability of the observation need to be computed. As presented in Sec. II-B, since $\mathbf{x}$ is considered as a realization of an MRF, its prior probability can be derived from

$$P(\mathbf{x}) = \frac{1}{Z}Exp\left[-\beta U(\mathbf{x})\right]. \tag{62}$$

The voxel intensity y_i is assumed to follow a Gaussian distribution with parameters $\theta_x = \{\mu_x, \sigma_x\}$, given the tissue type label x_i:

$$p(y_i|x_i) = g(y_i; \theta_{x_i}) = \frac{1}{\sqrt{2\pi\sigma_{x_i}^2}}Exp\left[-\frac{(y_i - \mu_{x_i}^2)}{2\sigma_{x_i}^2}\right]. \tag{63}$$

Based on the conditional independence assumption of $\mathbf{y}$ (see Eq. 17), the joint likelihood probability takes the form of

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i \in S} p(y_i|x_i),$$

so,

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i \in S}\left(\frac{1}{\sqrt{2\pi}}Exp\left[-\frac{(y_i - \mu_l^2)}{2\sigma_l^2} - \log(\sigma_{x_i})\right]\right),$$

which can be written as

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z'} Exp\left[-U(\mathbf{y}|\mathbf{x})\right], \tag{64}$$

with the *likelihood energy*

$$U(\mathbf{y}|\mathbf{x}) = \sum_{i \in S} U(y_i|x_i) = \sum_{i \in S} \left(\frac{(y_i - \mu_{x_i}^2)}{2\sigma_{x_i}^2} + \log(\sigma_{x_i}) \right), \tag{65}$$

and the constant normalization term $Z' = (2\pi)^{(\mathcal{N}/2)}$. It appears that

$$\log(P(\mathbf{x}|\mathbf{y}) \propto -U(\mathbf{x}|\mathbf{y}), \tag{66}$$

where

$$U(\mathbf{x}|\mathbf{y}) = U(\mathbf{y}|\mathbf{x}) + U(\mathbf{x}) + const \tag{67}$$

is the *posterior energy*. The MAP estimation is equivalent to minimizing the posterior energy function

$$\hat{\mathbf{x}} = arg \min_{x \in \mathcal{X}} \{U(\mathbf{y}|\mathbf{x}) + \beta U(\mathbf{x})\}. \tag{68}$$

This minimization problem is mathematically simple but computationally infeasible. However, optimal solutions can be computed using iterative minimization techniques such as *iterated conditional modes (ICM)* [9].