A Fully Automatic and Robust Brain MRI Tissue Classification Method

Chris A. Cocosco [*], Alex P. Zijdenbos, Alan C. Evans

McConnell Brain Imaging Centre, Montreal Neurological Institute,
McGill University, 3801 University Street, Montréal, Québec, H3A 2B4, Canada

Abstract

A novel, fully automatic, adaptive, robust procedure for brain tissue classification from 3D magnetic resonance head images (MRI) is described in this paper. The procedure is adaptive in that it customizes a training set, by using a "pruning" strategy, such that the classification is robust against anatomical variability and pathology. Starting from a set of samples generated from prior tissue probability maps (a "model") in a standard, brain-based coordinate system ("stereotaxic space"), the method first reduces the fraction of incorrectly labeled samples in this set by using a minimum spanning tree graph-theoretic approach. Then, the corrected set of samples is used by a supervised kNN classifier for classifying the entire 3D image. The classification procedure is robust against variability in the image quality through a non-parametric implementation: no assumptions are made about the tissue intensity distributions. The performance of this brain tissue classification procedure is demonstrated through quantitative and qualitative validation experiments on both simulated MRI data (10 subjects) and real MRI data (43 subjects). A significant improvement in output quality was observed on subjects who exhibit morphological deviations from the model due to aging and pathology.

Key words: Automatic brain MRI classification, Automatic brain tissue segmentation, Morphological variability insensitivity, Pruning, Non-parametric method

[*] Corresponding author. Address: Chris A. Cocosco, Philips Research Laboratories, Division Technical Systems, Roentgenstrasse 24-26, D-22335 Hamburg, Germany.
URL: `http://www.bic.mni.mcgill.ca/users/crisco/` (Chris A. Cocosco).

Preprint of article published in: Medical Image Analysis vol.7(4), p513–527, Dec 2003.

1 Introduction

Fully automatic brain tissue classification from magnetic resonance images (MRI) is of great importance for research and clinical studies of the normal and diseased human brain (e.g.: Collins et al., 2001; MacDonald et al., 2000; Paus et al., 1999; Rapoport et al., 1999; Zijdenbos et al., 2002). Operator-assisted classification methods are non-reproducible, and also are impractical for the large amounts of data required for a meaningful statistical analysis. Methods for fully automatic brain tissue classification typically rely on an existing anatomical model for localizing a training set for each tissue class to be labelled, e.g. gray matter, white matter, and CSF (cerebro-spinal fluid). This assumption of normal anatomical distribution of tissue types makes them sensitive to any deviations from the model due to pathology, or simply due to normal anatomical variability between individuals. Also, there may be situations when the only model available was constructed from a different human population than the image to be classified. This paper presents a novel, fully automatic classification procedure that is robust against morphological deviations from the model. Moreover, the procedure does not make any assumptions about the MRI tissue intensity distributions.

Many kinds of computerized analyses can be used to extract information from three-dimensional (3D) MRI data of the human head. The application that concerns this paper is the classification, or labeling, of individual voxels of a 3D anatomical MR image (MRI) as one of the three main tissue classes in the brain: CSF, grey matter, and white matter; a fourth class is defined as "background", denoting everything else (skull, skin, fat, air surrounding the subject's head, and so on). An attractive feature of MRI is that different contrasts between tissue types (multi-spectral image data of the same subject) can be easily obtained. Accurate and robust tissue classification is the basis for many applications such as the quantitative analysis of tissue volume in healthy and diseased populations (Collins et al., 2001; Rapoport et al., 1999; Zijdenbos et al., 1998, 2002), cortical thickness measurements (Fischl and Dale, 2000; Jones et al., 2000; MacDonald et al., 2000), morphological analysis (e.g. voxel-based morphometry: Paus et al., 1999; Wright et al., 1995), and visualization.

Several types of medical image segmentation methods can be applied to anatomical brain MRI. Intensity-based classification methods generally operate in a multi-dimensional feature space ($d \geq 1$). Each feature consists of an image intensity at the spatial location (voxel) to be classified; all the features are derived from the same subject. Such classification techniques are, in fact, not medical imaging specific – an extensive coverage of classifiers is given in Duda et al. (2001).

Many researchers have applied to brain MRI classic methods such as the

Bayes (maximum likelihood) classifier (Collins et al., 2001; Kamber et al., 1995), or non-parametric classifiers like kNN (k nearest neighbors) (Warfield et al., 2000) and ANN (artificial neural network) (Zijdenbos et al., 1998). Expectation-Maximization (EM) is a popular statistical classification scheme for this application. Originally proposed in a brain MRI context by Wells III et al. (1996), and further improved by many others (e.g.: Ashburner, 2000; Ashburner and Friston, 2000; Guillemaud and Brady, 1997; Held et al., 1997; Pohl et al., 2002; Schroeter et al., 1998; Van Leemput et al., 1999a,b), these methods interleave intensity non-uniformity field estimation (correction) and classification, in an iterative fashion.

All MRI classification methods are sensitive to overlap in the tissue intensity distributions. Such overlaps are caused by inherent limitations of the image acquisition process, such as noise, intensity non-uniformity (INU, also known as bias field) (Sled and Pike, 1998), and partial volume effect (as a consequence of the finite resolution of the imaging process, the image voxels may contain a mixture of more than one tissue type, which all contribute to the measured signal). Several approaches have been proposed to address this limitation of intensity-based classification. For example, post-classification morphological operations or contextual classifiers (Choi et al., 1991; Held et al., 1997; Rajapakse et al., 1997; Udupa and Samarasekera, 1996; Yan and Karp, 1995). Moreover, a number of researchers have proposed continuous classifiers which attempt to estimate the mixing proportions of several tissues in a voxel (i.e. the partial volume effect) (Choi et al., 1991; Laidlaw et al., 1998; Pham and Prince, 1999; Schroeter et al., 1998; Van Leemput et al., 2002). Another approach, in which the limitations of intensity-based classification are addressed by constraining it with the non-linearly deformed anatomical template, was proposed by Warfield et al. (2000).

The main contribution of the work we present here is a novel method for fully automatic generation of correct training samples for tissue classification. The method is non-parametric, hence does not make any assumptions about the feature space distributions. It is based on a prior tissue probability map in a standard, brain-based coordinate system (the “model”), and is designed to accommodate subject anatomies that are significantly different than the model (the difference can be due to aging, due to pathology, or even due to anatomical variability between normal individuals of similar age).

In Section 2 we explain the problem we address here, and why previously existing solutions are unsatisfactory. In Section 3 we describe our new method. Section 4 presents validation experiments and their results. We conclude with a discussion and comparison with other approaches in Section 5.

2 Problem Statement

2.1 Fully automatic classification

Manual, or even semi-automatic, classification performed by a trained expert is labor-intensive (hence impractical for processing large amounts of data), highly subjective, and non-reproducible (Zijdenbos et al., 1998, 2002). Fully automatic, robust tissue classification is required for the quantitative analysis of MRI data from large-scale (150 – 1000 subjects), possibly multi-site clinical trials or research projects (e.g.: Zijdenbos et al., 1998, 2002).

The MRI intensity scale has no absolute, physical meaning: the image values and contrast are dependent on the pulse sequence, and other variable scanner and post-processing parameters. Thus, the ability of a tissue classification method to automatically adapt to a new MRI dataset is especially important when the data is collected at multiple sites or with several different MRI scanners.

An aspect that is often ignored by brain MRI classification schemes is how to adapt to a new MRI dataset in a fully automated manner. Some researchers have addressed this issue:

- The use of stereotaxic space tissue probability maps (a probabilistic brain atlas) for automating supervised classification algorithms was originally proposed by Kamber et al. (1995), and subsequently used by others (Kollokian, 1996; Zijdenbos et al., 1998, 2002).
- Automatic implementations of the popular Expectation-Maximization (EM) statistical classification scheme were proposed by Van Leemput et al. (1999a,b) and by Ashburner (2000), Ashburner and Friston (2000). These methods use a probabilistic brain atlas to initialize, and also to constrain, the iterative EM process.

However these methods can fail for “atypical” brain scans (significantly different from the atlas), such as child brains, or brains with large pathological abnormalities[1] . All these classification methods start by spatially registering the subject’s MRI to the probabilistic atlas using a linear (rigid body) transformation. Work very recently presented (D’Agostino et al., 2002; Pohl et al., 2002) showed that using elastic (non-linear) registration of the subject to the atlas can improve the performance of Van Leemput’s method.

However, as of this writing these results are preliminary and comprehensive

[1] This was shown by our experience with the method of Kamber et al. (1995), and was reported by both Van Leemput et al. (1999b) and Ashburner (2000).

validation is still pending.

2.2 Feature space distributions

The classification procedure we describe in this paper is *non-parametric*. Many of the brain tissue classification methods proposed in the literature employ *parametric* classifiers – they assume the data distributions in feature space follow a certain model (notable exceptions: Udupa and Samarasekera, 1996; Warfield et al., 2000; Zijdenbos et al., 1998). Typically, the multi-variate Gaussian model ("Normal" distribution) is used. If the features are MR signal intensities from various MRI modalities (such as T1, T2, PD),

then the Gaussian model assumption can be poor (Clarke et al., 1993; DeCarli et al., 1992; Schellenberg et al., 1990) : besides biological causes such as the intrinsic heterogeneity within the tissue classes that concern this paper (CSF, grey matter, white matter), the MRI acquisition artifacts also affect the intensity distributions – i.e. result in deviations from a Normal distribution (Ashburner, 2000; Kollokian, 1996; Schellenberg et al., 1990). While intensity non-uniformity can be reduced by retrospective correction methods (e.g.: Sled et al., 1998), the partial volume effect cannot.

Even if the normal-distribution assumption would be acceptable for some data, it is safer not to make it if the automatic classification method aims to be robust against variability in the imaging data quality. Robustness is especially important for unsupervised processing of data collected in large-scale, multi-site research projects or clinical trials.

2.3 Model-based training set selection

In this paper, "stereotaxic space" is a standard frame of reference defined by anatomical landmarks; it allows for the removal of affine differences (rotation, translation, scale) between brains. A stereotaxic space tissue probability map (TPM)[2] of a given tissue is a spatial probability distribution representing a certain subject population. For each spatial location in the stereotaxic space, the TPM value at that location is the probability of the given tissue to be observed there, for that particular population.

Once imaging data is spatially registered (normalized) to the stereotaxic space by means of an affine transformation, the TPM-s provide an a-priori spatial

[2] A TPM is also sometimes referred in the literature as "Statistical Probability of Anatomy Map (SPAM)", or "probabilistic brain atlas".

probability distribution for each tissue (Fig. 1). This distribution can be used to automatically produce a training set for the supervised classifier (Kamber et al., 1995). For example, choose spatial locations that have a TPM value $\geq \tau = 0.99$ (99%); the TPM will provide the class label for the training sample, and the actual MR image value(s) at that spatial location will provide the sample's feature vector. However, this simplistic approach has two limitations:

Mis-labeled samples: Even among the locations with very high a-priori probability of being a given tissue, some of them will in fact be from another class. There are several reasons for this.

First, the morphology of the human brain is highly variable and the TPM is created from a finite sample.

Second, in practice the automatic linear registration of the subject to the stereotaxic space will not be perfect. Lastly, in practice the TPM is not computed from ground-truth (which cannot be obtained in-vivo) but from semi-automatic segmentations of MR images (Kamber et al., 1995; Kollokian, 1996); any systematic errors or bias of the method will propagate into the TPM.

The fraction of mis-labeled samples in the training set will increase when the minimum prior probability threshold τ is decreased. Also, for a given τ, this fraction will be larger when the subject is from a different population than the population statistically represented by the TPM [3].

Intensity distribution estimation: For highest τ (where the rate of mis-labeled samples is lowest) the qualifying sample points give a very limited coverage of the brain area (Fig. 2). Using these points will not yield a good estimate of the true tissue intensity distributions (which is needed by a supervised classifier), for two reasons:

- MRI artifacts, such as intensity non-uniformity (INU), introduce spatial variations in the image intensity of any given tissue type.
- Due to the underlying biology, the MRI signal intensity of the main brain tissue types is not homogeneous throughout the brain (Kandel et al., 2000).

Thus, more spatial coverage, by sampling at a lower τ, would be beneficial for the intensity distribution estimation provided by the training set. However, lowering τ also results in an increased number of incorrectly labeled samples in this set.

[3] Pathology is a common cause of significant deviations of a subject's anatomy from a normal brain model.

In the following, we present a procedure which allows for a lower τ while limiting the rate of mis-labeled training samples. Specifically, an automatic "pruning" of the raw set of points obtained from the TPM is performed.

3 Method

Our fully automatic, non-parametric, brain tissue classification procedure (Fig. 3) consists of two stages:

(1) A novel semi-supervised classifier, using a minimum spanning tree graph-theoretic method and stereotaxic space prior information. It produces a set of training samples customized for the particular individual anatomy subjected to classification. This stage will be referred to as the "pruning" stage, and is described in Section 3.1.

(2) A supervised non-parametric classifier trained on the set of samples produced by the first stage. This training set provides an estimate of the tissue intensity distributions in the actual MR dataset subjected to classification. This stage is described in Section 3.2.

The features used are signal intensities of one or more MRI modalities. If multiple MR contrasts of the same subject are used, the different acquisitions need to be spatially aligned to each other in a pre-processing step.

The feature space proximity measure we used is the common Euclidean distance in d-dimensional space. In a pre-processing step, the intensity values along each dimension are normalized by a histogram range-matching procedure[4] .

3.1 Pruning stage

The pruning works on an input set of spatial locations that are selected through random sampling from the qualifying spatial locations in the respective tissue probability map (TPM); an equal number of samples is selected for each tissue class (background, CSF, grey matter, white matter). The qualifying locations are locations where the TPM value (i.e. the prior probability) is $\geq \tau$ (Section 2.3).

[4] Points located a small percentile away from the absolute minimum/maximum are used as robust estimators of the histogram's range. We heuristically determined that 4/0.5/4% percentiles are adequate for T1/T2/PD.

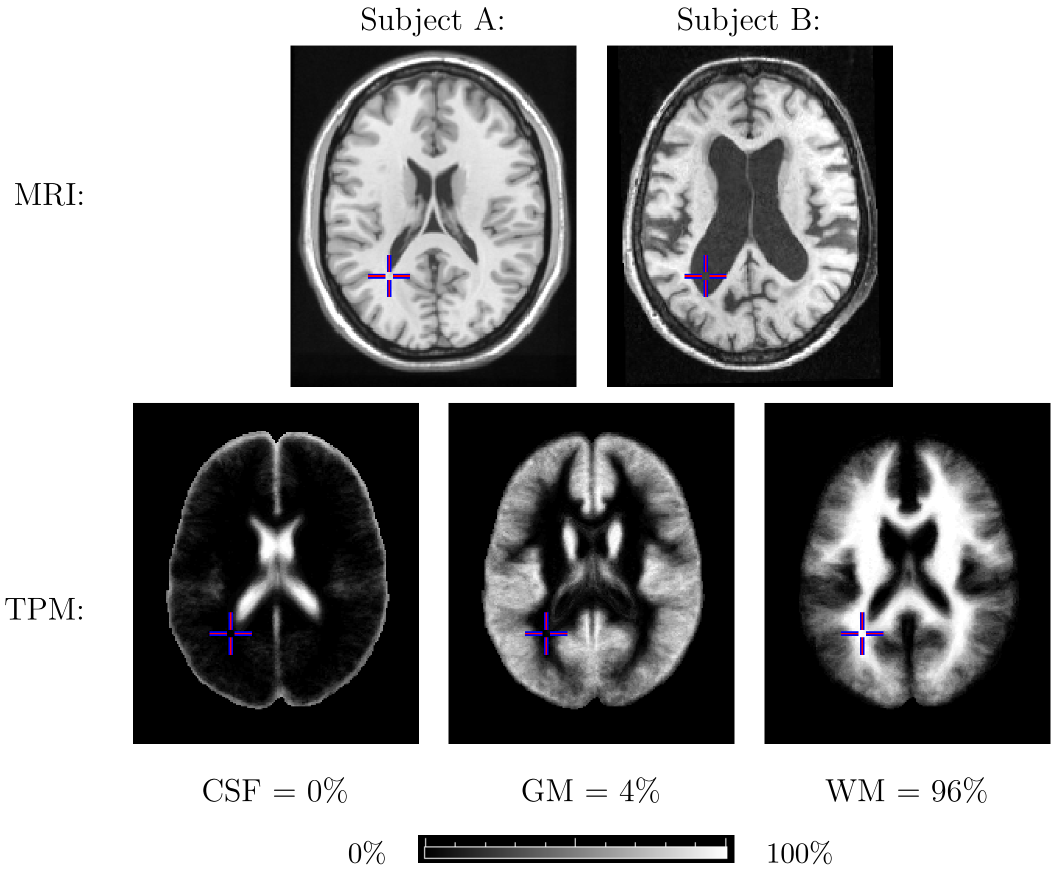

Fig. 1. For a given spatial location in the 3D stereotaxic space, Tissue Probability Maps (TPM) provide an *a priori* probability for a tissue to be found there. In this example, the TPM-s are computed from a young-normal population (N=53), subject A is a young-normal individual, and subject B is an elderly Alzheimer's Disease patient (who exhibits significant brain atrophy with enlarged ventricles). Note that a location ('+' in images) with very high prior probability to be white matter is in fact CSF in subject B. Although 2D sections are shown, all these data are 3D.

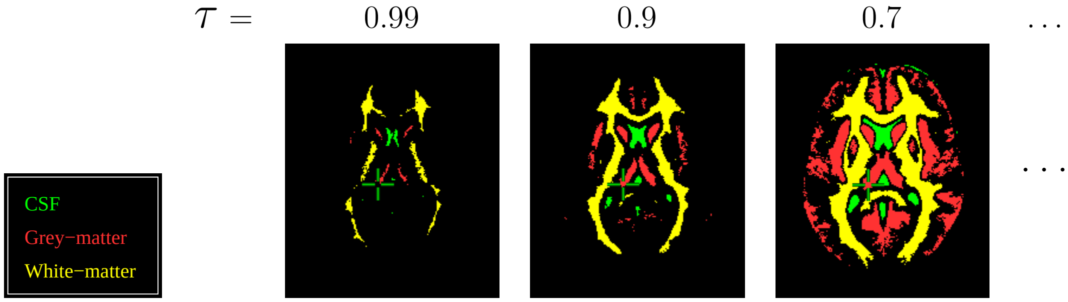

Fig. 2. Brain coverage of the training set obtained from the 3D prior tissue probability maps (TPM) for different $\mathcal{T}$ thresholds (training samples are selected from spatial locations with TPM value $\geq \mathcal{T}$). Note the very limited spatial coverage at high values of $\mathcal{T}$, especially for CSF.

The pruning technique makes use of a minimum spanning tree (MST) in feature space[5]. We refer to this method as "semi-supervised" because, unlike in traditional unsupervised classification, some prior information exists in this application: the number of main clusters[6] is known (to be 4), and each input point has an initial labeling suggested by the TPM-based sample selection process. The purpose of the pruning is to reject the points with incorrect labeling.

Here are the three main steps of the pruning method, followed by a more detailed description of the important parts:

(1) The minimum spanning tree of the input set of points is constructed in feature space (Fig. 4-top).
(2) In an iterative loop, the graph is broken into smaller connected components (clusters) by removing "inconsistent" edges from the initial MST. At each iteration, the *main clusters* are identified and labeled by using prior knowledge, and a stop condition is tested on them. If the condition is not satisfied, the graph breaking loop is continued.
(3) After the iterative process stops, the points that are in the "correct" cluster (i.e. have the same initial labeling as their cluster) are deemed to be correctly labeled and are kept; all the other points are deemed to be incorrectly labeled and are discarded[7].

MST computation: This work uses Kruskal's algorithm (Albertson and Hutchinson, 1988; Kruskal, 1956). By employing union-by-rank and path-compression methods (Cormen et al., 1990) for an efficient connected components implementation, the time complexity of our implementation (for an arbitrary feature space dimensionality d) is $O(n^2 \log n)$ for large n, where n is the number of input points. This complexity can be further reduced to $O(n \log n)$ in 2-dimensional space using the property that the Euclidean-distance minimum spanning tree is a subset of the Delaunay triangulation (O'Rourke, 1998). In 1-dimensional space, computing the Euclidean-distance MST is equivalent to sorting, which is $O(n \log n)$.

MST breaking loop: A heuristic method (inspired by Duda et al., 2001) was implemented and experimentally evaluated (in Section 4). It uses a thresh-

[5] A MST of a set of points in d-dimensional space is defined as a graph that connects all the points, has no cycles, and whose sum of all edge lengths is as small as possible.

[6] There could be other, smaller, clusters produced by acquisition artifacts, or by other brain tissue classes than the main four (background, CSF, GM, WM), such as fat or skull.

[7] Justification: their prior probability (given by the TPM) is very low.

old value R, which is decreased at each iteration of the algorithm and tested on all edges of the graph in parallel:

- an edge (i, j) is removed if $length(i, j) > R \times A(i)$ or if $length(i, j) > R \times A(j)$, where $A(i)$ is the average length of all the other edges incident on node i (Fig. 4-bottom).

Note that smaller R values will result in more edges being removed, hence result in more, and smaller, clusters. To minimize unnecessary graph fragmentation, it is desirable to use the largest value of R that satisfies our stop condition. The appropriate value of R is adaptive: for each particular input MRI dataset, an approximation of R is automatically computed by our algorithm in an iterative loop that tests progressively smaller R values until the stop condition is satisfied.

Main clusters identification: The main clusters are the best guesses for the true background, CSF, grey matter, and white matter clusters in feature space. Under the assumption that the majority of points have correct initial labels, the best guess for each class is the cluster which contains the largest number of points labeled as that class.

Note that early in the iterative process some of these main clusters will not be distinct (they are still connected for the current value of R); in other words, the same cluster will contain most samples from class i, but also most samples from class j.

Stop condition: The loop stops when the main clusters, identified as above, are four disconnected clusters.

3.2 Final classification stage

The supervised classifier we used in our implementation is the classic k nearest-neighbor (kNN) classifier. For each data point to be classified, kNN computes this point's closest k training samples in feature space; then, the data is classified with the label most represented among these k nearest neighbors. kNN is attractive because it is a "non-parametric" classifier: it can learn, from the training set, data feature distributions of arbitrary shape (see also Section 2.2).

For good performance, the size of the training set (n) should be as large as is practical in order to get a good estimate of the true feature space class distributions. It was shown (Devroye, 1981; Stone, 1977) that if k satisfies both of the conditions:

$$\lim_{n\to\infty} k = \infty$$

$$\lim_{n\to\infty} \frac{k}{n} = 0$$

then the kNN classifier (for $n \to \infty$) will have an error probability equal to the optimal "Bayes risk" R^* – the minimum possible error probability of a generic classifier (achieved when the classifier knows the true data distributions).

For example, a k that satisfies the above conditions is $k = \sqrt{n}$ (for small n, Enas and Choi (1986) suggest $k = n^{0.25} \dots n^{0.375}$). In the experiments presented in this paper (Section 4) we chose $k = 45$, and used $n > 3000$ training samples per class (i.e. $k \approx n^{0.48}$). This large n might bring only marginal classification improvements over smaller values (and the computational requirements are proportional to n). However, the goal of our validation experiments was to assess the performance of the novel pruning stage; therefore, we conservatively over-specified the final supervised classification stage in an attempt to reduce its contribution to the errors of the overall two-stage method (Fig. 3).

A likely reason why the kNN classifier was not used more in the past is that its straight-forward (naive) implementation is slow. However, kNN's computational requirement can be reduced by several techniques (see Duda et al., 2001, Section 4.5.5), and the computing power of commonly available computers steadily and quickly increases. Our implementation uses a fast nearest-neighbor lookup library (publically available: Mount and Arya, 1998), which pre-processes the training set using box-decomposition (BD) trees (Arya and Mount, 1993; Arya et al., 1998).

By employing this library we experienced a reduction of about 100 times in the necessary computation time compared to the straight-forward kNN implementation.

Post-processing: The classifier sometimes incorrectly labels as brain tissue some regions which are outside the brain itself, such as dura, skin, or fat. This is an inherent limitation (especially when only one MR modality is available) of any "pure" brain tissue classification method that considers each voxel independently and uses only the MRI intensity information at that voxel. However, such non-brain tissue can be masked out by a skull removal procedure (also known as skull stripping, or brain extraction) that uses spatial position, and possibly also voxel neighbourhood information, in addition to the MRI intensity. Several such automatic procedures exist, for example: Hahn and Peitgen

(2000); Justice et al. (1997); MacDonald et al. (2000); Smith (2002); Stokking et al. (2000).

4 Experiments and Results

The prior stereotaxic space probabilistic anatomy model (TPM) used in these experiments (Fig. 1) was produced (Kamber et al., 1995; Kollokian, 1996) from a set of T1/T2/PD MRI scans of 53 young individuals (aged 18 to 35) using a semi-automatic tissue classification method. The stereotaxic space is the Talaraich space (Talairach and Tournoux, 1988) and is defined by brain anatomical landmarks. The 53 individual scans were spatially registered to this standard frame of reference using a linear (rigid body) transformation with 9 degrees of freedom (Collins et al., 1994). The class-C TPM value of each voxel was computed as the frequency with which that location was labeled as C among the 53 individuals.

We performed validation experiments on both simulated and real data, and on both single-spectral (T1) and multi-spectral (T1+T2+PD) MRI data. Moreover, to validate the robustness against morphological deviations from the TPM, we tested the classification method on elderly and on diseased subjects – aging and pathology typically cause significant deviations from a young-normal model. Specifically, experiments were performed on the following data:

(1) realistic MRI simulations driven by a new custom set of 10 brain "phantoms" resembling elderly brains.
(2) a real MRI dataset from a young normal individual, which was fully manually segmented by a human expert.
(3) real multi-spectral MRI scans of 31 ischemia patients.
(4) real multi-spectral MRI scans of 11 Alzheimer's Disease elderly patients.

Although single-feature (T1 only) experiments were performed on the datasets 1 and 2, the results were qualitatively similar to the multi-feature (T1-T2-PD) experiments, so in the interest of brevity we will not show them here.

The quantitative measurements were performed with repetitions to assess the statistical significance of each data point – the repetitions were over 10 different "subjects" (for data 1), and over 10 different raw sets of samples (for data 2).

To validate the novel pruning stage, we also performed classification experiments without pruning (termed as "raw" in the following): the supervised kNN classifier was directly trained with the raw samples extracted from the TPM (see Fig. 3).

For a quantitative measure of classification performance, we computed the *Kappa* metric against a "gold standard." *Kappa* (Cohen, 1960) is a chance-corrected similarity measure between two labelings, defined as follows: given a C-class classification problem for a set of N samples, if ($\forall$ class i) we denote by a_i the total number of samples labeled as i by both classifiers, and by s_i, t_i the total number of samples labeled as i by the first, respectively by the second classifier, then:

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed proportion of agreement (also known as "accuracy"):

$$P_o = \frac{1}{N} \sum_{i=1}^{C} a_i$$

and P_e is the expected (due to chance) proportion of agreement:

$$P_e = \frac{1}{N^2} \sum_{i=1}^{C} s_i t_i$$

The maximum *Kappa* value of 1 corresponds to a perfect agreement ($P_o = 1$), and a value of 0 corresponds to agreement due to chance alone ($P_o = P_e$, i.e. the two labelings are completely independent events). We should point out that P_o (and consequently *Kappa*) does not necessarily give the same weight to the labeling accuracy of each of the individual classes – e.g. the label similarity for an over-abundant class i will have more effect on the overall *Kappa* value than the similarity for a less-abundant class j. Instead, P_o gives the same weight to the label similarity of each individual sample. Nevertheless, the ratio of the brain CSF : Grey : White image voxels is typically 1 : 2.3 : 1.9 (for the stereotaxic space we used), so over-abundance of one class is not a significant issue in this work.

4.1 Elderly brain simulated MRI

These data were produced by a MRI simulator (Kwan et al., 1999) that produces realistic synthetic MRI images based on an anatomical model (a "phantom")[8]. 10 phantoms resembling 10 different elderly brains were created as

[8] This simulator is accessible through the Internet (Cocosco et al., 1997; MNI, 1997), and is widely used by the research community for validation.

follows:

(1) T1 scans of 10 individuals (60-70 years old, 5 males, 5 females)[9] were non-linearly spatially registered (Collins and Evans, 1997) to a previously available standard phantom (Collins et al., 1998; MNI, 1997).
(2) The resulted deformation field was inverted and used to deform the standard phantom, such that it looks similar to the source individual brain (Fig. 5).

Then, T1, T2, and PD MRI-s were simulated as $1\,mm^3$ voxel acquisitions, with 3% noise and 20% INU (intensity non-uniformity). These values were previously reported as typical artifact severity[10] (Kollokian, 1996; Sled et al., 1998; Zijdenbos et al., 2002). Sample MRI images are in Fig. 7.

We computed the Kappa classification performance measure, over the entire brain area, against the digital phantom used to drive the simulations.

Fig. 6 shows that the pruning brings a statistically significant improvement in the Kappa metric over the "raw" method. Also, the plot does not show any significant variation in the Kappa for our method ("pruned") when $0.30 \leq \tau < 1.00$. Sample classification outputs are in Fig. 7.

4.2 Real data: young-normal individual

We also performed a quantitative validation on a real multi-spectral MRI dataset of a 36 year old healthy individual. The T1-weighted $1\,mm^3$ 3D scan was completely manually classified (except the cerebellum) (Kabani et al., 1997, 1998) by a trained neuroanatomist[11]. T2 and PD scans were also acquired as $2\,mm$ thick sagittal slices ($1\,mm^2$ in plane), in two acquisitions offset by $1\,mm$; the two paired scans were spatially co-registered and averaged together in order to improve the image resolution. All three MRI modalities were then spatially registered to the Talairach stereotaxic space using automatic rigid-body (9-parameter) registration software (Collins et al., 1994). Also, INU correction was performed using the method of Sled et al. (1998).

[9] Data source: Dr. Ryuta Kawashima, Sendai, Japan.

[10] In a practical automatic image analysis system, an INU correction procedure (such as Sled et al., 1998) would typically be employed. However, here we wanted to simulate the worst case scenario: when the INU correction is not available or it fails.

[11] This segmentation was done completely manually, without the aid of any semi-automatic intensity-based tools. Besides the MR image intensity, the neuroanatomist used additional information such as spatial context and expert neuroanatomical knowledge.

We used this manual segmentation as the "gold standard" for computing the Kappa figure of merit of the classification (Fig. 8). The plot does not show any significant variation in the performance of our method ("pruned") for choices of τ in the range $0.50 \leq \tau < 1.00$ (the median Kappa only varies from 0.772 to 0.783, which is negligible).

Moreover, these quantitative measurements do not show an overall statistically significant improvement of the pruned classification over the not-pruned ("raw") one. Nevertheless, this individual is morphologically similar to the TPM used, so the "raw" classification is already of good quality[12] .

4.3 Real data: ischemia patients

These data are multi-spectral T1-T2-PD MRI datasets of 31 patients diagnosed with ischemia. T1: $1\,mm^3$ resolution, T2/PD: 1x1x3.5$\,mm$ resolution; same pre-processing (spatial normalization, INU correction) as for the dataset of Section 4.2. Ischemia leads to brain atrophy, hence to a significant morphological difference from the young-normal TPM.

We qualitatively evaluated the classification results by visual inspection. On some of these 31 datasets the classification result without pruning was poor, and the addition of the pruning stage significantly improved the result (Fig. 9). On the other datasets the pruning produced only subtle improvements (the no-pruning classification was already of acceptable quality).

4.4 Real data: Alzheimer's Disease elderly patients

Multi-spectral MRI scans (T1-T2-PD) were acquired for 11 elderly patients (aged > 60) diagnosed with Alzheimer's Disease (which causes cortical atrophy), as part of a clinical study. The T1 was of $1\,mm^3$ resolution. The T2 and PD were acquired as 8 interleaved sagittal-slice acquisitions; the 8 scans were then spatially co-registered and spliced together in order to obtain a single final 3D image of $1\,mm^3$ resolution. Finally, the same pre-classification processing as for the dataset of Section 4.2 (spatial normalization, INU correction) was applied to all three MRI modalities.

A qualitative evaluation of the classification results revealed a clear improve-

[12] Moreover, preliminary results (which are beyond the scope of this paper) indicate that, with a large training set and an appropriate k (Section 3.2), the kNN classifier is more robust against incorrectly labeled training samples than other commonly used classifiers, e.g. maximum-likelihood with a Gaussian density estimator.

ment on some datasets when the pruning stage was employed (Fig. 10). While the improvements were subtle on the other datasets, pruning never degraded the classification result.

5 Discussion and Conclusion

5.1 Contributions

- We described and validated a completely automatic procedure for brain tissue classification from MR anatomical images. The procedure can take as input any number of MR imaging modalities of the same subject.
- Our procedure is robust against significant morphological differences between the subject of the classification and the particular probabilistic anatomical model used for initialization. Moreover, the implementation is non-parametric in that it does not make any assumptions about the tissue image intensity distributions.
- The performance of the procedure was demonstrated by quantitative and qualitative experiments on both simulated and real MRI data, and on subjects who are both similar and dissimilar to the anatomical model used for initialization.
- Although this procedure requires a probabilistic anatomical atlas (model) for initialization, it is not restricted to the particular model that we used, nor to the method we used to spatially register the subject to the atlas. Any other prior model and spatial registration procedure can be used; the only requirement is that the majority of samples in the initial training set (provided by the model) have correct tissue labels. More specifically, the requirement is that the "main cluster identification" guess (page 10) is valid.
- Our pruning procedure is general: one is not restricted to using kNN for the second (supervised) classification stage. The implementation we validated in these experiments uses kNN (for the reasons presented in Section 3.2), but another supervised classifier could be used instead. A non-parametric classifier is advisable, for the reasons presented in Section 2.2.

5.2 Other approaches

The advantage of MR-intensity-based tissue classification over other image segmentation methods (e.g. deformable models) is its ability to produce high quality definition of tissue boundaries. This is especially important for human brain tissue classification, where highly curved interfaces between tissues (such as between gray and white matter) can be challenging to recover from finite resolution images.

In Section 2 we explained the shortcomings of previously existing methods for fully automatic MRI tissue classification. Our novel method is designed to accomodate "atypical" brain anatomy (i.e. subjects significantly different from the brain model). In particular, the validation experiments showed that our new method performs better on atypical subjects than the traditional "raw" method – the latter method is similar to the one originally proposed by Kamber et al. (1995), but implemented using a non-parametric classifier (kNN). In addition, our new procedure is more robust against imaging artifacts and variability in the MRI data quality because it has a non-parametric implementation: as explained in Section 2.2, the Normal-distribution assumption (made by parametric methods such as: Ashburner and Friston, 2000; Van Leemput et al., 1999a,b) is not safe for brain MRI tissue classification.

The methods of Van Leemput et al. (1999a,b) and of Ashburner and Friston (2000), Ashburner (2000) use the prior probability (at that particular spatial location in stereotaxic space) in the classification decision for each voxel. This approach prevents the resulting classification from being significantly different from the prior probability model. If the subject is seriously "atypical" (e.g. a patient with severe brain atrophy) then there will be serious errors in the classification result. Our method is more adaptive because it uses the spatial prior only for initialization, and not in the classification decision – which is based only on actual image information.

The method by Warfield et al. (2000) improves on the inherent limitations of intensity-based classification by combining it with an elastically deformed atlas. This technique has an advantage for situations when intensity data alone is insufficient to distinguish the tissues (e.g. neonate brains). However, this method currently needs manual supervision, and it still needs to be investigated how reliably can this method be automated.

Moreover, current elastic (non-linear) registration methods are not robust when there is a significant topological difference between the subject and the target brain. Another limitation of elastic matching is that in practice the deformation field is band-limited and cannot fully match the cortical folding patterns of the human brain.

Consequently, even if one uses an elastic deformation of the subject MRI to the probabilistic atlas (such as in recent preliminary work: D'Agostino et al., 2002; Pohl et al., 2002), this may not completely eliminate the mis-labeled samples in the training set suggested by the atlas [13]. Hence, our pruning approach

[13] In theory, one would have no mis-labeled samples if the prior probability atlas would be generated from a very large ("infinite") sample, and if the elastic deformation field would have sufficiently high spatial frequencies. In practice however, such an atlas would require vast resources for data acquisition and processing, and such a deformation would take a very long time to compute. The pruning approach

will still be needed in order to have a robust classification procedure.

An intuitive approach for classifying MRI data for a large set of individuals from a certain population is to first obtain a probabilistic anatomy atlas representing that particular population. However, our method is more general (hence, more reproducible): its pruning strategy allows one to use a less specific atlas instead – e.g. use a young-normal atlas for classifying subjects from a broad range of ages. Besides, the generation of a new probabilistic atlas for a certain population requires non-trivial resources.

5.3 Limitations and future work

In a typical anatomical MRI the tissue intensity distributions partially overlap. Consequently, our pruning method eliminates some of the training set samples that are positioned in feature space in between the clusters (note that this overlap is an inherent problem for any intensity-based discrete voxel labeling method). One cause of this distribution overlap is the finite spatial resolution of the image acquisition: voxels at the boundary between tissue types have more than one tissue contributing to the measured signal (partial volume effect). It would be desirable for the pruning method to put less trust in the intensity of such voxels than in the intensity of pure tissue voxels. The boundary (partial volume) voxels will be located in high gradient areas of the MR image, thus the incorporation of local gradient information into the algorithm is worth exploring.

The classification method we presented here is designed for head MRI-s without significant amounts of pathological brain tissue (e.g. lesions, tumors); the extension of our method to such pathological subjects is a topic of future research. Nevertheless, if the volume of pathological tissue is relatively small compared to the healthy tissue then the presented algorithm will still correctly label the healthy tissue voxels – the cluster identification procedure (page 10) will never select the relatively small cluster representing the pathological tissue as one of the "main clusters."

Acknowledgements

We would like to thank: John Sled, Steve Robbins, Peter Neelin, Louis Collins, Jean-François Mangin for valuable feedback on earlier versions of this paper; Noor Kabani for neuroanatomy expertise and for the manual segmentation

we presented here is a more economical solution.

of Section 4.2; the anonymous reviewers for useful comments and suggestions; Guido Gerig and David Gering for their comments on the MICCAI conference version of this paper. In addition, the first author would like to thank Godfried Toussaint for his inspiring "Pattern Recognition" course.

A Implementation details

The implementation of the method described in Section 3 has to take into account practical limitations on the input size (n) for the graph-theoretic pruning method (Section 3.1). There are two causes for this:

(1) The computational requirement — the MST computational complexity is $O(n^2 \log n)$ for multi-spectral input consisting of three or more MRI modalities.
(2) The limited precision of the MRI data representation — 12-bit, or even 8-bit, integer data is typical (only 4096, respectively 256, possible values). A large n will result in an over-population of the discretely-sampled feature space, and consequently in a reduced ability to compare distances between points in this space.

However, the final supervised kNN classifier is less affected by the above two issues. Our practical implementation solution (used for all the experiments presented in this paper) was to perform multiple independent pruning steps on smaller sets of "raw" samples, and then to merge all the resulting pruned samples into a larger set of samples that is fed to the kNN (Fig. A.1).

B Box Plots

Figures 6 and 8 present experimental results using "box and whisker" plots (produced using Matlab, The MathWorks Inc., 1998). The meaning of the various graphical symbols on the boxes is as follows:

- box has (horizontal) lines at the lower quartile (25% percentile), median, and upper quartile (75% percentile) values of the data sample.
- whiskers show the extent of the rest of the data; their length is $\leq 1.5\times$ the height of the box.
- outliers (represented as "+") are data points beyond the ends of the whiskers.

A side-by-side comparison of two boxes is the graphical equivalent of a t-test: if the boxes do not overlap in their vertical extent, then $p < 0.05$.

References

Albertson, M., Hutchinson, J., 1988. Discrete Mathematics with Algorithms. Wiley, New York.

Arya, S., Mount, D., 1993. Algorithms for fast vector quantization. In: Storer, J., Cohn, M. (Eds.), Proc. of DCC '93: Data Compression Conference. IEEE Computer Society Press, pp. 381–390.

Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A., 1998. An optimal algorithm for approximate nearest neighbor searching. Journal of the ACM 45 (891-923).

Ashburner, J., 2000. Computational neuroanatomy. Ph.D. thesis, University College London.

Ashburner, J., Friston, K. J., Jun. 2000. Voxel-based morphometry — the methods. Neuroimage 11 (6), 805–821.

Choi, H. S., Haynor, D. R., Kim, Y., Sep. 1991. Partial volume tissue classification of multichannel magnetic resonance images - a mixel model. IEEE Trans Med Imaging 10 (3), 395–407.

Clarke, L. P., Velthuizen, R. P., Phuphanich, S., Schellenberg, J. D., Arrington, J. A., Silbiger, M., 1993. MRI: stability of three supervised segmentation techniques. Magn Reson Imaging 11 (1), 95–106.

Cocosco, C., Kollokian, V., Kwan, R.-S., Evans, A., May 1997. Brainweb: Online interface to a 3D MRI simulated brain database. In: NeuroImage (Proc. of HBM'97). Vol. 5 (4, part2/4). Academic Press, p. S425.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurements 20, 37–46.

Collins, D., Evans, A., Dec. 1997. Animal: validation and applications of non-linear registration-based segmentation. International Journal of Pattern Recognition and Artificial Intelligence 11 (8), 1271–1294.

Collins, D., Montagnat, J., Zijdenbos, A., Evans, A., Arnold, D., Jun. 2001. Automated estimation of brain volume in multiple sclerosis with BICCR. In: Insana, M. F., Leahy, R. M. (Eds.), Proc. of IPMI 2001. Vol. 2082 of LNCS. Springer-Verlag, pp. 141–147.

Collins, D. L., Neelin, P., Peters, T. M., Evans, A. C., Mar./Apr. 1994. Automatic 3D inter-subject registration of MR volumetric data in standardized Talairach space. Journal Comput Assist Tomogr 18 (2), 192–205.

Collins, D. L., Zijdenbos, A. P., Kollokian, V., Sled, J. G., Kabani, N. J., Holmes, C. J., Evans, A. C., Jun. 1998. Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 17 (3), 463–8.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., 1990. Introduction to Algorithms. MIT Press, Cambridge, MA.

D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2002. A viscous fluid model for multimodal non-rigid image registration using mutual information. In: Dohi, T., Kikinis, R. (Eds.), Proc. of MICCAI 2002. Vol. 2489 of LNCS. Springer-Verlag, pp. 541–548.

DeCarli, C., Maisog, J., Murphy, D. G. M., Teichberg, D., Rapoport, S. I., Horwitz, B., Mar./Apr. 1992. Method for quantification of brain, ventricular and subarachnoid CSF volumes from MR images. Journal Comput Assist Tomogr 16 (2), 274–284.

Devroye, L., 1981. On the almost everywhere convergence of nonparametric regression function estimates. Annals of Statistics 9, 1310–1319.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern classification, 2nd Edition. Wiley.

Enas, G., Choi, S., 1986. Choice of the smoothing parameter and efficiency of k-nearest neighbour classification. Computers and Mathematics with Applications 12A (2), 235–244.

Fischl, B., Dale, A. M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences 97 (20), 11044–11049.

Guillemaud, R., Brady, M., Jun. 1997. Estimating the bias field of MR images. IEEE Trans Med Imaging 16 (3), 238–51.

Hahn, H. K., Peitgen, H.-O., Oct. 2000. The skull stripping problem in MRI solved by a single 3D watershed transform. In: Delp, S. L., DiGioia, A. M., Jaramaz, B. (Eds.), Proc. of MICCAI 2000. Vol. 1935 of LNCS. Springer-Verlag, pp. 134–143.

Held, K., Kops, E. R., Krause, B. J., Wells 3rd., W. M., Kikinis, R., Muller-Gartner, H. W., Dec. 1997. Markov random field segmentation of brain MR images. IEEE Trans Med Imaging 16 (6), 878–86.

Jones, S. E., Buchbinder, B. R., Aharon, I., 2000. Three-dimensional mapping of cortical thickness using Laplace's equation. Hum Brain Mapp 11, 12–32.

Justice, R. K., Stokeley, E. M., Strobel, J. S., Ideker, R. E., Smith, W. M., Feb. 1997. Medical image segmentation using 3-D seeded region growing. In: Proc. of SPIE: Image Processing. Vol. 3034. pp. 900–910.

Kabani, N., Collins, L., Evans, A., Oct. 1997. Hemispheric differences in gray matter volume of adult human brain. In: Proc. of Society for Neuroscience Annual Meeting. New Orleans-LA, USA.

Kabani, N., MacDonald, D., Holmes, C., Evans, A., Jun. 1998. 3D atlas of the human brain. In: Neuroimage (Proc. of HBM'98). Vol. 7(4). Academic Press.

Kamber, M., Shinghal, R., Collins, D. L., Francis, G. S., Evans, A. C., Sep. 1995. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. IEEE Trans Med Imaging 14 (3), 442–53.

Kandel, E. R., Schwartz, J. H., Jessel, T. M., 2000. Principles of Neural Science, 4th Edition. McGraw Hill.

Kollokian, V., Nov. 1996. Performance analysis of automatic techniques for tissue classification in magnetic resonance images of the human brain. Master's thesis, Computer Science, Concordia University, Montreal, Quebec, Canada.

Kruskal, J., 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society 7, 48–50.

Kwan, R. K., Evans, A. C., Pike, G. B., Nov. 1999. MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans Med Imaging 18 (11), 1085–97.

Laidlaw, D. H., Fleischer, K. W., Barr, A. H., Feb. 1998. Partial-volume bayesian classification of material mixtures in MR volume data using voxel histograms. IEEE Trans Med Imaging 17 (1), 74–86, comment in: IEEE Trans Med Imaging 1998 Dec;17(6):1094-6.

MacDonald, D., Kabani, N., Avis, D., Evans, A. C., Sep. 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3), 340–56.

MNI, 1997. Brainweb – Simulated Brain Database. http://www.bic.mni.mcgill.ca/brainweb/.

Mount, D., Arya, S., 1998. ANN: Library for approximate nearest neighbor searching. http://www.cs.umd.edu/∼mount/ANN/.

O'Rourke, J., 1998. Computational Geometry in C, 2nd Edition. Cambridge University Press.

Paus, T., Zijdenbos, A., Worsley, K., Collins, D. L., Blumenthal, J., Giedd, J. N., Rapoport, J. L., Evans, A. C., Mar. 1999. Structural maturation of neural pathways in children and adolescents: in vivo study. Science 283 (5409), 1908–11.

Pham, D. L., Prince, J. L., Sep. 1999. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Trans Med Imaging 18 (9), 737–52.

Pohl, K., Wells, W., Guimond, A., Kasai, K., Shenton, M., Kikinis, R., Grimson, W., Warfield, S., 2002. Incorporating non-rigid registration into expectation maximization algorithm to segment MR images. In: Dohi, T., Kikinis, R. (Eds.), Proc. of MICCAI 2002. Vol. 2488 of LNCS. Springer-Verlag, pp. 564–571.

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

Rapoport, J. L., Giedd, J. N., Blumenthal, J., Hamburger, S., Jeffries, N., Fernandez, T., Nicolson, R., Bedwell, J., Lenane, M., Zijdenbos, A., Paus, T., Evans, A., Jul. 1999. Progressive cortical change during adolescence in childhood-onset schizophrenia. a longitudinal magnetic resonance imaging study. Arch Gen Psychiatry 56 (7), 649–54.

Schellenberg, J. D., Naylor, W. C., Clarke, L. P., Jun. 1990. Application of artificial neural networks for tissue classification from multispectral magnetic resonance images of the head. In: Proc. of IEEE Symposium on Computer-Based Medical Systems. IEEE, pp. 350–357.

Schroeter, P., Vesin, J. M., Langenberger, T., Meuli, R., Apr. 1998. Robust parameter estimation of intensity distributions for brain magnetic resonance images. IEEE Trans Med Imaging 17 (2), 172–86.

Sled, J. G., Pike, G. B., Aug. 1998. Standing-wave and RF penetration artifacts caused by elliptic geometry: an electrodynamic analysis of MRI. IEEE Trans Med Imaging 17 (4), 653–662.

Sled, J. G., Zijdenbos, A. P., Evans, A. C., Feb. 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17 (1), 87–97.

Smith, S., 2002. Fast robust automated brain extraction. Hum Brain Mapp 17 (3), 143–155.

Stokking, R., Vincken, K. L., Viergever, M. A., Dec. 2000. Automatic morphology-based brain segmentation (MBRASE) from MRI-T1 data. Neuroimage 12 (6), 726–38.

Stone, C., 1977. Consistent nonparametric regression. Annals of Statistics 5, 595–645, (with discussion).

Talairach, J., Tournoux, P., 1988. Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging. Thieme Medical Publishers, New York, NY.

The MathWorks Inc., 1998. Matlab. http://www.mathworks.com.

Udupa, J. K., Samarasekera, S., May 1996. Fuzzy connectedness and object definition: theory, algorithms, and applications in image segmentation. Graphical Models and Image Processing 58(3), 246–61.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., Oct. 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Trans Med Imaging 18 (10), 885–96.

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

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., Dec. 2002. A unifying framework for partial volume segmentation of brain MR images. IEEE Trans Med Imaging 21 (11), 105–119.

Warfield, S. K., Kaus, M., Jolesz, F. A., Kikinis, R., Mar 2000. Adaptive, template moderated, spatially varying statistical classification. Med Image Anal 4 (1), 43–55.

Wells III, W. M., Grimson, W. E. L., Kikinis, R., Jolesz, F. A., Aug. 1996. Adaptive segmentation of MRI data. IEEE Trans Med Imaging 15 (4), 429–442.

Wright, I. C., McGuire, P. K., Poline, J.-B., Travere, J. M., Murray, R. M., Frith, C. D., Frackowiak, R. S. J., Friston, K. J., 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 2, 244–252.

Yan, M. X. H., Karp, J. S., Jun. 1995. An adaptive bayesian approach to three-dimensional MR brain segmentation. In: Bizais, Y., Barillot, C., Paola, R. D. (Eds.), Proc. of Information Processing in Medical Imaging (IPMI). Kluwer, pp. 201–213.

Zijdenbos, A., Forghani, R., Evans, A., Oct. 1998. Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT. In: Wells, W. M., Colchester, A. C. F., Delp, S. (Eds.), Proc. of MICCAI'98. Vol. 1496 of LNCS. Springer-Verlag, pp. 439–448.

Zijdenbos, A. P., Forghani, R., Evans, A. C., Oct. 2002. Automatic 'pipeline' analysis of 3D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imaging 21 (10), 1280–91.

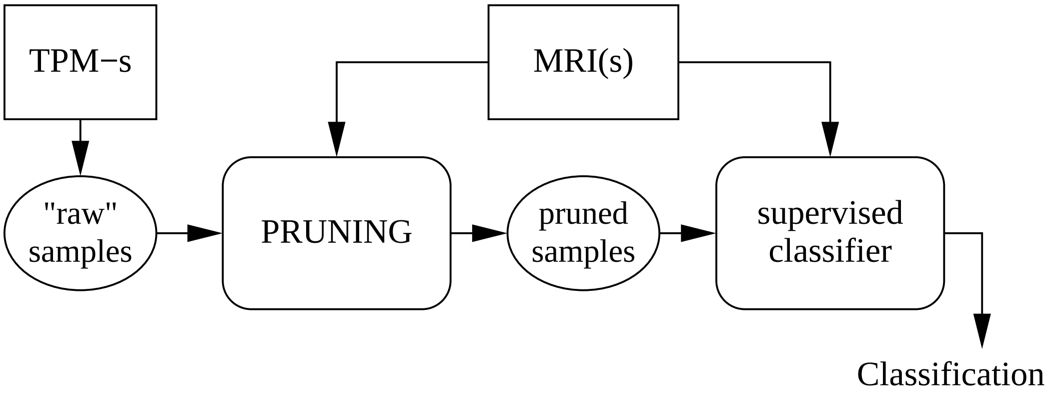

Fig. 3. Diagram of our novel classification procedure described in Section 3. The spatial probabilistic prior (TPM) is only used for generating the raw set of samples; the two-stage classification procedure (pruning and final classification) uses only MR image intensities.

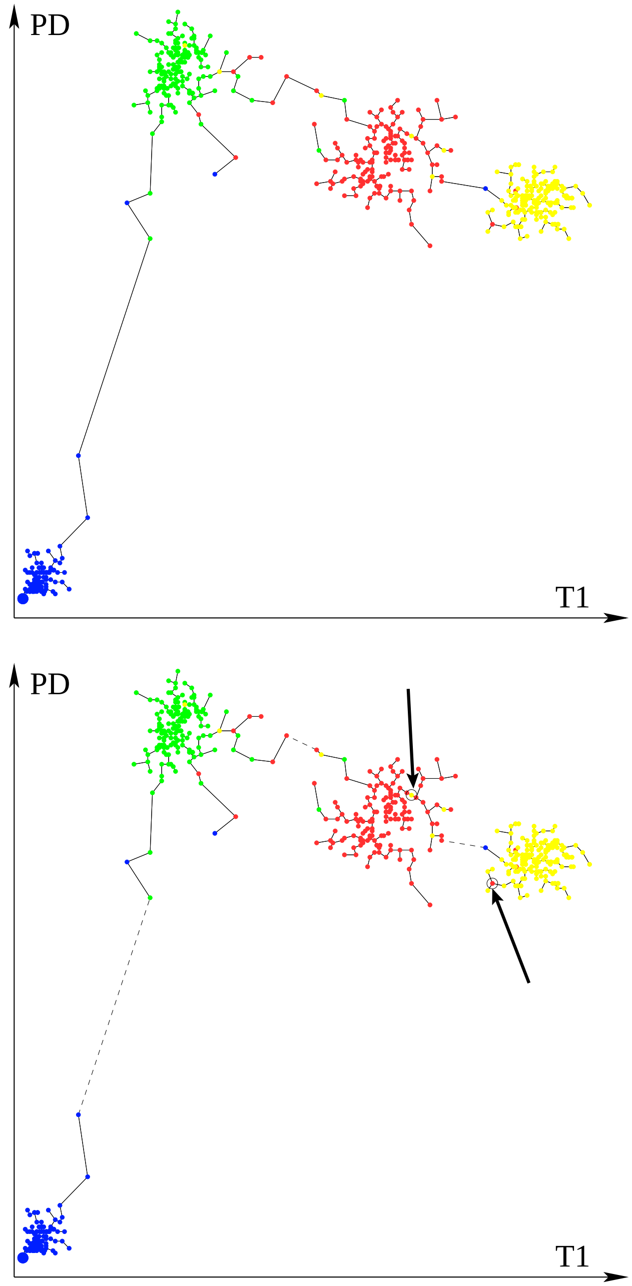

Fig. 4. **Top:** Minimum spanning tree (MST) of a set of points in feature space. The T1-PD data are simulated MRI-s (Section 4.1). The initial labeling of each sample point is indicated by its shade. **Bottom:** Clusters resulted after the iterative graph breaking procedure (Section 3.1) stopped at $R = 4.92$. The arrows indicate points with incorrect initial labeling.

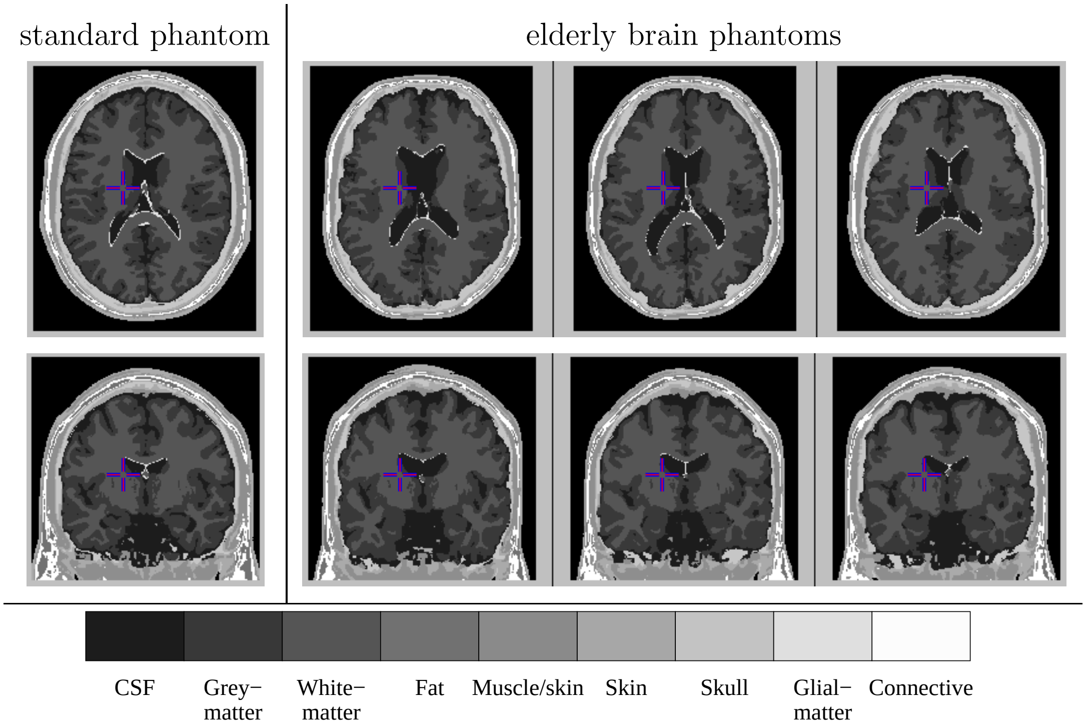

Fig. 5. *(Left to Right:)* Standard phantom (column 1), and three sample “elderly brain” phantoms (columns 2-4), produced as described in Section 4.1. Transverse and coronal sections are shown, but the phantoms are 3D. Compared to the standard phantom (constructed from a young-normal scan), the “elderly” phantoms exhibit enlarged ventricles and overall brain atrophy (typical of aging brains).

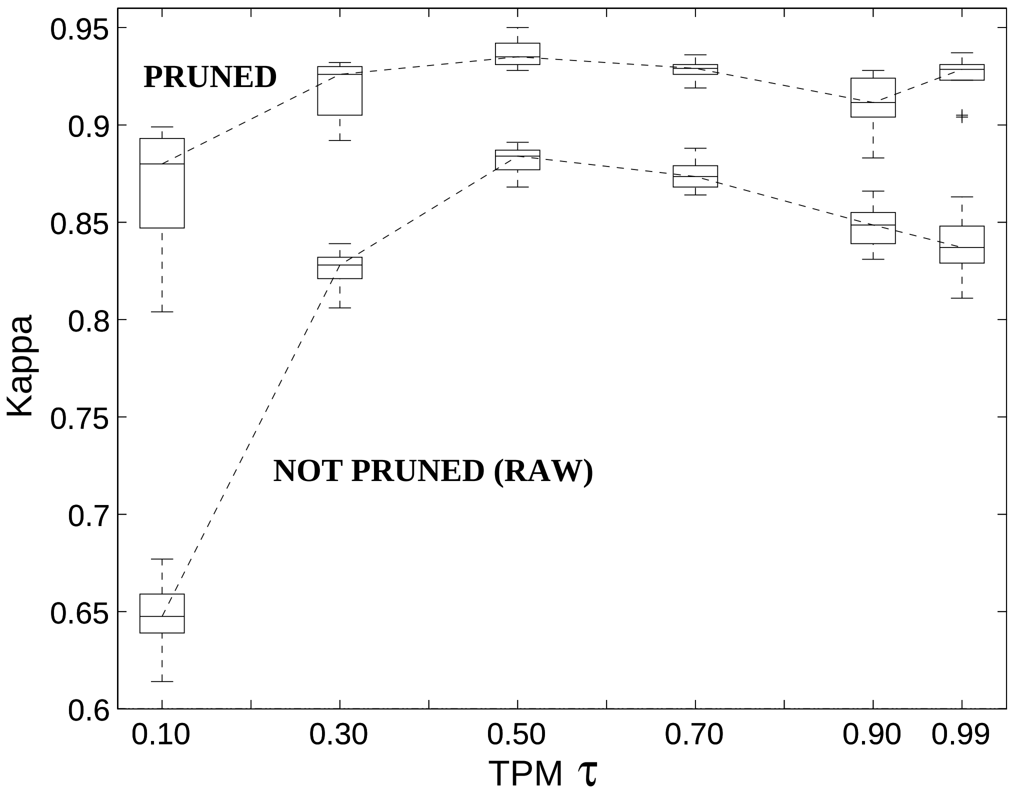

Fig. 6. Quantitative evaluation of classification performance on elderly brain multispectral T1-T2-PD simulated MRI (Section 4.1). $\mathcal{T}$ is the prior probability threshold for extracting the training samples using the TPM. Experiment was repeated with 10 anatomically different elderly brain digital phantoms. Note that for $0.30 \leq \tau < 1.00$ the Kappa of the "pruned" method does not vary significantly with τ. Also, for all $\tau \geq 0.30$ the lowest median Kappa of "pruned" is higher than any median Kappa of "raw", and the difference is statistically significant. (A description of box plots is given in Appendix B.) See also Fig. 7.

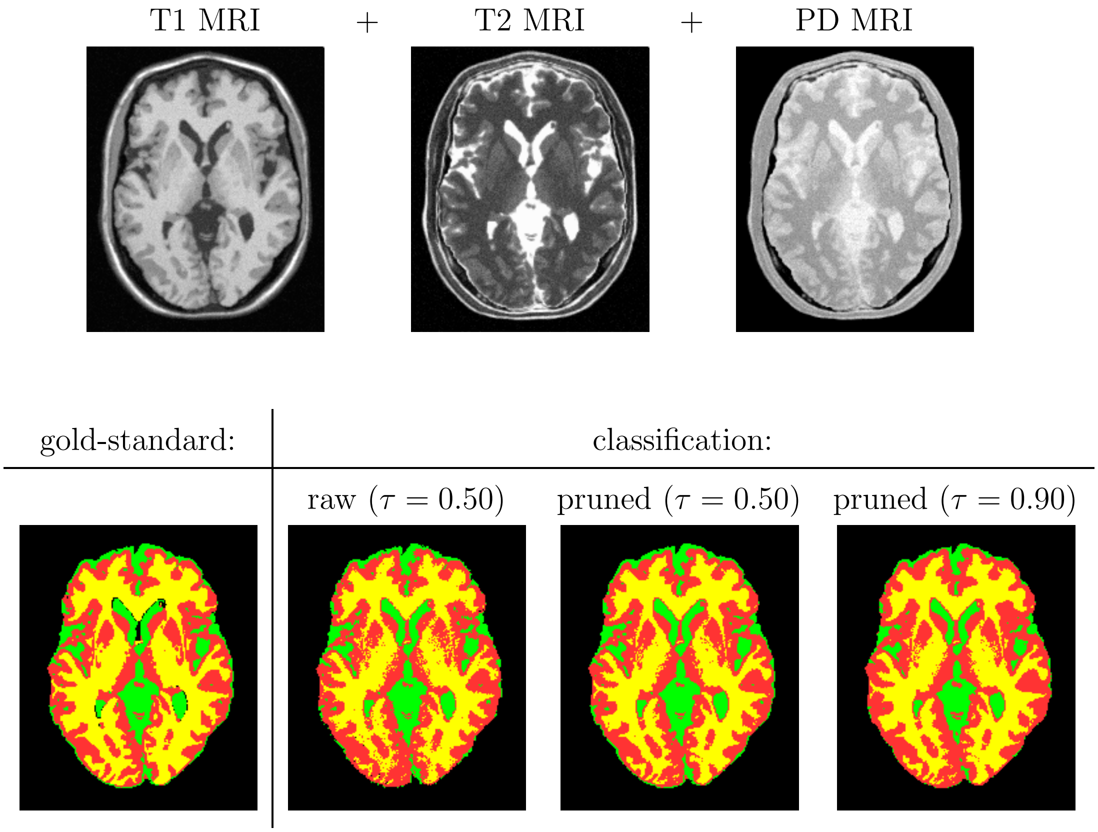

Fig. 7. Gold-standard (digital brain phantom) and classification output for one elderly brain multispectral simulated MRI (Section 4.1). The “pruned” method provided a qualitative improvement in classification compared to the “raw” method – e.g. in the left and right putamen (center of images), and in the left posterior white-matter (lower-left of images). Although only transverse slices are shown, the entire processing was done on the full 3D data volume.

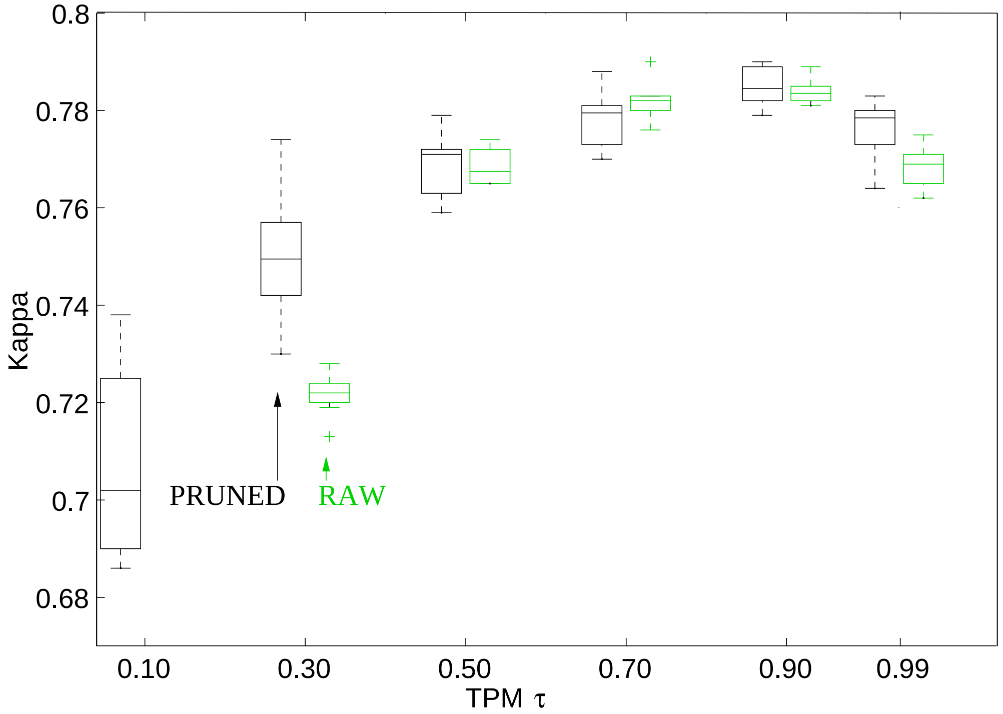

Fig. 8. Quantitative evaluation of classification performance on multispectral (T1-T2-PD) real MRI of a young-normal individual (Section 4.2). Measurements were repeated 10 times with different raw training sets extracted by random sampling of the qualifying spatial locations indicated by the TPM (Section 3.1). The "raw" (no-pruning) result for $\tau = 0.10$ has very low Kappa and is not shown.

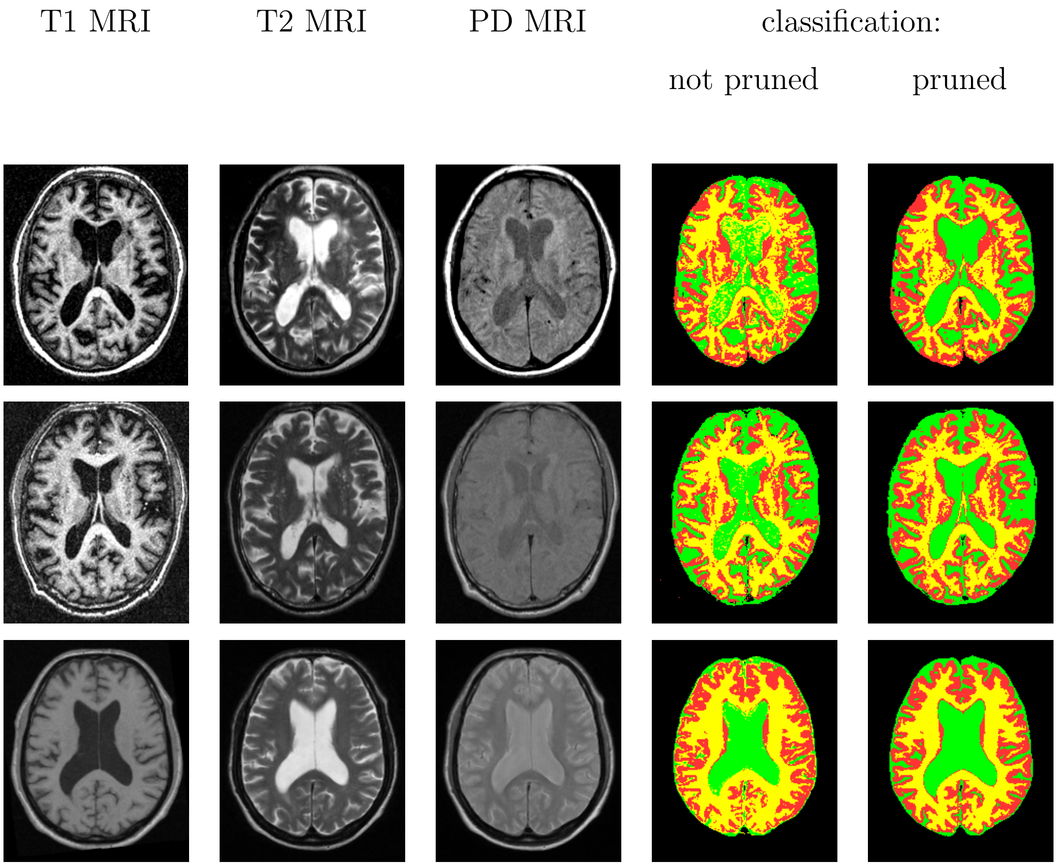

Fig. 9. Classification results on real multi-spectral MRI scans of 3 ischemia patients, shown one per row (Section 4.3). The "raw" (not pruned) classification is poor inside the ventricles (top and middle patients), in the anterior cortex (middle patient, top of the axial slice), and outside the cortex (bottom patient). In comparison, the "pruned" classification is qualitatively improved.

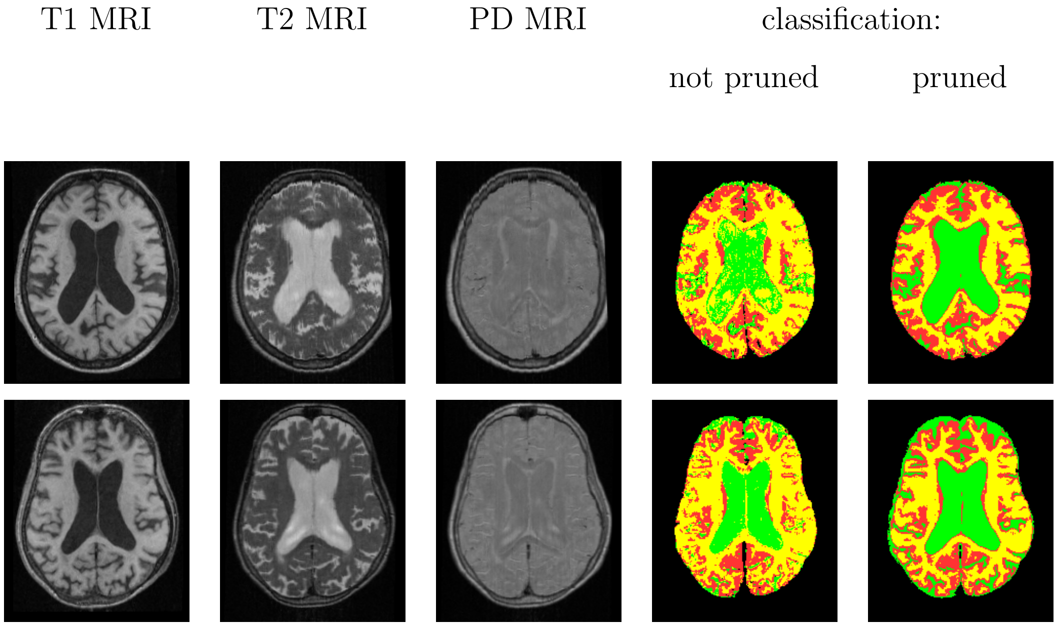

Fig. 10. Classification results on real multi-spectral MRI scans of 2 Alzheimer's Disease elderly patients, shown one per row (Section 4.4). The "raw" (not pruned) classification is poor inside the ventricles (especially for the top patient), and around the lateral cortex – where the white-matter is over-estimated, and grey-matter and external CSF are under-estimated. In comparison, the "pruned" classification is qualitatively improved.

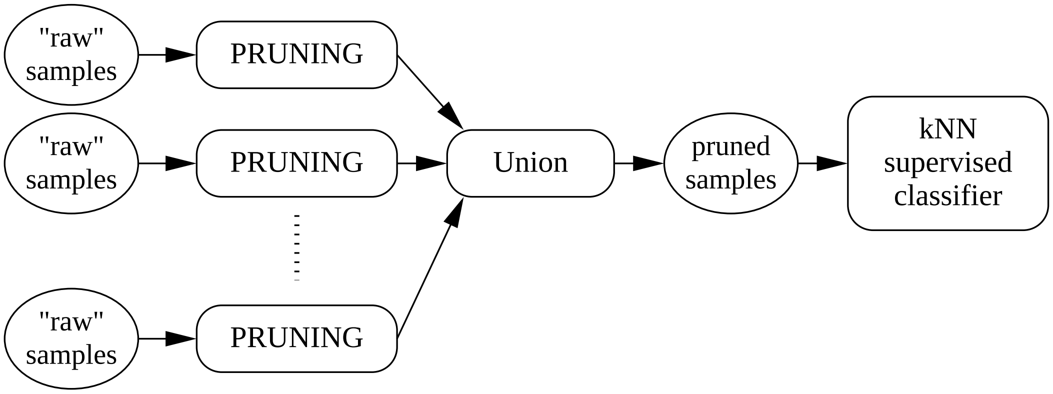

Fig. A.1. Practical implementation (see Appendix A) of the generic classification method from Fig. 3.