Comparison of Image Intensity, Local and Multi-Atlas Priors in Brain Tissue Classification

Wang, Liping; Labrosse, Frederic; Zwiggelaar, Reyer

Published in:
Medical Physics

DOI:
10.1002/mp.12511

Publication date:
2017

Citation for published version (APA):
Wang, L., Labrosse, F., & Zwiggelaar, R. (2017). Comparison of Image Intensity, Local and Multi-Atlas Priors in Brain Tissue Classification. *Medical Physics*, *44*(11), 5782-5794. https://doi.org/10.1002/mp.12511

Comparison of Image Intensity, Local and Multi-Atlas Priors in Brain Tissue Classification

Liping Wang [a)], Frédéric Labrosse, and Reyer Zwiggelaar
Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB, UK

Purpose: Automated and accurate tissue classification in 3D brain Magnetic Resonance images is essential in volumetric morphometry or as a preprocessing step for diagnosing brain diseases. However, noise, intensity inhomogeneity and partial volume effects limit the classification accuracy of existing methods. This paper provides a comparative study on the contributions of three commonly used image information priors for tissue classification in normal brains: image intensity, local and multi-atlas priors.
Methods: We compared the effectiveness of the three priors by comparing the four methods modelling them: K-Means (KM), KM combined with a Markov Random Field (KM-MRF), multi-atlas segmentation (MAS) and the combination of KM, MRF and MAS (KM-MRF-MAS). The key parameters and factors in each of the four methods are analysed and the performance of all the models is compared quantitatively and qualitatively on both simulated and real data.
Results: The KM-MRF-MAS model that combines the three image information priors performs best.
Conclusions: The image intensity prior is insufficient to generate reasonable results for a few images. Introducing local and multi-atlas priors results in improved brain tissue classification. This study provides a general guide on what image information priors can be used for effective brain tissue classification.

Keywords: Classification, MRI, MRF, multi-atlas, priors

1. Introduction

Brain tissue classification aims to segment brain tissues, in our case the three primary tissue types: white matter (WM), grey matter (GM) and cerebral spinal fluid (CSF). Such segmentation is essential for diagnosing some brain disorders such as Alzheimer or Schizophrenia by quantitative analysis[1,2]. It assists with some applications in medical image analysis like image registration[3], lesion segmentation[4] and cortical surface extraction[5]. T1-weighted magnetic resonance (MR) imaging has been widely used in this area owing to its excellent soft tissue contrast in images[6,7]. Manual delineation of brain tissue is time-consuming due to the large volume of data. Also, manual delineation can lead to intra- and inter-expert variability[8]. Numerous supervised and unsupervised segmentation/classification algorithms have been developed in the last decade[9–11]. However, accurate and robust tissue classification remains challenging due to noise, intensity inhomogeneity and partial volume effects existing in brain MR images[6,12,13].

Different image information priors can be adopted to drive the tissue classification process. Using the information at a voxel level is the most intuitive approach. For example, intensity information is widely used in clustering methods such as K-Means (KM)[14], Fuzzy C-Means (FCM)[15] and Gaussian Mixture Model (GMM)[16]. Other features can also be extracted from images and used in classification methods that use machine learning algorithms[17–20].

Considering the continuity of each tissue type in brain images, the central voxel and its neighbours in a local neighbourhood tend to belong to the same tissue class. This spatial information can be modelled by a Markov Random Field model (MRF) and combined in the classification framework as a constraint to improve the accuracy, especially in cases where severe image noise and intensity inhomogeneity exist[6,7,21,22].

In addition to the prior information obtained from a single target image, population-specific atlases can also be introduced as a prior to benefit the classification. This process requires the registration between the atlases and the target image. In the simplest case, the classification is achieved by propagating the label map of a topological atlas to the target image[23]. To account for the inter-subject variability, a probabilistic atlas can be generated by averaging a series of label maps of other images and these probabilities are then utilised in the classification framework[7,24,25]. Recently, multi-atlas segmentation (MAS) has drawn attention because of its superior performance[23,25–28]. Instead of using a probabilistic atlas constructed from all the available label maps, a selection process can be applied to select the most relevant atlases, hence not biasing the classification by the relatively less relevant atlases. Atlas selection is carried out based on image similarity or meta-information[29]. After that, the selected atlases are combined using global or local label fusion methods for the final tissue classification[26,30–34]. Different from modelling MAS as a label prior, a statistical non-parametric regression framework was proposed[35,36] to model MAS in the high-dimensional space of images. The expected segmentation error of the regression estimator was characterised as a function of the size of the atlas database and optimised by estimating a set of parameters fundamental to the specific segmentation task. In addition, a globally optimal label fusion method was recently proposed[37], which combined MAS and graph-based segmentation. The shape priors were modelled by the graph-based method and incorporated into the label fusion approach to achieve a globally optimal segmentation.

The image intensity, local and multi-atlas priors were first combined into an energy functional, which was minimised by graph cuts, to segment the hippocampus in brain images[38]. It was then improved by adopting a modified graph cuts and Expectation Maximisation (EM) for brain structures segmentation[27]. This work was further extended[39] with the prior knowledge of neighbouring structures incorporated by a global and stationary MRF. The extension led to improved

performance on brain structures segmentation.

The contribution of this paper lies in providing a direct and unbiased comparison of the effects of the three image information priors described above. To achieve this, the performance of four methods modelling them is compared: the image intensity is utilised by KM to obtain a preliminary classification; the neighbouring information is modelled by a MRF and combined with KM by modelling each resulted tissue cluster with a Gaussian distribution; the multi-atlas prior is used in MAS by applying multi-atlas registration and global/local label fusion techniques; finally, we combined the multi-atlas prior with the image intensity and local prior into an overall KM-MRF-MAS framework. These four methods are chosen to ensure that each prior is modelled exactly in the same way in the involved methods so that the comparison of these three priors is not biased by using different modelling approaches. In MAS, the importance of the atlas selection is validated and the classification accuracies of applying various label fusion schemes are compared. The effects of a range of parameters in each model are analysed in detail and the performance of all four approaches is compared quantitatively and qualitatively on both simulated and real data.

2. Methods

In this section, we formulate the modelling of local[6] and multi-atlas priors[34].

2.A. Intensity and Local Prior Model: KM-MRF

The voxels of a 3D brain MR image are indexed with $i \in \mathcal{S} = \{1, 2, ..., N\}$ where N is the number of voxels. Each voxel in $\mathcal{S}$ is associated with $y_i \in \mathbb{R}$, the intensity value of the i^{th} voxel. The set of y_i is the observed image denoted by $y = \{y_1, y_2, ..., y_N\} \in \mathbb{R}^N$. Our tissue classification in normal brains aims to classify each voxel in $\mathcal{S}$ into one of the classes labelled by $\mathcal{L}$ = {CSF, GM, WM}. The tissue class of the i^{th} voxel is denoted by $x_i \in \mathcal{L}$ and $x = \{x_1, x_2, ..., x_N\} \in \mathcal{L}^N$ is a classification of the image. Our task is to find the best classification $x^* = \{x_1^*, x_2^*, ..., x_N^*\} \in \mathcal{L}^N$ given the image intensity y formulated as a maximum a posterior (MAP) problem:

$$\begin{aligned} x^* &= \arg\max_{x\in\mathcal{L}^N} P(x|y) = \arg\max_{x\in\mathcal{L}^N} \frac{P(y|x)P(x)}{P(y)} \\ &= \arg\max_{x\in\mathcal{L}^N} P(y|x)P(x) \end{aligned} \tag{1}$$

where $P(x|y)$ is the probability of the classification x given the image intensity y. $P(y)$ is independent from the classification x and hence ignored in the optimisation. By taking the negative logarithm of equation (1), the MAP problem is converted to the minimisation of an energy functional:

$$\begin{aligned} x^* &= \arg\min_{x\in\mathcal{L}^N} (-lnP(y|x) - lnP(x)) \\ &= \arg\min_{x\in\mathcal{L}^N} (E_{in}(x) + E_{pr}(x)). \end{aligned} \tag{2}$$

E_{in} represents the intensity energy prior which models the intensity distributions of three tissue classes. It measures how well the current classification x explains the image y. E_{pr} represents the prior knowledge of the classification.

The intensity prior energy E_{in} can be formulated as

$$E_{in}(x) = -\sum_{i\in\mathcal{S}} \ln P(y_i|x_i) = \sum_{i\in\mathcal{S}} \left(\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} + \ln\sqrt{2\pi} + \ln\sigma_{x_i} \right) \tag{3}$$

where $P(y_i|x_i)$ is the probability density function of y_i given the tissue class x_i and the model parameters $\theta_{x_i} = \{\mu_{x_i}, \sigma_{x_i}\}$ are the mean and standard deviation of the Gaussian distribution for the tissue class x_i.

The local prior energy E_{pr} can be formulated as

$$E_{pr} = \sum_{i\in\mathcal{S}} \left(\frac{\beta}{2} \sum_{j\in\mathcal{N}_i} \frac{\delta(x_i, x_j)}{d(i,j)} \right) + \ln Z, \tag{4}$$

where Z is the normalisation factor; β is the spatial parameter; $d(i,j)$ measures the distance between the two voxels at $\mathcal{S}_i$ and $\mathcal{S}_j$; $\delta(x_i, x_j)$ is a weighting function and defined as

$$\delta(x_i, x_j) = \begin{cases} -1 & x_i = x_j \\ +1 & x_i \neq x_j. \end{cases} \tag{5}$$

Then equation (2) can be rewritten as

$$x^* = \arg\min_{x\in\mathcal{L}^N} \sum_{i\in\mathcal{S}} \left(\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} + \ln\sigma_{x_i} + \frac{\beta}{2} \sum_{j\in\mathcal{N}_i} \frac{\delta(x_i, x_j)}{d(i,j)} \right) \tag{6}$$

with the constants removed.

2.B. Multi-Atlas Segmentation: MAS

In MAS, only the atlas priors (i.e. the segmentations of other images) are used to drive the tissue classification. First, multiple atlas images are nonrigidly registered to the target image using demons registration[40,41]. Then, the similarities between the registered atlas images and the target are measured and only the atlases with high similarities are selected to contribute to the classification. Next, the label maps of the selected atlases are propagated to the target image. Finally, global or local label fusion methods are applied to the transformed label maps to infer the classification of the target image. A variety of label fusion methods have been proposed[26]. We use representative approaches to compare global and local label fusion techniques in MAS. Other methods such as STAPLE[42] or SIMPLE[43] could be used as alternatives.

Mutual information (MI) and normalised cross correlation are commonly used for atlas selection[26,29]. In this paper, we used mutual information[44,45] to calculate the similarity between the registered atlas and the target image. Four label fusion strategies were compared in this study: majority voting (MV), MI, the probabilistic patch-based model (PPBM) and MI-PPBM. MV and MI are global label fusion methods, which estimate the classification accuracy of each atlas globally and assign a single weight to each of the selected

atlases. MV is the most commonly used[26,46]. It treats every atlas equally and assigns to each voxel the class label that most atlases agree on. MI takes image intensities into account and assigns to each atlas a global weight. Local label fusion methods like PPBM[32,34] estimate the classification accuracy at each voxel in a local neighbourhood and assign weights accordingly[46]. The model PPBM accounts for the potential registration error by considering a local neighbourhood in the atlases by estimating the intensity likelihood and the label likelihood simultaneously. It is straightforward to assign a global weight to each atlas in the local label fusion strategies such as MI-PPBM, which combines MI with PPBM.

The multi-atlas prior can be combined with KM-MRF by reformulating the energy functional in equation (2) as

$$\begin{aligned} x^* &= \arg\min_{x\in\mathcal{L}^N}(-\ln P(y|x) - \ln P(x)) \\ &= \arg\min_{x\in\mathcal{L}^N}(-\ln P(y|x) - \ln(P_{Lpr}(x)P_{Apr}(x))) \\ &\propto \arg\min_{x\in\mathcal{L}^N}(E_{in}(x) + E_{Lpr}(x) + \gamma E_{Apr}(x)). \end{aligned} \quad (7)$$

The probability of the prior classification $P(x)$ is determined by both the local and multi-atlas prior probabilities which are represented as P_{Lpr} and P_{Apr}, respectively. Then the two prior probabilities are rewritten as energy terms $E_{Lpr}(x)$ and $E_{Apr}(x)$. The effect of the local prior is balanced with that of the image intensity by the spatial parameter β embodied in $E_{Lpr}(x)$. γ is introduced to balance the effect of the multi-atlas prior with that of the intensity and neighbouring information. The local prior $E_{Lpr}(x)$ is defined by equation (4) and the multi-atlas prior $E_{Apr}(x)$ based on each label fusion strategy, is defined in Appendix A. When combining the 3 energy terms for minimisation, all the constants can be removed.

The tissue classification is iterated with parameters estimation in an EM framework[6]. In each iteration, the energy functional defined in equation (6) or (7) is minimised with the Iterated Conditional Modes (ICM) algorithm[47].

3. Experiments and Results

3.A. Databases

BrainWeb is a simulated brain database[48]. We used 18 T1-weighted brain MR images with various levels of noise (0%, 1%, 3%, 5%, 7% and 9%) and intensity inhomogeneity (0%, 20% and 40%). The size of each volume is $181 \times 217 \times 181$ and the image resolution is $1mm \times 1mm \times 1mm$. The segmentation of a normal anatomical model provided in the database was used to perform the skull stripping and generate the ground truth for the test data. The brain mask was obtained from combining the segments of WM, GM and CSF and then applied to all the simulated images. The ground truth of the three tissues was also produced from their fuzzy models by assigning each voxel the dominant tissue class.

The Internet Brain Segmentation Repository dataset IBSR18 and its manual segmentations were provided by the Center for Morphometric Analysis (CMA) at Massachusetts General Hospital[49]. It consists of 18 T1-weighted images with their tissue classifications. The size of each image is $256\times256\times128$ with the resolution varying between $0.8mm \times 0.8mm \times 1.5mm$ and $1mm \times 1mm \times 1.5mm$. These images have been 'positionally normalised' into the Talairach orientation and bias field corrected by the CMA 'autoseg' routines. The brain mask was generated by filling the holes of the combination of the manually segmented brain tissues. The filled voxels were then combined into the ground truth of CSF.

The IBSR project has provided a second database IBSR20, which consists of 20 normal T1-weighted brain images with their manual segmentations. The size of each volume is typically $256\times256\times61$ with the resolution $1mm\times1mm\times3mm$. Similar to IBSR18, the brain images were skull stripped and the new ground truth of CSF was generated.

MRBrainS13 is a clinical brain image database used in the Grand Challenge on MR Brain Image Segmentation workshop at the international conference on MICCAI in 2013[50]. Twenty 3T scans are available including 5 cases for training and 15 for testing. We only used T1-weighted scans with volume size $240 \times 240 \times 48$ and resolution $0.958mm \times 0.958mm \times 3mm$. All scans were bias corrected. For 5 training cases, manual segmentations of 8 brain structures and 3 main tissue types were completed using techniques based on the contour segmentation objects tool from Mevislab[51].

As mentioned above, for BrainWeb and IBSR databases, the ground truth of brain segmentation was used for all the images. We did not apply any other skull stripping step in order to make sure the tissue classification performance is not affected by the skull stripping methods. For MRBrainS13, because the ground truth for the test images were not provided, the MAS method was applied to remove the skulls from the whole-head images. The 5 training images were considered as atlases. The brain mask of each training image was generated from the label map of tissue segmentation. The voxels of 3 tissues (WM, GM and CSF) were considered as foreground and the other voxels were considered as background. By registering each training image to the test image, 5 candidate brain masks could be obtained for each test image. These candidate brain masks were combined by applying majority voting to generate the final brain mask for the test images. Image intensity normalisation was performed for all four databases in the process of atlas selection because it is necessary for image registration. Other preprocessing steps, such as intensity inhomogeneity or noise removal, were not performed.

3.B. Evaluation Measures

For each tissue class, we use the Dice similarity coefficient (DSC)[52] to measure the spatial overlap accuracy between the classification result and the ground truth, defined by

$$DSC = 2 \cdot TP/(2 \cdot TP + FP + FN) \quad (8)$$

where TP, FP and FN represent the true positives, false positives and false negatives. The overall accuracy for classifying the three tissues is defined by

$$AC = (TP_{WM} + TP_{GM} + TP_{CSF})/|S| \quad (9)$$

where $|S|$ denotes the number of voxels in the brain mask S. It calculates the proportion of all the correctly classified voxels

TABLE I. Optimised parameters from applying k-fold cross validation

Method		KM-MRF		MAS					KM-MRF-MAS
Priors		intensity and local prior		multi-atlas prior					intensity, local and multi-atlas priors
Parameter		β	$\lvert\mathcal{N}\rvert$	n_{atlas}	$\lvert B\rvert$	$\lvert\mathcal{M}\rvert$	σ_1	σ_2	γ
Database	BrainWeb	[1, 1.5]	6	-	-	-	-	-	-
	IBSR18	[100, $+\infty$)	10	5, 6	11	19	0.5	[1, 1.5]	[400, $+\infty$)
	IBSR20	[20, $+\infty$)	10	3, 5, 7	11	19	[20, $+\infty$)	10	[400, $+\infty$)
	MRBrainS13	0.3	6	3	7	7	0.1	2	0.3

- $\lvert\mathcal{N}\rvert$ represents the size of the local neighbourhood in KM-MRF; $\lvert B\rvert$ and $\lvert\mathcal{M}\rvert$ represent the patch size and the size of the local neighbourhood considered in MAS using PPBM for label fusion; σ_1 and σ_2 are also the parameters of PPBM.
- '[ ]' or '[ )' is used to represent the range of the parameter values. '-' means the method is not applied on the corresponding database.

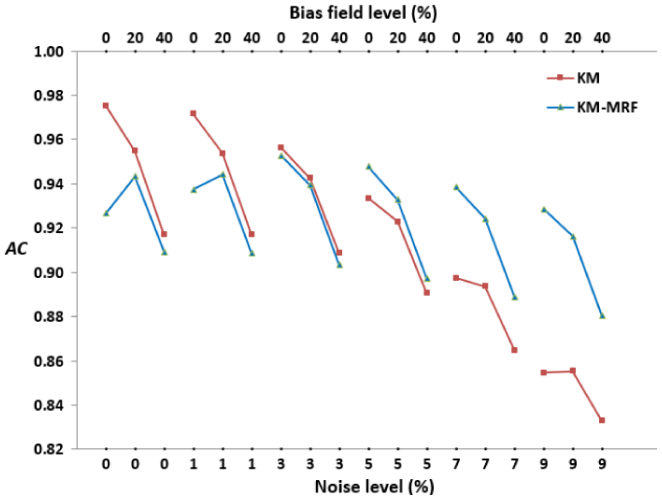


FIG. 1. Comparison of the AC (see equation (9)) values between applying KM-MRF and KM on the brain images with various levels of noise and bias field in BrainWeb.

within the brain mask. The values of both metrics are in the range [0, 1]. In our experiments, both metrics are used to measure the classification on BrainWeb and IBSR databases. For MRBrainS13, only the evaluation results measured by DSC are provided by the challenge.

In our experiments, a paired-sample t-test[53] is adopted to analyse the impacts of the parameter settings on the performance of specific methods. It measures the difference between two sets of classification results with the significance level set at 0.05. The two sets of results are produced by the same method with different parameters and measured by the same metric on the same database.

When comparing the performance of different methods with regard to several metrics on the same database, repeated-measures MANOVA (Multivariate Analysis of Variance)[54] is applied. The multivariate analysis tells if significant differences exist among the performance of different methods with respect to the combination of all metrics. If significant differences exist, the univariate analyses are undertaken to explore whether the significant differences exist among the methods in terms of each metric. Then for each metric where significant differences exist, the Bonferroni *post hoc* test is utilised to conduct all the pairwise comparisons to determine which pairs of the results are significantly different with the significance level set at 0.05. Finally, based on all the pairwise comparison results with respect to each metric, the performance of different methods is compared taking all metrics into account.

3.C. Parameters Optimisation

For BrainWeb and IBSR databases, we performed k-fold cross validation to optimise the parameters for each method. The choice of k depends on the number of brain volumes in the database. We used 3-fold cross validation for BrainWeb and IBSR18 (both contain 18 volumes so 6 per fold), while 4-fold cross validation was used for IBSR20 (containing 20 volumes so 5 per fold). For MRBrainS13, the 5 training images were used to optimise the parameters involved in all methods.

For KM, we set the number of clusters K as 3 because we aim to segment the brain tissue into three classes. For KM-MRF, a range of values were tested for the parameters β and the size of the local neighbourhood $\lvert\mathcal{N}\rvert$. At first, we set an initial value for β and evaluated the performance of KM-MRF when varying $\lvert\mathcal{N}\rvert$ between 6 and 26. After applying k-fold cross validation, the optimal $\lvert\mathcal{N}\rvert$ was obtained. Then we set $\lvert\mathcal{N}\rvert$ to the optimal value and tuned β in a range around its initial value, for which we again use k-fold cross validation. Finally, for each database, the brain tissue classification results were produced from applying KM-MRF with the corresponding optimised parameters. For MAS, we optimised the number of atlases fused n_{atlas}. In addition, 4 parameters involved in PPBM and MI-PPBM, the size of the local neighbourhood $\lvert\mathcal{M}\rvert$, the patch size $\lvert B\rvert$, σ_1 and σ_2, were also optimised. A similar approach was used as for KM-MRF: when optimising one parameter, the others were set to the initial or optimised values. A range of values were tested in optimising each parameter. For KM-MRF-MAS, we used the parameters optimised for KM-MRF and MAS. The additional parameter γ was optimised by evaluating the performance of KM-MRF-MAS with varying γ around its initial value.

In the optimisation process, different initial values for β

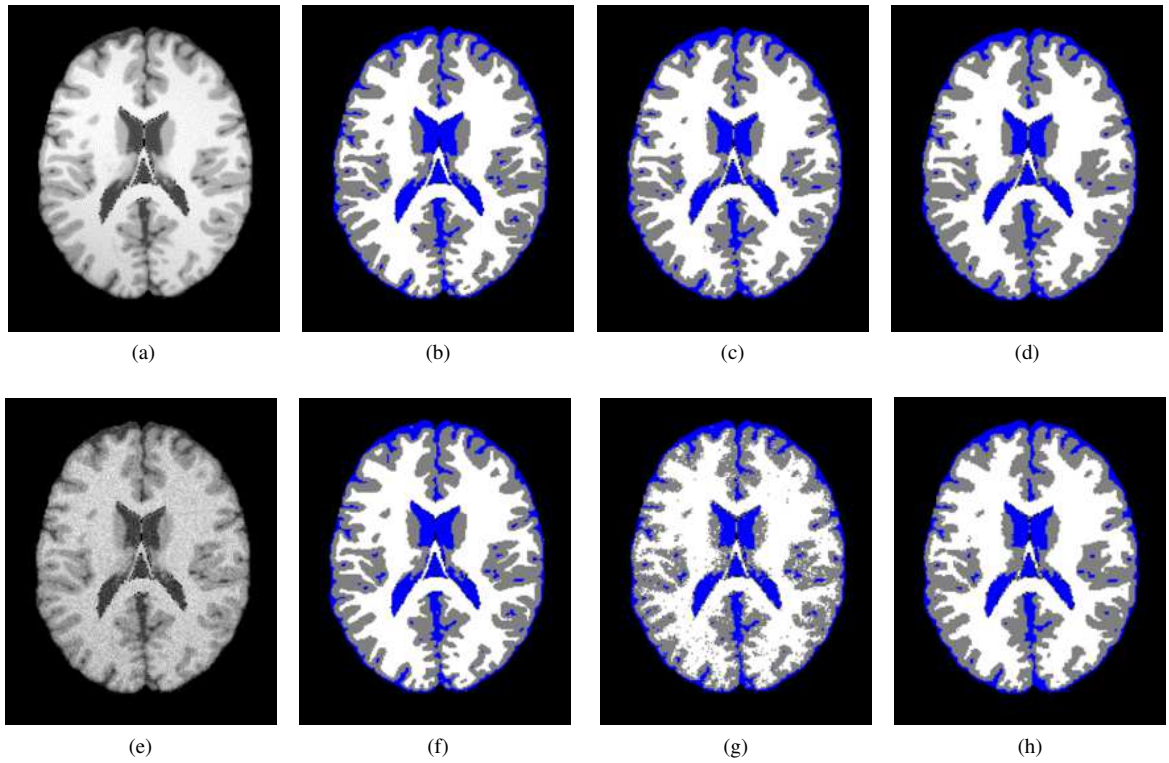


FIG. 2. Axial slices of two volumes (a) with 1% noise and 40% bias field, (e) with 7% noise and 20% bias field; their ground truth (b) and (f); the classification results of applying KM ((c) and (g)) and KM-MRF ((d) and (h)). Three tissues (CSF, GM and WM) in the ground truth and classifications are labelled in blue, grey and white, respectively. The overall accuracy AC and the accuracies of classifying three tissues DSC_{CSF}, DSC_{GM} and DSC_{WM} measured for each classified volume are: 0.92, 0.94, 0.91 and 0.91 for (c); 0.91, 0.93, 0.90 and 0.90 for (d); 0.89, 0.92, 0.88 and 0.89 for (g); 0.93, 0.93, 0.92 and 0.93 for (h).

(e.g. 1, 10, 100) were tested and we chose the range which produces the best segmentation results as the optimal parameter. The validation was repeated three times in optimising each parameter. It was found that the standard deviations among the segmentation results generated from repeating the optimisation process are very small.

In Table I, we list the optimal values for all the parameters involved in all the methods applying on various databases.

3.D. Validation of KM-MRF

3.D.1. Experiments on simulated data

Firstly, we tested the KM-MRF model described in Section 2.A on the 18 simulated BrainWeb images. The spatial parameter β was set in the range [1, 1.5] and 6 nearest neighbours were taken into account at each voxel. For each brain image with specific levels of noise and intensity homogeneity, the classification accuracy AC was compared between applying KM-MRF and KM which uses the image intensity only. As illustrated in Fig. 1, when the noise level is lower than 5%, KM outperforms KM-MRF; especially when both the noise and the bias field levels are very low (e.g. the images with 0% noise 0% bias field and 1% noise 0% bias field), KM-MRF performs much worse than KM. However, when the noise level increases (5% or higher), the classification benefits greatly from the neighbouring information. For most images with a certain level of noise, both methods deteriorate from increasing the bias field level; for most images with a certain level of bias field, both methods also perform worse when adding more noise but the AC of applying KM declines more steeply than when using KM-MRF especially when the noise level reaches 5%.

Fig. 2 shows a qualitative comparison. For the image with low noise levels (a), KM slightly outperforms KM-MRF and we hardly observe any difference between their classification results based on a single slice ((c) and (d)). However, for higher noise levels (e), KM-MRF distinctly improves the performance with the local prior taken into account ((g) and (h)).

3.D.2. Experiments on real data

We also tested KM-MRF on three real databases. The values of the parameters β and $|\mathcal{N}|$ applied on each database are listed in Table I. β was set to 100 and 20 for IBSR18 and IBSR20, respectively. The impacts of $|\mathcal{N}|$ and β on the classification accuracy are analysed in Appendixes B and C, respectively.

The classification results are compared with those produced by KM. The repeated-measures MANOVA is applied to test

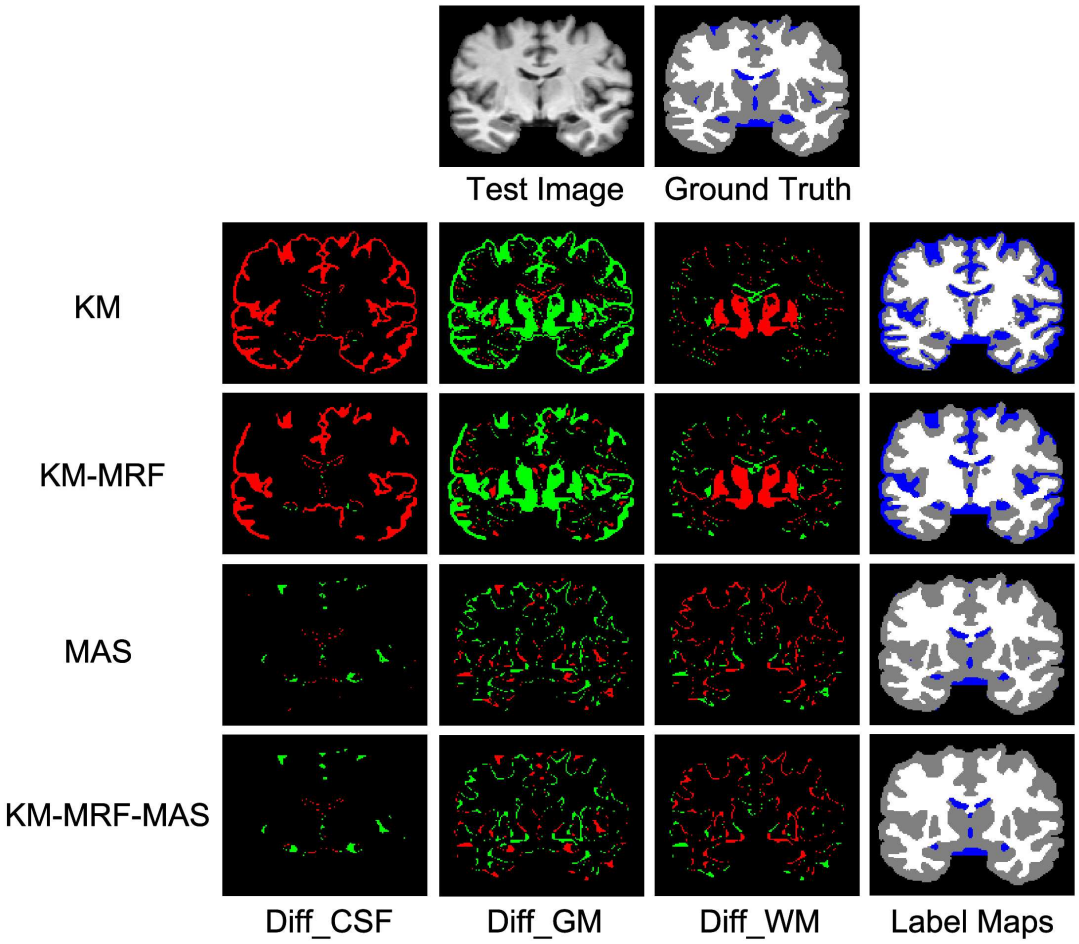


FIG. 3. Top row: coronal slice of a representative volume in IBSR18 and its ground truth of tissue classification. The difference maps of three tissues (Diff_CSF, Diff_GM and Diff_WM) and the label maps are produced by KM, KM-MRF, MAS and KM-MRF-MAS, respectively. Three tissues (CSF, GM and WM) in the ground truth and classifications are labelled in blue, grey and white colours, respectively. The red and green points in the difference maps represent the false positives and false negatives, respectively. The classification accuracies of the volumes produced by the four methods KM, KM-MRF, MAS and KM-MRF-MAS are: 0.71, 0.75, 0.92 and 0.92 measured by AC; 0.18, 0.24, 0.74 and 0.74 measured by DSC_{CSF}; 0.69, 0.75, 0.94 and 0.93 measured by DSC_{GM}; 0.90, 0.88, 0.92 and 0.91 measured by DSC_{WM}.

the differences between the performance of two methods and the results are listed in Table IV. The multivariate outcome indicates that significant differences exist between the performance of two methods with respect to the combination of all metrics. The univariate outcomes demonstrate that KM-MRF performs significantly better than KM in terms of the overall accuracy (AC) and the classification accuracies of CSF (DSC_{CSF}) and GM (DSC_{GM}) for both IBSR databases. Consistent results are obtained from the qualitative comparisons illustrated in Fig. 3 where the difference maps between the classification results generated by each method and the ground truth are displayed for each tissue class. We can see that the label map generated by KM-MRF is more contiguous than that produced by KM because the local prior encourages each tissue class to be more continuous; the false positives are decreased in CSF; a number of false negatives are eliminated in GM. The results in Table IV also shows that KM-MRF performs worse (non-significantly on IBSR18 and significantly on IBSR20) than KM in the classification of WM; however, taking the classification of all three tissues into account, the overall classification accuracy of KM-MRF is significantly higher than KM. For MRBrainS13, KM-MRF performs significantly better than KM in terms of the classification accuracies of all three tissue types. Based on these comparisons, the conclusion is drawn that the classification is significantly improved by taking the local prior into account on real data.

TABLE II. Analysis of the impact of atlas selection

Database	Approach	n_{atlas}	AC	p-value
IBSR18	Ranked	5, 6	0.87 ± 0.06	< 0.001
	Random	6, 8, 11	0.79 ± 0.08	
IBSR20	Ranked	3, 5, 7	0.78 ± 0.05	0.13
	Random	18, 19	0.77 ± 0.08	

- n_{atlas} represents the optimal number of atlases; MV is used for label fusion.

TABLE III. Comparisons of the performance of MAS using different label fusion strategies

Database	Method	AC	DSC_{CSF}	DSC_{GM}	DSC_{WM}
IBSR18	MV	0.87 ± 0.06	0.56 ± 0.13	0.90 ± 0.04	0.83 ± 0.14
	MI*	$0.87 \pm 0.06^{\star}$	$0.58 \pm 0.12^{\star}$	$0.90 \pm 0.04^{\star}$	0.83 ± 0.14
	PPBM**	$0.88 \pm 0.06^{\star\star}$	$0.70 \pm 0.13^{\star\star}$	$0.91 \pm 0.03^{\star\star}$	0.83 ± 0.14
	MI-PPBM**	$0.89 \pm 0.06^{\star\star}$	$0.70 \pm 0.13^{\star\star}$	$0.92 \pm 0.03^{\star\star}$	0.83 ± 0.14
IBSR20	MV	0.78 ± 0.05	0.39 ± 0.08	0.83 ± 0.04	0.76 ± 0.06
	MI*	$0.79 \pm 0.04^{\star}$	$0.41 \pm 0.08^{\star}$	$0.83 \pm 0.03^{\star}$	0.76 ± 0.06
	PPBM**	$0.82 \pm 0.04^{\star\star}$	$0.57 \pm 0.09^{\star\star}$	$0.86 \pm 0.03^{\star\star}$	$0.76 \pm 0.06^{\star}$
	MI-PPBM**	$0.82 \pm 0.04^{\star\star}$	$0.57 \pm 0.09^{\star\star}$	$0.86 \pm 0.03^{\star\star}$	$0.76 \pm 0.05^{\star}$

- Each value represents the mean accuracy and the standard deviation of applying MAS based on each label fusion strategy on the database.
- The $\star$s at the top right of each value indicate the significant differences of the corresponding result compared with all the others derived by applying other label fusion methods on the same database and measured by the same metric. The $\star$s at the top right of each method (in the methods column) indicate the significant differences of the performance of the corresponding method compared with all the other methods. The results or methods labelled with more $\star$s are significantly better than those with fewer or no $\star$s and the results or methods labelled with equal number of $\star$s have no significant difference.

3.E. MAS Based on Various Label Fusion Strategies

The performance of MAS based on various label fusion strategies was validated on the real data. For every brain image, the other images in the same database together with their label maps were considered as the atlases. All the atlas images were nonrigidly registered to the target image. Subsequently the similarity between each pair of the registered image and the target was calculated based on the mutual information and the atlases were ranked according to the similarities. The top ranked atlases were selected and their label maps were propagated to the target image to perform the tissue classification. The validation was not conducted on BrainWeb since all the simulated images in this database have the same ground truth.

3.E.1. Impact of the atlas selection

To investigate the necessity of the atlas selection, we tested MAS by increasing the number of ranked and randomly selected atlases on both IBSR18 and IBSR20. MV is used for multi-atlas label fusion due to its simplicity. The classification accuracy is measured by AC and averaged over the database. The paired-sample t-test is used to test the difference between any two sets of classification results produced by methods using the same number of ranked and random atlases.

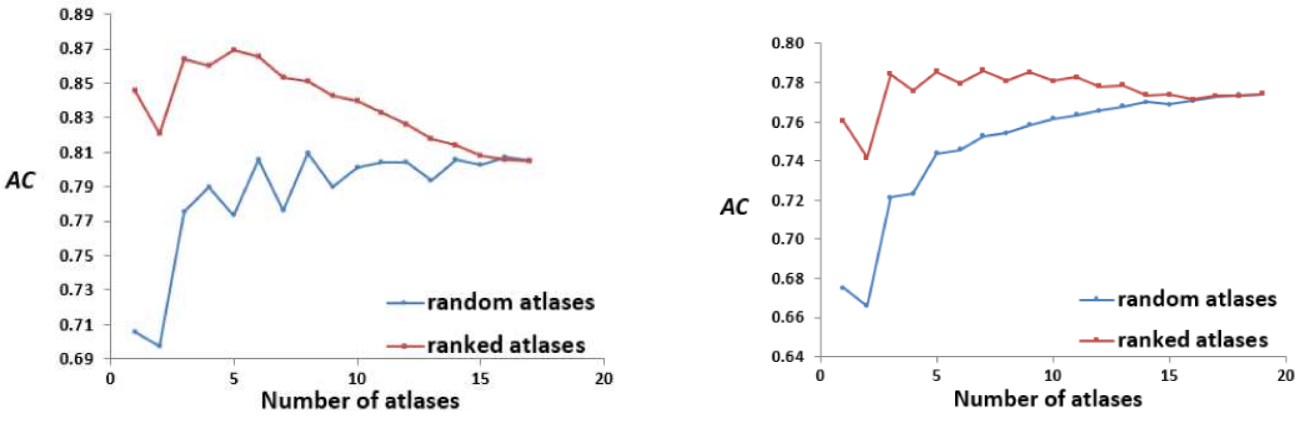
From Fig. 4, it is observed that at first MAS using the ranked atlases significantly outperforms that using the same number of random atlases on both databases. When increasing the number of atlases fused, the difference between the AC values obtained from fusing the ranked and random atlases decreases until it becomes nonsignificant (when fusing 14 atlases or more) and finally vanishes (when all the atlases are fused). Thus it is concluded that the atlases contribute unequally to the classification of brain images and the atlas selection approach picks out the most effective ones.

Then we run k-fold cross validation to optimise the number of ranked and random atlases fused in MAS on each database. In Table II, we listed the comparisons of the two approaches applied on the two databases. The tissue classification performance was measured by AC, which is the mean value across the whole database in three validations. The paired-sample t-test was used to test the difference between two sets of AC values generated from two approaches. It is shown that for IBSR18, applying atlas selection not only decreases the number of atlases fused but also significantly improves the segmentation performance. For IBSR20, no significant difference is found between the two sets of classification results measured by AC. However, the number of atlases fused is substantially reduced by applying atlas selection.

Therefore from a performance point of view, applying atlas selection significantly improves the classification accuracy or decreases the number of atlases to be fused without deteriorating the result. Reducing the number of atlases involved in MAS reduces computation for label fusion; nevertheless, the amount of registration required for atlas selection still makes it time-consuming. This is the main weakness of adopting the atlas selection in MAS. In addition to optimising the number of atlases fused in MAS by applying cross validation, an alternative approach was proposed, which models MAS as a non-parametric regression problem[35,36] and predicts the number of atlases required to keep the segmentation error below a specified tolerance level.

3.E.2. Impact of the label fusion strategy

Four label fusion strategies described in Section 2.B were considered for MAS and the classification results were compared on both IBSR18 and IBSR20. Five and three top ranked atlases were fused in MAS applied on IBSR18 and IBSR20, respectively. The other parameter settings of PPBM on IBSR18 and IBSR20 are listed in Table I. σ_1 was set to 20 when applying PPBM on IBSR20; σ_2 was set to 1 when applying PPBM on IBSR18. The parameters of MI-PPBM were the same as those of PPBM on each database.



(a) (b)

FIG. 4. AC values of applying MAS by increasing the number of ranked and random atlases on IBSR18 (a) and IBSR20 (b). MV is used for multi-atlas label fusion.

The repeated-measures MANOVA is applied to test the performance differences of MAS using four label fusion strategies and the results are listed in Table III. Significant differences are detected among the four methods with respect to the combination of four metrics on both databases. For IBSR18, the performance of four strategies is significantly different in terms of the overall accuracy (AC) and the classification accuracies of CSF (DSC_{CSF}) and GM (DSC_{GM}). The *post hoc* pairwise comparison results of the four methods are consistent with respect to these three metrics: MI significantly outperforms MV; PPBM and MI-PPBM perform equally and they perform better than both MV and MI. No significant differences exist among the four strategies in the classification of WM. For IBSR20, significant differences exist among the performance of four methods in terms of each metric. The pairwise comparison results with regard to AC, DSC_{CSF} and DSC_{GM} are the same with those for IBSR18. In the classification of WM, PPBM performs equally with MI-PPBM and they perform better than both MV and MI. Based on the comparison results on IBSR20, the performance of the four methods is ranked exactly the same as on IBSR18: MI outperforms MV; PPBM outperforms both MV and MI; MI-PPBM performs equally to PPBM.

We can draw the conclusion that in MAS, the mutual information based global weighting scheme MI performs better than MV; the probabilistic local label fusion model PPBM achieves better results than global models; finally the combination of global and local label fusion strategies MI-PPBM performs equally to PPBM.

3.F. Comparisons of Classification Models Using Different Image Information

Based on the conclusions we drew above, PPBM was adopted in the combined KM-MRF-MAS model. The impact of γ on the classification accuracy of KM-MRF-MAS is analysed in Appendix D. We compared the performance of the four tissue classification models on the three real databases. The optimal values were used for the parameters involved in each method applied on each database as listed in Table I. The repeated-measures MANOVA was applied to test the performance differences among the four models on the three databases and the results are listed in Table IV. The multivariate outcome tells that significant differences exist among the four methods with respect to the combination of the four metrics on all three databases. For IBSR18, the univariate outcomes demonstrate that significant differences exist among the four methods in terms of the overall accuracy (AC) and the classification accuracies of CSF (DSC_{CSF}) and GM (DSC_{GM}). The four models perform similarly in the classification of WM. The *post hoc* pairwise comparisons of the four models are conducted with respect to the metrics AC, DSC_{CSF} and DSC_{GM}. For IBSR20, significant differences exist among the four methods with respect to all four metrics. The results of *post hoc* pairwise comparisons of the four models regarding AC, DSC_{CSF} and DSC_{GM} are the same with those for IBSR18. For the classification accuracy of WM (DSC_{WM}), it shows that KM-MRF performs significantly worse than the other methods.

By comparing the performance of the three methods KM, KM-MRF and KM-MRF-MAS applied on IBSR databases, it is observed that introducing additional information including the local and multi-atlas priors significantly improves the tissue classification performance with respect to the overall accuracy and the classification accuracies of CSF and GM. Fig. 3 gives an example of the qualitative comparisons on a typical volume of IBSR18, which are consistent with the quantitative comparisons. Comparing the difference maps produced by KM, KM-MRF and KM-MRF-MAS, we observe that the total numbers of false positives and false negatives are dramatically decreased in the classification of CSF and GM by introducing additional information. From the label map generated by each method, we see that acceptable tissue classification is achieved by KM based on the image intensity prior, which provides an initialisation for KM-MRF; adding the local prior makes each tissue class more contiguous by encouraging the voxels to belong to the same class as their

TABLE IV. Comparisons of the performance of different tissue classification models on real data

Database	Method	AC	DSC_{CSF}	DSC_{GM}	DSC_{WM}
IBSR18	KM	0.68 ± 0.16	0.20 ± 0.08	0.70 ± 0.15	0.82 ± 0.21
	KM-MRF*	$0.72 \pm 0.18^{\star}$	$0.28 \pm 0.14^{\star}$	$0.75 \pm 0.18^{\star}$	0.80 ± 0.20
	MAS**	$0.88 \pm 0.06^{\star\star}$	$0.70 \pm 0.13^{\star\star}$	$0.91 \pm 0.03^{\star\star}$	0.83 ± 0.14
	KM-MRF-MAS***	$0.88 \pm 0.06^{\star\star}$	$0.71 \pm 0.12^{\star\star\star}$	$0.91 \pm 0.03^{\star\star}$	0.82 ± 0.15
IBSR20	KM	0.65 ± 0.16	0.21 ± 0.07	0.68 ± 0.15	$0.74 \pm 0.19^{\star}$
	KM-MRF*	$0.67 \pm 0.18^{\star}$	$0.34 \pm 0.15^{\star}$	$0.72 \pm 0.17^{\star}$	0.66 ± 0.17
	MAS**	$0.82 \pm 0.04^{\star\star}$	$0.57 \pm 0.09^{\star\star}$	$0.86 \pm 0.03^{\star\star}$	$0.76 \pm 0.06^{\star}$
	KM-MRF-MAS***	$0.82 \pm 0.04^{\star\star}$	$0.57 \pm 0.09^{\star\star\star}$	$0.86 \pm 0.03^{\star\star}$	$0.76 \pm 0.06^{\star}$
MRBrainS13	KM*	–	$0.75 \pm 0.04^{\star}$	$0.76 \pm 0.03^{\star}$	$0.84 \pm 0.03^{\star}$
	KM-MRF**	–	$0.76 \pm 0.04^{\star\star}$	$0.80 \pm 0.02^{\star\star}$	$0.87 \pm 0.02^{\star\star}$
	MAS	–	0.71 ± 0.06	0.70 ± 0.07	0.73 ± 0.08
	KM-MRF-MAS**	–	$0.77 \pm 0.04^{\star\star}$	$0.81 \pm 0.02^{\star\star}$	$0.87 \pm 0.02^{\star\star}$

- Each value represents the mean accuracy and the standard deviation of applying each method on the database.
- The meaning of the $\star$s at the top right of each value or method is the same with that in Table III.

neighbours; introducing the multi-atlas prior refines the misclassified structures. By combining all the information, KM-MRF-MAS produces the most accurate classification. Most misclassification is left at the intersection between two tissues which is probably caused by the misregistration and the partial volume effect (a single voxel contains more than one tissue). From Table IV, we also observe that in contrast with KM and KM-MRF, MAS achieves better or comparable results with regard to all the metrics. On the basis of the classification accuracy obtained by MAS, the additional intensity and local prior information do not significantly benefit the classification in terms of AC, DSC_{GM} and DSC_{WM}. However, the classification accuracy of CSF is significantly improved by combining all three priors. From Fig. 3, we see that on that specific slice the differences between MAS and KM-MRF-MAS are negligible. These results indicate that the multi-atlas prior contributes more than the image intensity and local prior in brain tissue classification. Finally, taking all the comparison results into account, the four classification models are ranked as: KM-MRF significantly outperforms KM; MAS performs better than both KM and KM-MRF; the combined model KM-MRF-MAS achieves the best performance.

For MRBrainS13, we do not include the evaluation results of AC since this metric is not used by the challenge. Significant differences exist among the 4 methods in terms of DSC_{CSF}, DSC_{GM} and DSC_{WM}. The results of *post hoc* pairwise comparisons with regard to each metric are consistent: KM-MRF performs significantly better than KM; MAS performs the worst among the four methods; KM-MRF-MAS produced segmentation results comparable to KM-MRF. The poor performance of MAS probably associates with the low number of available atlases, with the selected atlases not similar enough to the test image. due to this, the incorporation of the multi-atlas prior does not significantly improve the tissue classification performance. Thus we can draw the conclusion that introducing the local prior significantly improve the tissue classification; the efficiency of the multi-atlas prior highly depends on the number and relevance of the selected atlases.

TABLE V. The performance of related tissue classification methods on IBSR18

Method	DSC_{CSF}	DSC_{GM}	DSC_{WM}
KVL	0.21 ± 0.17	0.79 ± 0.05	0.86 ± 0.02
CGMM	0.26 ± 0.16	0.80 ± 0.06	0.86 ± 0.04
AMM	–	0.81 ± 0.04	0.89 ± 0.02

- KVL represents the Leemput's method[55]; CGMM represents the constrained GMM[16] and AMM is the adaptive Markov model[56].
- Each value represents the mean accuracy and the standard deviation of applying each method on IBSR18. For KVL and CGMM, the results on 15 images were reported[16] and for AMM, the accuracy of segmenting CSF was not reported[56].

4. Discussion

As described above, the image intensity prior can be utilised by applying simple clustering algorithms like KM for initial classification. However, the performance really depends on the quality of the images. When the noise and intensity inhomogeneity levels are low, the tissues are classified with high accuracy; when the images contain high levels of artefacts, the classification can severely deteriorate. Introducing the local prior significantly benefits the tissue classification in particular when severe noise and intensity inhomogeneity exist in the images.

For the methods incorporating the multi-atlas prior, the most relevant atlases are selected so that the classification is not biased by irrelevant atlases. In our experiments on the IBSR databases, the atlases were taken from the same database. We tried using atlases from both databases. However, the results showed that the best segmentation is obtained by using the atlases from the same database due to the higher similarity among images. In addition to atlas selection, a sufficient number of available atlases are crucial for the efficiency of the multi-atlas prior. The preprocessing such as bias field correction and spatial normalisation applied to the origi-

nal IBSR databases could contribute to the similarity between images. But the large number of available atlases greatly increase the chances of selecting the most effective ones for segmenting test images, which significantly benefits MAS. This has been confirmed by the evaluation on the MRBrainS13 database, where the effect of the multi-atlas prior deteriorates severely due to the low number of atlases provided.

Some methods in the literature are related to the approaches used in our work. For example, Leemput et al. proposed a model-based brain tissue classification method[55] which uses a digital atlas to initialise the tissue segmentation. The algorithm interleaves tissue classification, intensity distribution parameter estimation, intensity inhomogeneity correction and MRF parameter estimation by applying a generalised EM approach. A constrained GMM was developed by Ruf et al.[16]. In this approach, each tissue type is modelled by a large number of Gaussian components and the parameters of the constrained GMM are estimated by applying EM. Awate et al. proposed an adaptive nonparametric Markov model for brain tissue classification[56]. The initial tissue classification is achieved using a probabilistic atlas; a stationary Markov model is used to model the neighbourhood information; the Markov image statistics are estimated by applying the nonparametric Parzen-window technique and the segmentation is optimised by maximising mutual information. These three methods were tested on IBSR18 and their reported tissue classification performance measured by DSC are listed in Table V. It can be observed that these three methods outperform KM-MRF for classifying both GM and WM since they are more sophisticated compared to KM-MRF. MAS and KM-MRF-MAS significantly outperform all these three methods for the segmentation of CSF and GM. They perform worse for the segmentation of WM. The reasons might be that the intensity inhomogeneity correction was not adopted in the methods we used and the partial volume effect was not taken into account. Besides, not all images in the dataset were used to generate the reported results listed in Table V, which also makes the comparison unfair.

We we did not consider the validity of the ground truth provided by the IBSR databases since they have been widely used for the validation of brain image analysis techniques. It has been pointed out that the sulcal CSF voxels in the IBSR datasets are considered as GM, which could affect the tissue classification accuracies measured by DSC[9]. In spite of the uncertainty existing in the ground truth of CSF and GM, we do not think it affects the conclusions we have drawn. First, the improvement of the classification by introducing the local prior has also been validated on two other databases: BrainWeb (see Fig. 1) and MRBrainS13 (see Table IV), for which we assume the ground truth provided is accurate. Second, the deviation in the ground truth of the IBSR datasets should not affect the performance of the multi-atlas prior. As described in Section 2.B, in a multi-atlas based segmentation method, the annotations of the images are actually used to achieve the segmentation. If the annotation is corrected, the segmentation will be adjusted accordingly.

The usage of the multi-atlas prior results in superior performance, but multiple atlases are not always available since human annotation takes great effort. The large amount of registration required in the atlas selection also restricts its application. A compromise was presented[29] and employed[34], which performs the atlas selection after the affine registration and applies the nonrigid registration to the top ranked atlases only. Moreover, registration between images might fail or the test images could not be similar enough to the images of atlases, in which case outliers will be included. A keypoint transfer segmentation approach was proposed to segment abdominal organs in computerised tomography (CT) images, which propagates the label maps according to the transformation calculated from the matched keypoints in the atlas and test images[57]. This approach requires no registration and yields a segmentation accuracy which compares favourably to that of state-of-the-art methods.

In addition to using the annotated images as atlases and performing the tissue classification in a MAS approach, we can also extract more advanced features such as Gaussian scale-space features[18,58], Gaussian derivative features[18] or 3D Haar-like features[19] from all the images and the classification is then achieved by training a classifier, such as a k-nearest neighbour classifier[58], a support vector machine[18] or a random forest[19], using these features. Some work has been proposed to extract the features from the tissue probability maps derived from classification results[19,20] and these features together with those extracted from the images are used for training the subsequent classifiers in a multi-stage tissue classification framework. Hence the tissue probability maps are refined at each classification stage. Also, the classification model resulting from applying the classifiers as described above can be combined with local and multi-atlas priors investigated in our paper, to further improve the classification. It has been stated that the classification derived from applying the multi-stage random forest can be combined with the anatomically-constrained multi-atlas segmentation approach[59] to reduce the possible anatomical errors[19]. Conversely, the multi-atlas based method can be applied first and the trained classifiers are used to refine the classification at 'ambiguous' voxels[60]. Alternatively, an appearance model can be obtained from training classifiers and then combined with the spatial model and the interaction potential generated from the multi-atlas segmentation and the neighbouring information, respectively[58], for the final tissue classification.

In contrast to using hand-crafted features and shallow machine learning algorithms, deep neural networks have produced impressive classification results by extracting higher-order features using a hierarchical data representation. A 3D deep convolutional neural network was introduced[61], which offers an end-to-end learning-based approach for segmenting brain structures. It jointly learns the abstract feature representation and multi-class classification. The probability map produced by the network and a fully connected conditional random field were used to generate the final segmentation. For brain tissue classification, a parallel multi-dimensional Long Short-Term Memory (MD-LSTM) network was proposed[62], which takes each pixel's entire spatio-temporal context into account. A deep voxelwise residual network was trained to generate more representative features[63]. Multi-modality and

multi-level contextual information were integrated to train the network. The classification performance can be further improved by combining the low-level appearance features and high-level context. These works provide state-of-the-art accuracy on the MRBrainS13 database[50]. Unfortunately, the multi-atlas based methods (MAS and KM-MRF-MAS) discussed in this paper do not perform efficiently on this database because of the low number of atlases provided.

Moreover, images taken from other modalities such as diffusion or T2-weighted MR imaging for the same subject provide more information for brain tissue classification[8]. Longitudinal data can also facilitate tissue classification by taking temporal smoothness into account. The preprocessing steps, such as noise removal[64], intensity inhomogeneity correction[65] and partial volume effect correction[66] can also benefit brain tissue classification.

5. Conclusions

This paper has investigated the effects of three main image priors including intensity, local and multi-atlas priors on the tissue classification in 3D T1-weighted brain MR images. The modelling of these image priors, the combination of the models constructed using one or two image priors and the tissue classification approach of each individual or combined model have been described. The performance of all the models using the three image priors has been validated on the simulated and real data. We have also discussed the impacts of varying the key parameters or factors of each model thoroughly. The classification results have been evaluated based on several metrics and the performance of all the models has been compared quantitatively and qualitatively.

The image intensity used by KM generates acceptable initial classification for most images to form the basis for further processing. However, its performance deteriorates from severe artefacts and it produces a few outliers in real data. Introducing the local prior overcomes the disadvantage of KM to some extent and provides an incremental performance improvement. The multi-atlas prior is beneficial in refining the misclassified structures contributing more in brain tissue classification than the image intensity and the local prior. Appropriate atlas selection and sufficient number of available atlases are crucial in MAS based approaches. The local weighting method performs better than the global methods in multi-atlas label fusion. The model combining all 3 image priors achieves better or equal results compared to the other methods.

A. The calculation of multi-atlas prior energy

B. Impact of the size of the local neighbourhood on KM-MRF

C. Impact of β on KM-MRF

D. Impact of γ on KM-MRF-MAS

All the appendixes can be found in the supplemental material.

Acknowledgments

This work is supported by China Scholarship Council and the Department of Computer Science, Aberystwyth University. This support is gratefully acknowledged.

a) Author to whom correspondence should be addressed. Electronic mail:liw20@aber.ac.uk.

[1] P. Vemuri, J. L. Gunter, M. L. Senjem, et al. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *NeuroImage*, 39(3):1186–1197, 2008.

[2] S. M. Smith, Y. Zhang, M. Jenkinson, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479–489, 2002.

[3] R. A. Heckemann, S. Keihaninejad, P. Aljabar, D. Rueckert, J. V. Hajnal, A. Hammers, Alzheimer's Disease Neuroimaging Initiative. Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. *NeuroImage*, 51(1):221–227, 2010.

[4] B. Johnston, M. S. Atkins, B. Mackiewich, and M. Anderson. Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. *IEEE Trans Med Imaging*, 15(2):154–169, 1996.

[5] J. S. Kim, V. Singh, J. K. Lee, et al. Automated 3-d extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage*, 27(1):210–221, 2005.

[6] M. B. Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J. P. Thiran. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *IEEE Trans Med Imaging*, 24(12):1548–1565, 2005.

[7] J. Tohka, I. D. Dinov, D. W. Shattuck, and A. W. Toga. Brain MRI tissue classification based on local Markov random fields. *Magn Reson Imaging*, 28(4):557–573, 2010.

[8] H. A. Vrooman, C. A. Cocosco, F. van der Lijn, et al. Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. *NeuroImage*, 37(1):71–81, 2007.

[9] S. Valverde, A. Oliver, M. Cabezas, E. Roura, and X. Lladó. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. *J Magn Reson Imaging*, 41(1):93–101, 2015.

[10] E. I. Zacharaki, S. Kanterakis, R. N. Bryan, and C. Davatzikos. Measuring brain lesion progression with a supervised tissue classification system. In *MICCAI*, volume 5241 of *Lecture Notes in Computer Science*, pages 620–627. Springer, 2008.

[11] A. W. C. Liew and H. Yan. Current methods in the automatic tissue segmentation of 3D magnetic resonance brain images. *Curr Med Imaging Rev*, 2(1):91–103, 2006.

[12] J. C. Rajapakse and F. Kruggel. Segmentation of MR images with intensity inhomogeneities. *Image Vision Comput*, 16(3):165–180, 1998.

[13] S. Ruan, C. Jaggi, J. Xue, J. Fadili, and D. Bloyet. Brain tissue classification of magnetic resonance images using partial volume modeling. *IEEE Trans Med Imaging*, 19(12):1179–1187, 2000.

[14] G. N. Abras and V. L. Ballarín. A weighted k-means algorithm applied to brain tissue classification. *J Comput Sci Tech*, 5, 2005.

[15] D. Q. Zhang and S. C. Chen. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artif Intell Med*, 32(1):37–50, 2004.

[16] A. Ruf, H. Greenspan, and J. Goldberger. Tissue classification of noisy MR brain images using constrained GMM. In *MICCAI*, volume 3750 of *Lecture Notes in Computer Science*, pages 790–797. Springer, 2005.

[17] H. A. Vrooman, C. A. Cocosco, R. Stokking, et al. kNN-based multispectral MRI brain tissue classification: manual training versus automated atlas-based training. In *SPIE Medical Imaging*, pages 61443L–61443L. International Society for Optics and Photonics, 2006.

[18] A. van Opbroek, F. van der Lijn, and M. de Bruijne. Automated brain-tissue segmentation by multi-feature SVM classification. *MICCAI Grand Challenge on MR Brain Image Segmentation Workshop*, 2013.

[19] L. Wang, Y. Gao, F. Shi, et al. Links: Learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage*, 108:160–172, 2015.

[20]P. Moeskops, M. A. Viergever, M. J. Benders, and I. Išgum. Evaluation of an automatic brain segmentation method developed for neonates on adult MR brain images. In *SPIE Medical Imaging*, pages 941315–941315. International Society for Optics and Photonics, 2015.

[21]S. Yousefi, R. Azmi, and M. Zahedi. Brain tissue segmentation in MR images based on a hybrid of MRF and social algorithms. *Med Image Anal*, 16(4):840–848, 2012.

[22]B. Fischl, D. H. Salat, E. Busa, M. Albert, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[23]M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Comput Meth Prog Bio*, 104(3):e158–e177, 2011.

[24]Z. Yi, A. Criminisi, J. Shotton, and A. Blake. Discriminative, semantic segmentation of brain tissue in MR images. In MICCAI, volume 5762 of *Lecture Notes in Computer Science*, pages 558–565. Springer, 2009.

[25]K. O. Babalola, B. Patenaude, P. Aljabar, et al. Comparison and evaluation of segmentation techniques for subcortical structures in brain MRI. In *MICCAI*, volume 5241 of *Lecture Notes in Computer Science*, pages 409–416. Springer Berlin Heidelberg, 2008.

[26]J. E. Iglesias and M. R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Med Image Anal*, 24(1):205–219, 2015.

[27]J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–2365, 2010.

[28]M. Sdika. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. *Med Image Anal*, 14(2):219–226, 2010.

[29]P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738, 2009.

[30]A. J. Asman and B. A. Landman. Non-local statistical label fusion for multi-atlas segmentation. *Med Image Anal*, 17(2):194–208, 2013.

[31]H. Wang, J. W. Suh, J. Pluta, M. Altinay, and P. Yushkevich. Optimal weights for multi-atlas label fusion. In *IPMI*, volume 6801 of *Lecture Notes in Computer Science*, pages 73–84. Springer, 2011.

[32]M. R. Sabuncu, B. T. Thomas Yeo, K. Van Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *IEEE Trans Med Imaging*, 29(10):1714–1729, 2010.

[33]G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Med Image Anal*, 18(6):881–890, 2014.

[34]W. Bai, W. Shi, D. P. O'Regan, et al. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. *IEEE Trans Med Imaging*, 32(7):1302–1315, 2013.

[35]S.P. Awate, P. Zhu and R.T. Whitaker. How many templates does it take for a good segmentation?: error analysis in multiatlas segmentation as a function of database size. In *International Workshop on Multimodal Brain Image Analysis. MBIA 2012*, volume 7509 of *Lecture Notes in Computer Science*, pages 103–114. Springer, Berlin, Heidelberg, 2012.

[36]S.P. Awate, and R.T. Whitaker. Multiatlas segmentation as nonparametric regression. *IEEE Trans Med Imaging*, 33(9):1803–1817, 2014.

[37]I. Oguz, S. Kashyap, H. Wang, P. Yushkevich and M. Sonka. Globally Optimal Label Fusion with Shape Priors. In *MICCAI*, volume 9901 of *Lecture Notes in Computer Science*, pages 538–546. Springer, 2016.

[38]F. van der Lijn, T. den Heijer, M. M. Breteler, and W. J. Niessen. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43(4):708–720, 2008.

[39]C. Ledig, R. Wolz, P. Aljabar, et al. Multi-class brain segmentation using atlas propagation and EM-based refinement. In *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 896–899. IEEE, 2012.

[40]J. P. Thirion. Fast non-rigid matching of 3D medical images. *[Research Report] RR-2547*, page 37, 1995.

[41]H. Wang, L. Dong, J. O'Daniel, et al. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. *Phys Med Biol*, 50(12):2887, 2005.

[42]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 Imaging*, 23(7):903–921, 2004.

[43]T. R. Langerak, U. A. van der Heide, A. N. Kotte, M. A. Viergever, M. van Vulpen, and J. P. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Trans Med Imaging*, 29(12):2000–2008, 2010.

[44]C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[45]T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[46]X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE Trans Med Imaging*, 28(8):1266–1277, 2009.

[47]J. Besag. On the statistical analysis of dirty pictures. *J R Stat Soc B*, pages 259–302, 1986.

[48]BrainWeb: Simulated Brain Database. http://www.bic.mni.mcgill.ca/brainweb/, accessed October 21, 2013.

[49]NITRC: IBSR. http://www.nitrc.org/projects/ibsr/, accessed July 1, 2015.

[50]A. M. Mendrik, K. L. Vincken, H. J. Kuijf, et al. MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans. *Comput Intel Neurosc*, 2015.

[51]MeVisLab. http://www.mevislab.de/, accessed April 5, 2016.

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

[53]P. Armitage, G. Berry, and J. N. S. Matthews. *Statistical methods in medical research*. John Wiley & Sons, 2008.

[54]A. Mayers. *Introduction to statistics and SPSS in psychology*. Pearson, 2013.

[55]K. Van Leemput, F. Maes, D. Vandermeulen and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Trans Med Imaging*, 18(10):897–908, 1999.

[56]S.P. Awate, T. Tasdizen, N. Foster and R.T Whitaker. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. *Med Image Anal*, 10(5):726–739, 2006.

[57]C. Wachinger, M. Toews, G. Langs, W. Wells and P. Golland. Keypoint transfer segmentation. In *IPMI*, volume 9123 of *Lecture Notes in Computer Science*, pages 233–245. Springer, 2015.

[58]F. Van der Lijn, M. De Bruijne, S. Klein, et al. Automated brain structure segmentation based on atlas registration and appearance models. *IEEE Trans Med Imaging*, 31(2):276–286, 2012.

[59]L. Wang, F. Shi, Y. Gao, et al. Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation. *NeuroImage*, 89:152–164, 2014.

[60]H. Xiao. Learning-boosted label fusion for multi-atlas auto-segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 17–24. Springer International Publishing, 2013.

[61]C. Wachinger, M. Reuter and T. Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 2017.

[62]M.F. Stollenga, W. Byeon, M. Liwicki and J. Schmidhuber. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In *Advances in Neural Information Processing Systems 2015*, pages 2998–3006, 2015.

[63]H. Chen, Q. Dou, L. Yu, J. Qin and P.A. Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 2017.

[64]M.A. Ertürk, P.A. Bottomley and A.M.M. El-Sharkawy. Denoising MRI using spectral subtraction. *IEEE Trans Bio Med Eng*, 60(6):1556–1562, 2013.

[65]N.J. Tustison, B.B. Avants, P.A. Cook, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, 29(6):1310–1320, 2010.

[66]K. Van Leemput, F. Maes, D. Vandermeulen and P. Suetens. A unifying framework for partial volume segmentation of brain MR images. *IEEE Trans Med Imaging*, 22(1):105–119, 2003.