Automatic Generation of Training Data for Brain Tissue Classification from MRI

Cristian A. Cocosco

Department of Electrical and Computer Engineering

McGill University, Montréal

April, 2002

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master's of Engineering.

Contents

List of Figures

List of Tables

Abstract

A fully automatic procedure for brain tissue classification from 3D magnetic resonance head images (MRI) is described. The procedure uses feature space proximity measures, and does not make any assumptions about the tissue intensity data distributions. As opposed to existing methods for automatic tissue classification, which are often sensitive to anatomical variability and pathology, the proposed procedure is robust against morphological deviations from the model. A novel method for automatic generation of classifier training samples, using a minimum spanning tree graph-theoretic approach, is proposed in this thesis. Starting from a set of samples generated from prior tissue probability maps (the "model") in a standard, brain-based coordinate system ("stereotaxic space"), the method reduces the fraction of incorrectly labelled samples in this set from 25% down to 2%. The corrected set of samples is then used by a supervised classifier for classifying the entire 3D image. Validation experiments were performed on both real and simulated MRI data; the kappa similarity measure increased from 0.90 to 0.95.

Résumé

Une procédure entièrement automatisée pour la classification de tissus cérébraux à partir d'images de résonance magnétique (IRM) 3D de la tête est décrite. Cette procédure utilise des mesures de proximité spatiale d'élements et ne fait aucune supposition sur les distributions des données d'intensité de tissus. Contrairement aux méthodes de classification automatique de tissus existantes, qui sont souvent sensibles aux variations anatomiques et aux pathologies, la procédure proposée est robuste relativement aux déviations morphologiques du modèle. Une nouvelle méthode pour la génération automatique d'échantillons d'entrainement de classificateur utilisant une approche théorique par arbre à étendue minimum est proposée dans ce mémoire. En se basant sur un ensemble d'échantillons générés à partir de cartes de probabilité de tissus (le "modèle") préalablement connues et exprimés dans un système de coordonnées standard lié au cerveau ("espace stéréotaxique"), la méthode réduit la fraction d'échantillons identifiés incorrectement de 25% à 2%. Cet ensemble d'échantillons corrigé est ensuite utilisé par un classificateur supervisé pour classifier toute l'image 3D. Des expériences de validation ont été effectuées autant sur des données IRM réelles que simulées; le coefficient de similarité Kappa a augmenté de 0.9 à 0.95.

Acknowledgments

Above all, I would like to thank my supervisor Alan C. Evans for guiding and supporting this work, and Alex P. Zijdenbos for co-supervision and guidance on this project. In addition, I am grateful to John Sled, Steve Robbins, and Peter Neelin for the prompt and thorough feedback, corrections, and valuable suggestions on the first draft of this thesis. Also, I would like to thank: Godfried Toussaint for his inspiring Pattern Recognition course, Noor Kabani for the full brain manual segmentation, Louis Collins for the non-linear registration software, Marguerite Wieckowska for the French version of the abstract, and Steve Robbins for proofreading. Last but not least, I would like to thank all the past and present "`geeks@BIC`" for the great computing environment at the McConnell Brain Imaging Centre, MNI — it really made this project possible.

Chapter 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. Operator-assisted segmentation (classification) methods are impractical for large amounts of data, and also are non-reproducible. Existing methods for fully automatic brain tissue classification typically rely on an existing anatomical model. This 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 completely different human population than the image to be classified.

This thesis 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.

Chapter 2

Brain Tissue Classification from MRI

Magnetic resonance imaging (MRI), also referred to as nuclear magnetic resonance (NMR), is a powerful and flexible medical imaging modality. Among many other capabilities, it can produce high-resolution images with good contrast of the different biological soft tissue types [32]. As a non-invasive technique, it is widely used in the clinical and research environments for imaging both anatomy and function.

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 thesis is the classification, or labeling, of individual voxels of a 3D anatomical MR image (aMRI) as one of the main tissue classes in the brain: cerebro-spinal fluid (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). A feature of MRI is that, by using different pulse sequences, different contrasts between tissue types (multi-spectral image data of the same subject) can be easily obtained.

An accurate and robust tissue classification is the basis for many applications such as: quantitative measurements of tissue volume in normal and diseased popu-

lations [19], morphological analysis (for example, of cortex folding patterns), or visualization. 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 [77]. Fully automatic, robust tissue classification is required for batch processing the data from large-scale, multi-site clinical trials or research projects (such as [77]).

2.1 AMRI artifacts

Acquisition artifacts in the MR images can be a significant challenge for automated tissue classification. The main artifacts affecting brain anatomical MRI (aMRI) scans are:

Intensity non-uniformity (INU) Also known as shading artifact or bias field, it is inherent to MRI. It was recently shown [60, 61] that the largest contributor to INU is the electromagnetic field inhomogeneity which is dependent on the particular shape of the subject being scanned. An extensive review of INU correction methods is given in [59, 63]; the correction (using post-acquisition image processing techniques) can be done separately [63], or in conjunction with image segmentation [5, 6, 31, 35, 54, 70, 75, 76].

Noise: The MRI noise is Rician distributed, and uncorrelated between voxels. The image acquisition exhibits a tradeoff between signal-to-noise ratio (SNR) on one side, and spatial image resolution and scan time on the other side – a higher SNR can be obtained by lowering the resolution, or by increasing the scan time. Noise tends to be 2-5% of the maximum signal intensity when using a modern MR scanner for a standard (in-vivo) human brain anatomical scan. MRI noise can be reduced by post-acquisition image processing techniques such as anisotropic filtering [30], which is an edge preserving smoothing

operation.

Partial volume: Technically it is not an "artifact" but merely a consequence of the finite resolution of the imaging process – the image voxels (typically about 1 mm^3 or larger) may contain a mixture of more than one tissue type, which all contribute to the measured signal. Since increasing the resolution is not always practical (longer scan time, increased noise), an alternate solution is to attempt to recover the mixing fractions in the image segmentation (classification) method [11, 47, 54, 72].

More information about MRI and its artifacts can be found in reference texts, such as [32].

2.2 Validation methods

Brain tissue classification (or segmentation) methods can be tested and evaluated on real MRI data, or on realistic simulations of MRI acquisitions. The evaluation can be:

qualitative: results are manually inspected and compared with knowledge of brain typical anatomy.

quantitative: some similarity measure is (automatically) computed between the classification result and a reference "gold standard" classification.

The MR image data used in this work for quantitative measurements is described below.

2.2.1 Simulations

These data were produced using a sophisticated MRI simulator [44, 45], using as input a realistic anatomical model ("phantom") [21]. The standard phantom is

based on a real scan of a young (30 year old) normal male. This simulated data is identical to what is publicly available through the BrainWeb Internet interface [1, 14].

It should be mentioned that these (BrainWeb) simulations assume homogeneous tissue MR properties throughout the brain. In practice, this is not the case [10, 41, 50]. However, these simulations provide a realistic approximation for testing MRI classification methods.

Moreover, partial volume (which reduces the cluster separation in feature space) is a challenge for any classification method. A simple classifier was used for estimating the tissue fractions in each voxel (partial volume) when creating the anatomical model used here [21]; a more sophisticated continuous classifier (section 2.3) may provide a better estimate.

One advantage of using simulated data is that the "answer" (the "gold standard") for the tissue classification procedure is known – it is the anatomical model ("phantom") that was used to produce the simulations. This allows the computation of accurate quantitative measures of performance. Other advantages of simulations are convenience, flexibility, and low cost.

Another advantage is in the case of multi-spectral MR data. With a real scanner, different MRI contrasts (such as T1- and T2-weighted scans) are acquired separately. The scans subsequently need to be spatially registered (aligned) to each other before the automatic classification, and the registration procedure can introduce alignment errors. Comparatively, simulated multi-modality image data is perfectly registered (as long as the same phantom is used for all modalities).

2.2.2 Real data, young individual

A real multi-spectral MRI scan of a 36 year old normal male was the basis of many of the measurements and experiments presented in this thesis. The T1-weighted, 1mm isotropic voxel, scan was completely manually segmented [37–39] by a human

expert – a trained neuroanatomist. T2 and PD scans were also acquired as 2mm thick sagittal slices; both acquisitions were repeated a second time, with a 1mm offset; the two paired scans were co-registered and averaged together in order to improve the image resolution.

All three modalities (shown in Figure 2.1) were registered (and re-sampled) to a stereotaxic space [20, 67] using linear registration software [20]; also, INU correction was performed using N3 [63].

For this dataset, the "gold standard" was the manual classification; this has several limitations. First, the human expert had the advantage of also having previous anatomical knowledge in addition to just the MR signal intensities. Second, the expert only performed the classification on the T1 data (some tissues, or CSF and air spaces, are difficult to distinguish on a T1). Lastly, this classification is discrete, not continuous (without any partial volume). Moreover, it is generally recognized that automatic classifications yield more consistent and reproducible region boundaries than manual classifications do.

This dataset presents additional challenges for automatic classification when all modalities are used together: the T2 and PD data are of a lower true resolution (1x1x2 mm, while the simulations were done at 1x1x1 mm), and there is a possibility of registration imperfections between the 3 scans.

2.2.3 Quantitative similarity measure

The terminology associated with a generic classification result is given in Table 2.1. This thesis uses the *Kappa* measure, which is a chance-corrected similarity measure between two labelings, originally proposed by Cohen [15]. This measure was also used by other researchers [5, 6, 42, 77] for quantitative measurements of brain MRI classification performance.

For a C-class classification problem, if ($\forall$ class i) we denote by a_i the number of true positives, by c_i the number of samples classified as i, and by t_i the true number

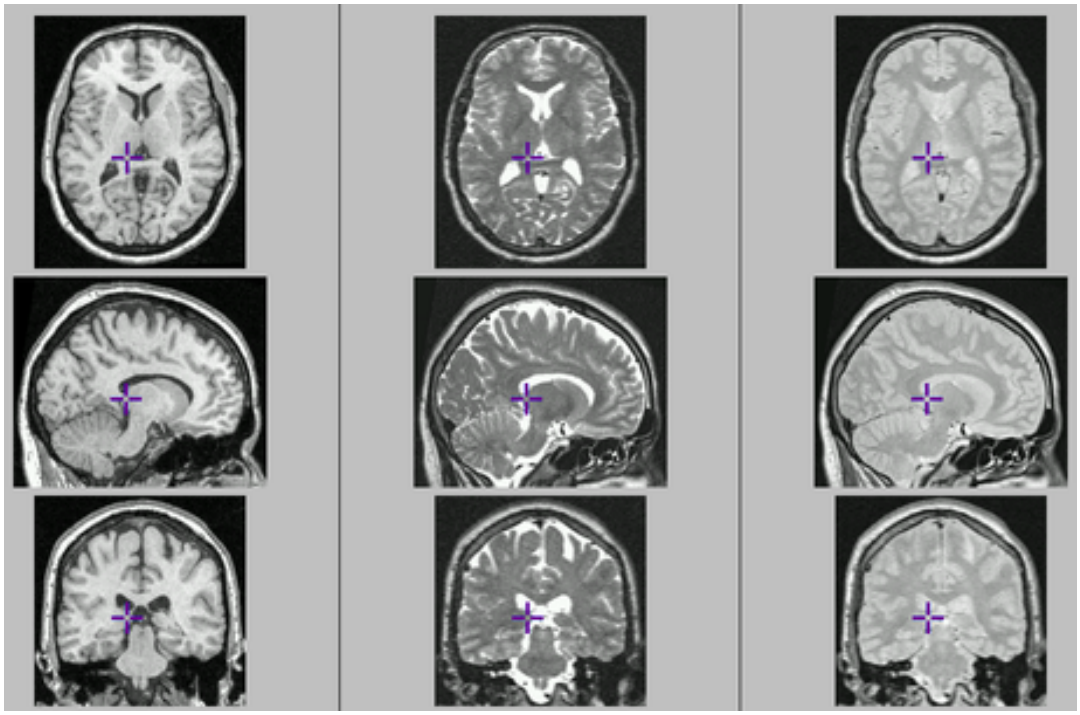


Figure 2.1: The "young normal" real multi-spectral MRI dataset (section 2.2.2). Left to right: T1, T2, PD. Top to bottom: transverse, sagittal, coronal 2D slices through the image volumes.

of samples in class i, and by N the total number of samples (in all classes), then kappa is defined as:

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

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:

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

The following should be pointed out regarding kappa:

- it treats all voxel labeling differences (and similarities) the same, regardless of where in the image the voxel is.
- its value is a relative and not an absolute one; that is, it is only meaningful in a comparison setting.

2.3 Review of the state of the art

An overview of medical imaging low-level segmentation methods is given in [23], and reviews of MRI segmentation are given in [8, 12, 78]. Segmentation methods can be image intensity-based (classification, region-based methods), or edge-based methods. Both have their limitations: the former are affected by any overlaps (between tissue types) in the tissue intensity distributions; for the latter, highly curved interfaces between tissues (such as the gray-white matter interface in the brain) can be challenging to recover from finite resolution images.

Classification methods operate in a multi-dimensional feature space. 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 [28], and some selected topics are also given in [26, 27]. Many researchers have applied to brain MRI classic methods such as the Bayes (maximum likelihood) classifier [19, 40], or non-parametric classifiers like kNN (k nearest neighbors) [74] and ANN (artificial neural network) [77].

Expectation-Maximization (EM) is a popular statistical classification scheme for this segmentation application; originally proposed (in a brain MRI context) by Wells [75], and further improved by others [5, 6, 31, 35, 58, 70, 71], these methods interleave INU field estimation (correction) and classification, in an iterative fashion. Some authors [31, 62] have suggested that Wells' original method [75] is very sensitive to the training samples, and also that high partial volume regions and outliers can confuse the bias field estimation. Moreover, the early methods [31, 35, 58, 75] need to be supplied with the class intensity distribution parameters (which are usually generated by manually selecting representative voxels), and these parameters are kept fixed during the operation. In order to address this issue, Van Leemput [70, 71] and Ashburner [5, 6] re-estimate the distribution parameters at each iterative step; also, both methods use a probabilistic brain atlas to provide the initial values of these parameters.

Any overlaps in the class feature space (intensity) distributions will lead to a certain amount of misclassifications, generally at the interface between tissues (where partial volume is also present). Solutions that attempt to reduce this problem are post-classification morphological operations, and contextual classifiers – such as "relaxation" optimization procedures using Markov random field (MRF) models [11, 35, 55, 76]. Udupa proposed a "fuzzy connectedness" method [69] which measures the strength of connectness between two voxels as a function of both their relative spatial location and the intensity similarity between them.

A method that combines the strengths of intensity (within region) information and edge information, in a multi-scale approach, was recently proposed by

Niessen [52]; however, the method requires some user assistance, hence is not fully automatic.

Besides the traditional discrete classifiers (that output a discrete class label for each voxel), there are also classifiers that output a fuzzy (continuous) class membership value. This can be considered as an estimate of the mixing proportions of several tissues in a voxel (the effect of partial volume). A number of researchers have proposed continuous classifiers (all modeling a mixture of multi-variate Normal intensity distributions): Choi [11] (the so called "mixel" model), Laidlaw [46–48] (using a Bayes classifier), Schroeter [58] (an improved EM algorithm, and also a genetic algorithm), and Pham [54] (an adaptive fuzzy c-means method that also estimates the INU field). An improvement was recently proposed by Van Leemput [72], who also suggested that previous methods may perform poorly in practical situations.

Another approach, typically used for segmenting specific anatomical structures, is to use higher-level information such as deformable models or atlases, or prior shape knowledge [64]. However, such methods can have difficulty with abnormal anatomies (or highly variable ones, like the human cortex).

An interesting method that combines classification and deformable models was proposed by Warfield [73, 74]. This method iteratively interleaves classification and non-linear registration; the typical limitations of classification are addressed by constraining it with the (deformed) anatomical template.

An aspect that is ignored by most brain MRI classification schemes is how to *fully automatically* produce a *correct* set of training samples for the classifier, when given a never-seen-before MRI brain dataset, possibly originating from a new site and MR scanner. Nevertheless, some researchers have attempted this:

- Harris [34] extracts samples from small brain areas with low intensity variance, then clusters them in feature space and rejects outliers using various heuristics. The method requires "typical" cluster mean and variance values; however,

these values are scanner, site, and even pulse sequence, specific.

- The use of stereotaxic space tissue probability maps (TPM-s) for automating supervised classification algorithms was originally proposed by Kamber [40], and subsequently used by other researchers [42, 77]. The TPM is used to select training samples from spatial locations that are very likely to contain a given tissue type. However, this approach is very sensitive to any deviations of the subject's anatomy from the statistical model represented by the TPM (more details about this approach will be given in section 3.1).

- Van Leemput [70, 71] uses a probabilistic brain atlas (a TPM) to initialize, and also constrain, their EM-style scheme. However, the author reports that the method fails on atypical (significantly different than the atlas) brain scans, such as child brains, or brains with large pathological abnormalities.

- A scheme very similar to the latter is used by Ashburner [5, 6]. Although a blurring of the prior probability map (with a 8mm FWHM Gaussian kernel) is performed in order to increase robustness, similar limitations are reported for pathological brains.

Moreover, most of the feature-space classification methods mentioned in this section (exceptions are [69, 74, 77]) assume multi-variate Normal (Gaussian) class intensity distributions – and, as will be argued in section 3.2, this is a questionable assumption.

Table 2.1: Agreement between a classification and truth ("gold standard"), for a generic two-class (binary) problem. For example, if a classification labels a sample as belonging to a certain class, then this answer is a "true positive" if it is correct, and a "false positive" if it is not. A C-class problem ($C > 2$) can be seen as a binary problem for each class i ("yes" corresponds to class i, "no" corresponds to the union of the other $C - 1$ classes); in other words, correctly labeled samples are "true positives", and incorrectly labeled samples are "false positives".

		Truth	
		Yes	No
Classification	Yes	true positive (TP)	false positive (FP)
	No	false negative (FN)	true negative (TN)

Chapter 3

Problem Statement

The contribution of this thesis 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 stereotaxic space (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).

The end of the previous Chapter reviewed existing methods for fully automatic brain tissue classification. Their limitations, as well as how these limitations are addressed by this thesis, are discussed in more detail below.

3.1 Model-based training set selection

The MRI intensity scale has no absolute meaning, and is dependent on the pulse sequence, and other variable scanner parameters. Thus, the ability of a tissue classification method to automatically adapt to a new MRI dataset is especially important for data collected from multiple sites. Most of the previously published classification methods (section 2.3) do not address this issue.

A stereotaxic space tissue probability map (TPM) of a given tissue is a spatial probability distribution representing a certain subject population. For each spatial location (voxel) in stereotaxic space, the TPM value at that location is the probability of the given tissue being observed there, for that particular population.

Once imaging data is spatially registered to a stereotaxic space, TPM-s provide an a-priori spatial probability distribution for each tissue (Figure 3.1). This distribution can be used to automatically produce a training set for the supervised classifier [40]: for example, choose spatial locations that have a TPM value $\geq \tau = 0.99$ (99%). The lower the τ, the more qualifying spatial locations there will be, as shown in Figure 3.2. However, this simplistic approach has limitations:

False positives: Since the morphology of the human brain is so complex and individual-specific, even among the locations with very high a-priori probability of being a given tissue, some of them will be "false positives" (that is, wrongly labeled as one tissue class, when in fact they are from another class). It was determined experimentally (details forthcoming in section 5.2) that for a subject drawn from the same population used to create the TPM-s, and for $\tau = 0.99$, the false positives amount to about 3% of all selected locations.

However, this fraction will be larger when the subject is from a different population (in a statistical sense) than the population represented by the TPM. This is illustrated by Figure 3.3: there is a visible morphological difference between the young and elderly groups of subjects[1]. For example, the enlarged ventricles (filled with CSF) will cause some spatial locations with high a-priori probability of being grey- or white-matter (according to the young-population TPM) to actually correspond to CSF in the elderly subject's brain.

Intensity distribution estimation: As seen in Figure 3.2, at $\tau = 0.99$ (where the false positive rate is lowest) the qualifying sample points give a very limited

[1]MRI data source: Dr. Ryuta Kawashima, Sendai, Japan.

coverage of the brain area (especially for CSF). Intuitively, this will not give a good estimate of the true tissue intensity distributions desired for the final supervised classification stage, for two reasons:

- Brain tissue is not homogeneous throughout the brain [10, 41, 50].
- MRI artifacts, such as INU, introduce additional spatial variations in the measured tissue signal.

Thus, sampling at a lower τ would be beneficial for the intensity distribution estimation; however, a lower τ also means more false positives.

The contribution of this thesis is a way to address these two limitations (of the simple TPM thresholding approach). Specifically, a "pruning" of the raw set of points obtained from the TPM is performed, with the goals of eliminating the false positives caused by anatomical difference, and of allowing for a lower TPM τ. Intuitively, the pruning should improve the tissue classification for subject morphologies which are "different" than the ones used to produce the TPM. Also it should give a better estimate of the tissue intensity distributions by allowing sampling at a lower TPM τ; the better estimate should ultimately help the final supervised classifier determine better decision boundaries.

From Figures 3.4 and 3.5, it can be observed that: 1. the densities estimated using $\tau = 0.51+$ pruning resemble the "truth" (manual segmentation) better than the ones estimated without pruning, and 2. the pruning is more important at $\tau = 0.51$ (the no-pruning distributions have severe overlap, due to the higher rate of false positives). The unusual shape of the left-most distribution (CSF) is likely because of the limitations of the manual segmentation procedure (see page 6).

The TPM-s used in this work (shown in Figure 3.1) were already available at the Montreal Neurological Institute (MNI) [40, 42]. They were produced as follows:

1. A set of 12 individual scans were registered to the Talairach stereotaxic space [20, 67], and then classified using a manually trained supervised classifier [40].

Then, for each of the three main tissue classes, the TPM value at each voxel (spatial location) was computed as the fraction of the 12 scans that had that voxel classified as that tissue. This was the first generation TPM.

2. A set of T1/T2/PD MRI scans of 53 young individuals (aged 18 to 35) were used to produce the second generation TPM-s. The supervised classifier was trained using the first generation TPM-s together with some manual assistance [42]. The resulting (second generation) TPM-s were the ones used in this thesis.

The final result of such a process is influenced by how the registration is done, and by the particular tissue classification method. In the above, the registration used a linear (rigid) transformation with 9 degrees of freedom, and the classification employed an artificial neural network (ANN) classifier.

Nevertheless, the particular TPM used is not a critical factor. The TPM is just an initial guess for the pruning algorithm; the only requirement is that the majority of training points it provides, for a given τ, are correctly classified ("true positives"). However, the more similar the morphology of the subject is to the average of the population represented by the TPM, the better the entire classification procedure should work.

3.2 Feature space data distributions

Non-parametric classifiers (supervised or not) are attractive because they don't make any assumptions about the underlying (feature space) data density functions. Several popular classifiers, such as the Bayes (maximum likelihood) statistical classifier, the k-means (c-means) classifier, and the minimum-distance classifier, are parametric classifiers – they assume the data distributions in feature space follow a certain model. Typically, the multi-variate Gaussian model ("Normal" distribution)

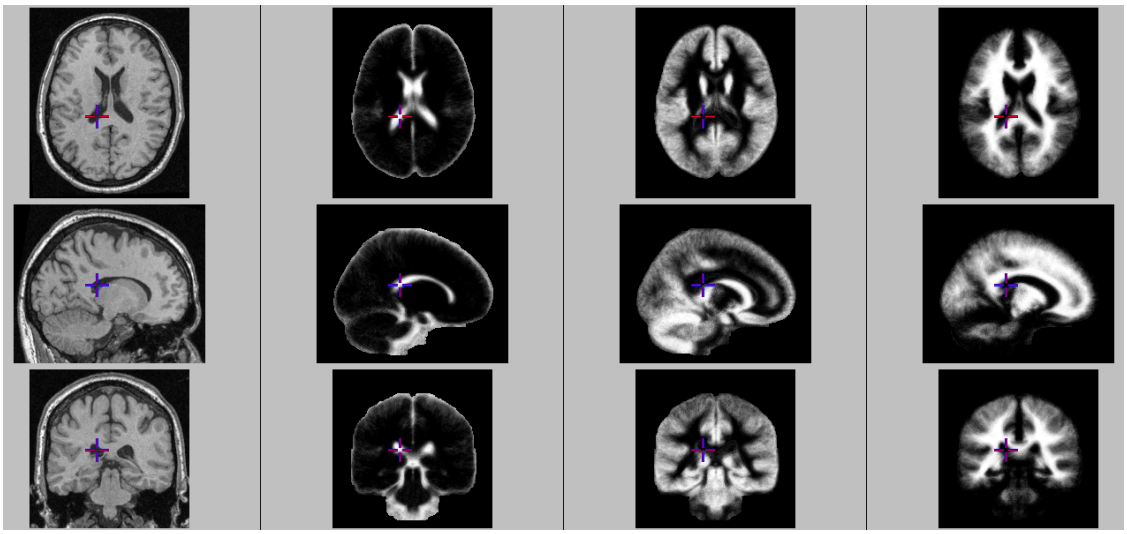


Figure 3.1: [left to right] A T1 MRI scan, and prior tissue probability maps (TPM-s) for CSF, grey matter, and white matter. All 3D images are registered to the same stereotaxic space.

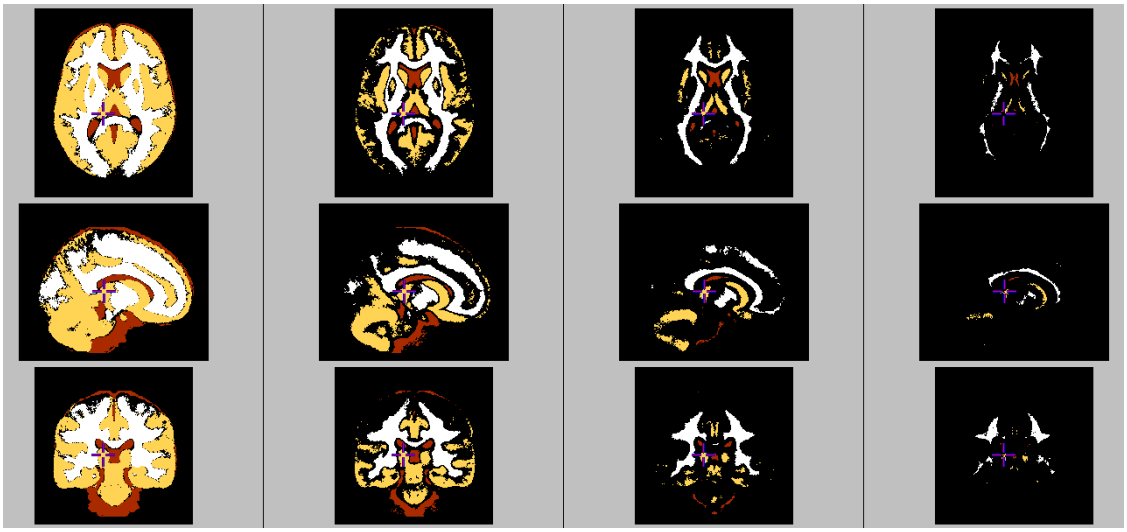


Figure 3.2: Spatial locations with prior tissue probability (TPM value) $\geq \tau$, where τ is (left to right): 0.50, 0.70, 0.90, 0.99. The three classes (CSF, grey matter, white matter) are represented as different gray shades (CSF is darkest, white matter is brightest).

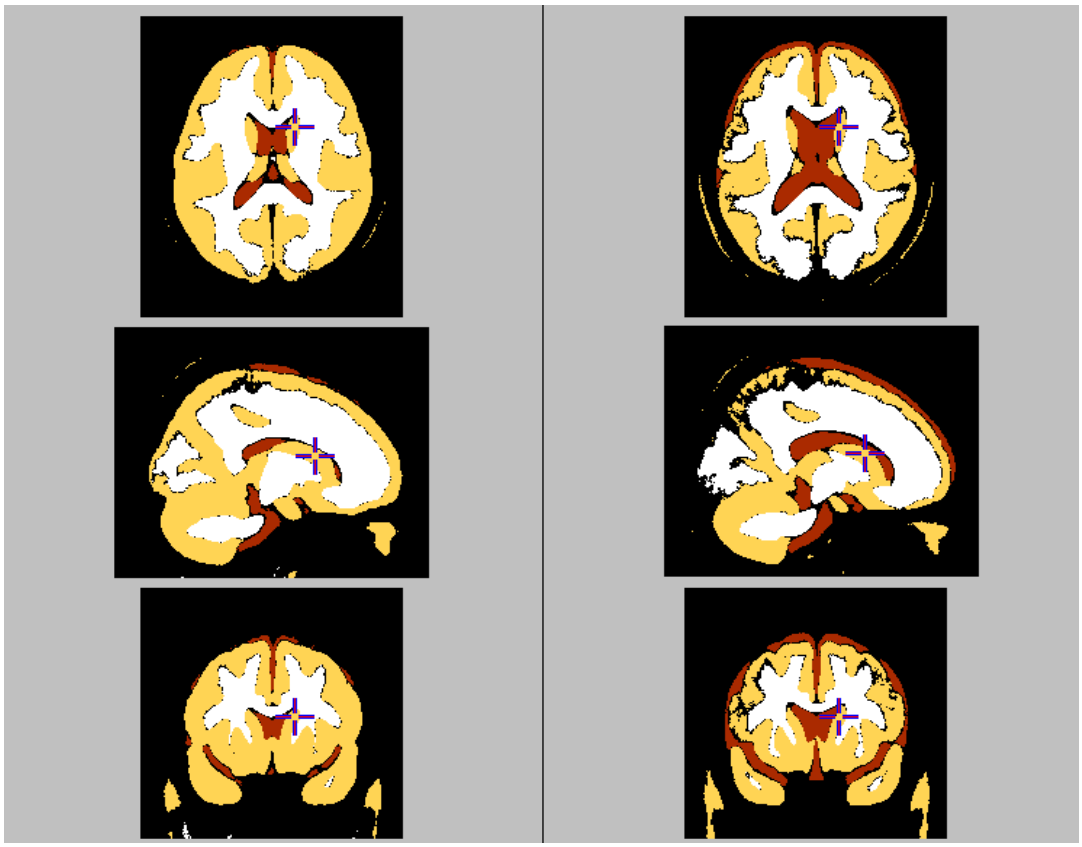


Figure 3.3: Average tissue classification results: the images show the spatial locations that were classified as a particular tissue in more than half of the scans in the given group. Left: group of young healthy subjects (18-35 years old). Right: group of elderly healthy subjects (60-80 years old), from the same population as the young ones. Note the enlarged ventricles, and the general increased atrophy of the brain, caused by natural aging. (CSF, grey matter, and white matter are shown as shades of grey of increasing brightness.)

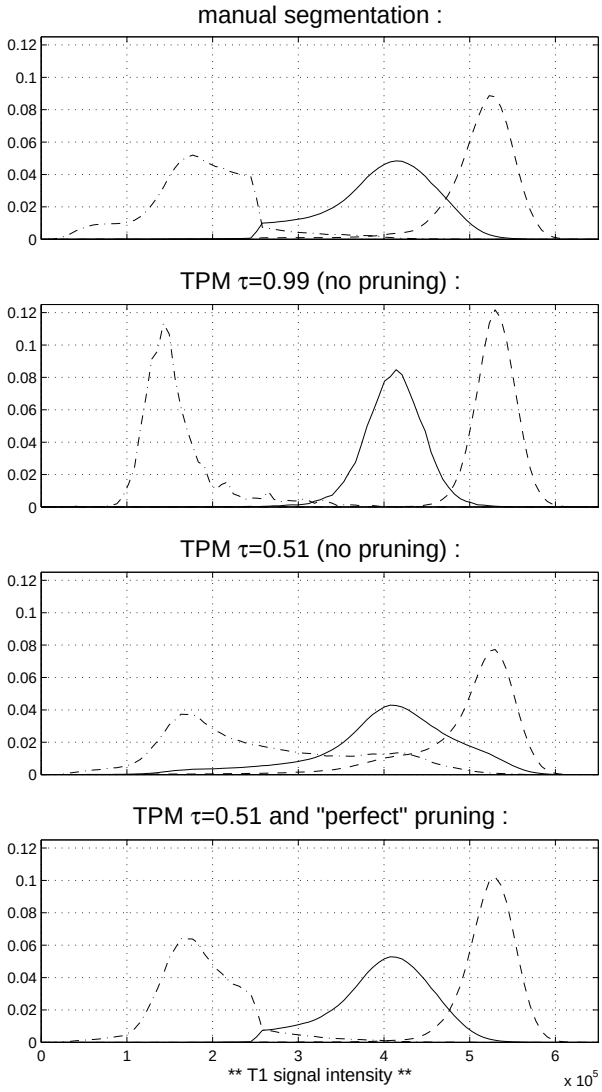


Figure 3.4: Single-dimensional feature space (T1 MRI intensity) probability densities for the three tissue classes (*left to right*: CSF, grey matter, white matter), estimated from the real dataset (section 2.2.2) using (*top to bottom*): the manual segmentation, the TPM thresholded at $\tau = 0.99$, the TPM at $\tau = 0.51$, and the latter followed by a "(perfect) pruning". The pruning removed the false positive data points, as indicated by the manual segmentation (considered "truth"). Note that the densities obtained with $\tau = 0.51+$ pruning resemble the manual segmentation significantly better than the ones for no-pruning.

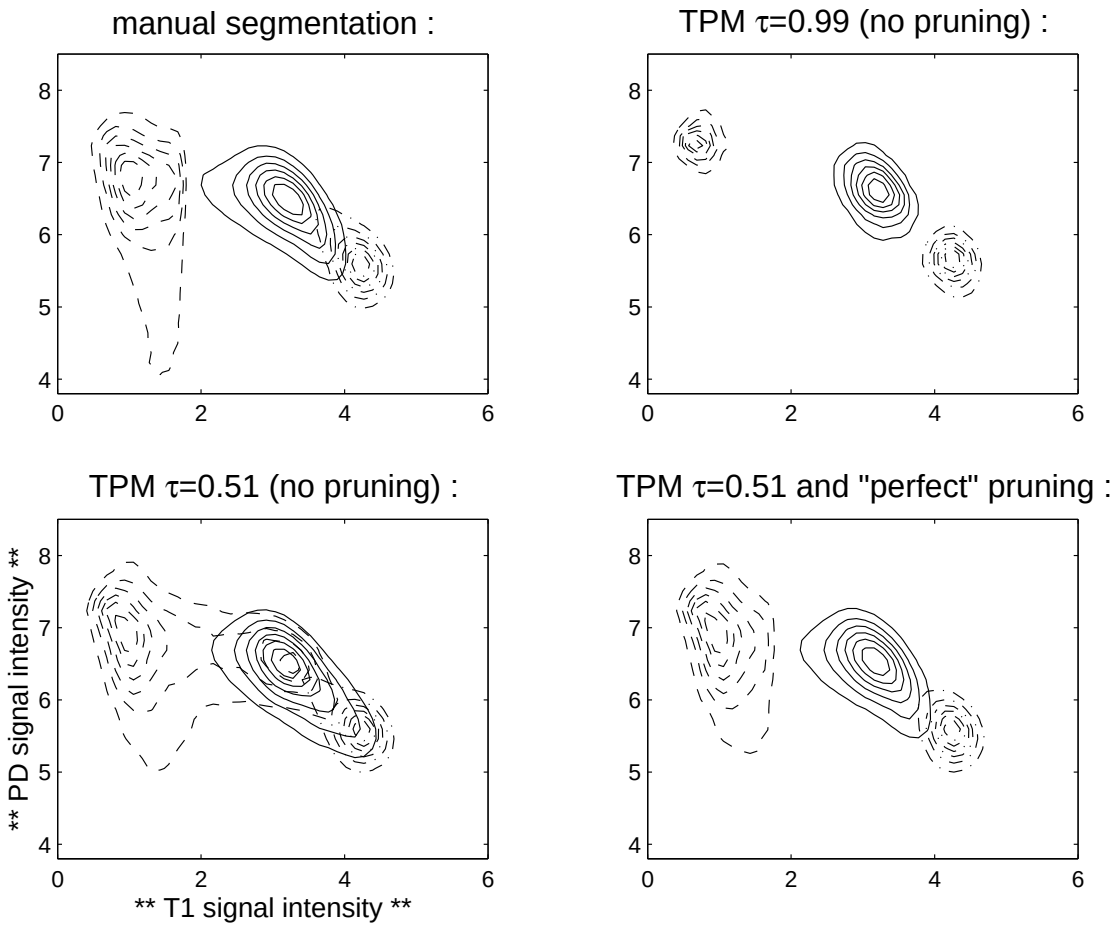


Figure 3.5: Two-dimensional feature space (T1 and PD MRI intensities) probability densities for the three tissue classes (*left to right*: CSF, grey matter, white matter), represented as iso-contours (contours correspond to equally spaced values between 0 and each cluster's maximum). The densities were estimated from the real dataset (section 2.2.2) using: the manual segmentation, the TPM thresholded at $\tau = 0.99$, the TPM at $\tau = 0.51$, and the latter followed by a "(perfect) pruning" (the pruning removed the false positive data points, as indicated by the manual segmentation, considered "truth"). Note that the densities obtained with $\tau = 0.51+$ pruning resemble the manual segmentation significantly better than the ones for no-pruning.

is used (see section 2.3).

If the features are MR signal intensities from various MRI modalities (such as T1, T2, PD), then the Gaussian model assumption can be poor. Other researchers have also suggested that MRI tissue intensity distributions are not Normal [13, 24, 57]. Besides biological causes such as the intrinsic heterogeneity within the tissue classes that concern this work (CSF, grey matter, white matter), the MRI acquisition artifacts also affect the intensity distributions [5, 42, 57]. In order to experimentally study their effect on feature space distributions, several artifacts (see section 2.1) were artificially added to a real MRI multi-spectral dataset: additional Rician distributed noise, additional multiplicative INU field (estimated from real MR data), and increased partial volume effect corresponding to thicker acquisition slices (simulated by blurring with a 1-dimensional box smoothing kernel). The results are presented in Figure 3.6. The INU artifact, and the partial volume effect (due to low spatial resolution) noticeably make the clusters deviate from the Normal shape.

A disadvantage of non-parametric classifiers is that they tend to require a larger training set for obtaining good performance; also, in general these classifiers are more computationally expensive. While this was likely the reason why they were not used more widely in the past, the computational demands are becoming less of an issue as computing power steadily increases.

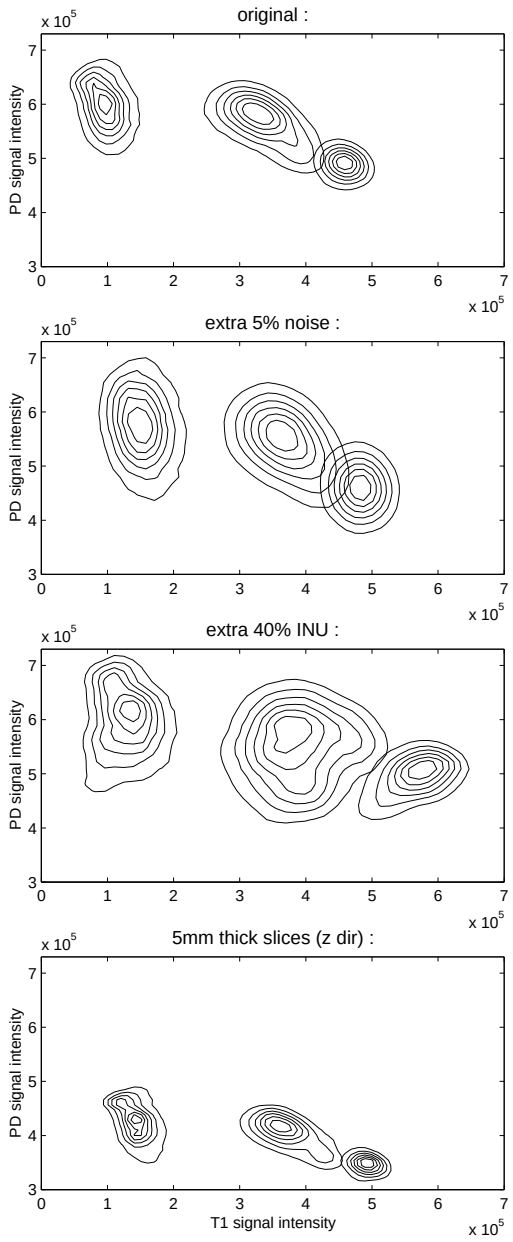


Figure 3.6: Effect of acquisition artifacts on the tissue probability densities in feature space (section 3.2). Densities (represented as iso-contours at equally spaced values between 0 and the cluster's maximum) were estimated from real T1+PD MRI data using a full-brain manual classification (see page 5), eroded once to reduce the initial partial volume. The INU and the partial volume (thick slices) noticeably make the clusters further deviate from the Normal shape.

Chapter 4

Method

The following presents a fully automatic, non-parametric, brain tissue classification procedure based on feature space proximity measures. It consists of two stages:

1. A 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.

2. A supervised classifier, using the k nearest neighbor (kNN) algorithm. It is trained on the set of samples produced by the first stage.

The main contribution of this thesis is stage 1, which will be referred to as the "pruning" stage.

The pruning works on a set of input sample points that are selected (through random sampling) from the qualifying 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 $\geq \tau$, where τ is the threshold parameter.

Both the pruning and the final classification are done in feature space. The features used in this work are only MRI signal intensities (image gray levels), but in general other features can be added – such as local gradient measures, spatial location information, various moments computed on a neighborhood centered at the voxel, and so on. The use of the TPM-s is a way of including prior anatomical knowledge in the intensity-based classification.

The feature space proximity measure used in this work is a distance metric – the common Euclidean distance (in d-dimensional space):

$$D(a,b) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}$$

However, any other distance metric can be used[1]. A *metric* distance measure must satisfy all of the following four conditions [28], $\forall a, b, c$ points in feature space:

1. non-negativity: $D(a,b) \geq 0$
2. reflexivity: $D(a,b) = 0$ if and only if $a = b$
3. symmetry: $D(a,b) = D(b,a)$
4. triangle inequality: $D(a,b) + D(b,c) \geq D(a,c)$

4.1 Pruning implementation

The pruning technique makes use of a minimum spanning tree (MST) in feature space. A MST of a set of points (in d-dimensional space) is defined as a tree that connects all the points, and whose sum of all edge lengths (or, more generally, edge "weights") is as small as possible.

[1]In fact, the minimum spanning tree can be constructed even for a feature space distance measure that is not a metric.

This method is referred to as "semi-supervised" because, unlike in traditional unsupervised classification (also known as clustering techniques), some prior information exists in this application: the number of main clusters, and their relative position in feature space is known (there could be other, smaller, clusters produced by acquisition artifacts, or by other brain tissue classes than the main three[2] such as fat, skull, or brain lesions). Furthermore, each sample point has an initial labeling suggested by the TPM-based point selection process (section 3.1); the assumption is that the majority of these initial labelings are correct. 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 (MST) of the input set of points is constructed, in feature space (see Figure 4.1).

2. Iteratively, the graph is broken into smaller trees (connected components, or clusters) by removing "long" edges from the initial MST. At each step, the *main clusters* are identified (and labeled) by using prior knowledge, and a stop condition is tested on them; if not satisfied, the graph breaking is continued.

3. At the end, the main cluster points that are in the right cluster (have the same initial labeling as their cluster) are deemed to be true positives and kept; all the other points are deemed to be incorrectly labeled (false positives) and discarded. Note that there may be more clusters than the main four (corresponding to the four classes sought for) – the minor, smaller clusters are considered to be false positives entirely.

MST computation: This work uses an implementation of Kruskal's algorithm [2, 43]. The computation time for the prototype implementation used here is $O(n^3)$

[2]CSF, grey matter, white matter.

for large n, where n is the number of points. However, by using union-by-rank and path-compression methods for an efficient connected components implementation, the time complexity can be reduced to $O(n^2 \log n)$ [22]. Furthermore, using the property that the (Euclidean distance) minimum spanning tree is a subset of the Delaunay triangulation, the computation can be reduced to $O(n \log n)$ in 2-dimensional space [53]. In 1-dimensional space, computing the (Euclidean distance) MST is equivalent to sorting, which is $O(n \log n)$.

MST breaking: The goal of MST breaking is to remove the "long", or "inconsistent", edges in order to separate the feature space clusters. Two heuristic methods (inspired by [28]) were implemented and experimentally evaluated (Chapter 5). Both use a threshold value T, which is decreased at each iteration of the algorithm and tested on all edges of the graph in parallel.

- METHOD A: an edge (i, j) is removed if $length(i, j) > T \times A(i)$ or if $length(i, j) > T \times A(j)$, where $A(i)$ is the average length of all the other edges incident on node i (see Figure 4.1).

- METHOD B: an edge (i, j) is removed if $length(i, j) > T$.

If the decreasing T reaches 1.0 for method A, or 0.0 for method B, and the stop condition is still not satisfied, then the pruning method signals "failure" and discards all the input points.

Main clusters identification: The main clusters are the best guesses for the true background, CSF, grey matter, and white matter clusters in feature space. Making 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. If this assumption is not valid (because the TPM point extraction threshold τ is too low), then the pruning result will be incorrect; various

τ values are experimentally explored in Chapter 5. Note that early in the iterative process some of these main clusters will not be distinct (because, for example, the gray and white clusters were not yet separated).

Stop condition: If the above determined main clusters are found to be four distinct clusters, and the relative cluster locations in feature space correspond to prior knowledge, the iterative graph breaking stops. The prior knowledge used in this work was the relative ordering of the tissue intensities on a T1 image (in increasing order: background, CSF, grey matter, white matter). The cluster locations are estimated as the cluster median, along the T1 feature axis (even for multi-dimensional feature spaces). The assumption here is that, even if the clusters overlap somewhat in feature space, their medians along the T1 axis are still in the correct relative order – this is a reasonable assumption to make (for adult human brains), even in the presence of strong artifacts, as shown in Figure 3.6. Other MRI contrast (such as T2) could be used for this purpose, as long as a similar ordering assumption is valid.

4.2 Practical problems

4.2.1 Data precision

In practice, the signal intensity data produced by the MRI scanner has a limited numerical precision – typically 12-bit, corresponding to at most 4096 distinct intensity levels. The various pre-processing operations (which traditionally use a fixed point 12-bit data representation for the intermediate files) may further reduce the data precision. The result is that when a large number of points are selected based on the TPM not all of them will have distinct intensity values – in other words, some points will be coincident in feature space. The problem is most severe when only one feature (MRI) is used: for example, on a typical T1 dataset, 400 TPM-selected

points was found to give only 200 distinct intensity values. When more features (multi-spectral MRI) are used the problem is reduced: each additional feature effectively increases the precision of the fixed-point intensity data representation (e.g. for two different MRI contrasts, each point will be represented by $2 \times 12 = 24$ bits in feature space).

Zero-length edges in the MST can confuse the edge-breaking method A, so the graph data structure was adapted to handle two kinds of relationships (edges) between the nodes: "neighbors" (normal edges), and "roommates" (degenerate zero-length edges) – see Figure 4.2. The MST operations (construction, breaking) ignore the "roommate" edges; however, these edges are considered by the rest of the pruning method (for example, when traversing all the points in a cluster).

Even if the "roommate" edges are ignored, due to the lack of resolution in feature space, the pruning does not work well on a large number of input points n; for example, in the single-feature case, for $n \to \infty$, any non-zero graph edge will have only one possible length: the intensity quantization step. It was observed experimentally that too many input points force the iterative pruning method (page 26) to advance to lower values of the threshold T, resulting in excessive fragmentation of the feature space clusters, and even failure of the pruning (due to the inability to separate and identify suitable main clusters). Consequently, the number of true positives (correctly labeled points) left after the pruning is reduced.

On the other hand, if the input set size n is too small the feature space data densities will not be adequately sampled. The practical upper bound on n for single-feature (T1) pruning was experimentally explored. The figure of merit chosen for this experiment was the percent of original true positives preserved. The results for simulated data (described in section 2.2.1) and for the real young-normal dataset (described in section 2.2.2) are presented in Table 4.1. Based on these data, 150 points per class were chosen for the pruning validation experiments on T1 simulated data, and 60 points per class for the T1-only real dataset (these validation

experiments are presented in Chapter 5).

4.2.2 Multi-dimensional feature space

If only one MRI image is to be used, then a T1-weighted scan is preferred, since it provides the best contrast between the main three brain tissue types. But using multi-spectral MRI data (that is, MRI data of different modalities, typically T1-, T2-, and PD-weighted) has advantages:

- it improves the cluster separation in feature space, especially in the presence of significant imaging artifacts – the noise and the INU are not correlated between T1 and T2 acquisitions, for example.
- it reduces the limited numerical precision problem (section 4.2.1).

Nevertheless, there are drawbacks too:

1. the possibility of registration errors: T2 and PD data are acquired separately and need to be registered to the T1 data.
2. T2 and PD data are typically lower resolution (2-3 mm thick slices instead of the 1 mm typical for the T1)[3], hence exhibit more partial volume effect.
3. the Euclidean distance in d-dimensions ($d > 1$) is *not* invariant to independent scaling of the different axes, and the MRI scanner raw output has no absolute nor guaranteed scale; if one of the MRI images has a range much smaller than the others, then it will not contribute much to the feature space distance metric [4].

[3]Commonly used T2- and PD-weighted MRI acquisition sessions are much longer than T1-weighted ones (for the same spatial resolution). Longer acquisition means more discomfort for the human subject, and also increased scanning costs.

[4]This could also be exploited: a certain feature's values can be artificially scaled down in order to reduce that feature's influence to the classification decisions.

Table 4.1: Single-feature (T1 MRI) pruning: influence of the input set size on the fraction of the original true positives that are preserved in the output. The observed decrease of this fraction is due to the limited numerical precision problem, which becomes more severe for larger input sets (see section 4.2.1 for details).

Simulated data		
input points (per class)	true positives preserved (%)	
	method A	method B
40	84	89
75	87	88
150	83	83
300	68	54
600	30	0

Real young-normal dataset		
input points (per class)	true positives preserved (%)	
	method A	method B
40	54	54
75	38	40
150	7	7

However, the latter problem can be addressed by a pre-processing step that adjusts ("normalizes") the ranges of the input MRI-s. The pre-processing used in this work was a simple range-matching procedure: the end-points of the intensity histograms were matched together between image modalities (specifically, the image intensities of the T2 and the PD were scaled to match the T1).

The end-points were not the absolute minimum and maximum values in the 3D images, but a certain small percentile away from the extrema of the intensity histograms. The rationale for this is the following: the MRI image can contain regions (caused by imaging artifacts, for example) which are much brighter or much darker than the intensities of the main tissue classes (CSF, grey matter, white matter); however, such regions amount to a small number of voxels compared to the main tissues, thus the above mentioned end-points are reliable estimators for the "shoulder" of the histogram peaks (corresponding to the main tissues). Based on examining a number of MRI scans, percentiles of 4/0.5/4% for, respectively, T1/T2/PD were determined as satisfactory estimators for the location of the histogram peaks' shoulders, on brain scans spatially registered to the Talairach stereotaxic space (if the skull was removed in a pre-processing procedure, 2/0.25/2% are adequate percentile values).

4.2.3 Large sets

This section concerns a practical implementation issue: the pruning algorithm cannot work on sets of sample points that are too large, for two reasons:

1. limited data precision (see section 4.2.1); this is a problem mostly for the single-feature case.

2. computational complexity (time, and also space) for MST computation; it is a limiting factor primarily for the multi-feature case (see section 4.1).

However, large sets of training points (thousands per tissue class) are needed by the final supervised non-parametric classifier in order to perform well. Moreover, the pruning method can occasionally fail on a particular input set of points (because it cannot separate the main clusters), in which case all input points are rejected.

These practical problems are addressed by the following scheme:

1. Generate many small sets of points (≤ 150 points per class, see section 4.2.1), or "chunks", using the TPM. The set generation is done, for each tissue class, by a uniformly-distributed random sampling of all the qualifying locations in the class TPM.

2. Prune each (small) set separately. This can be done in parallel, if appropriate computing facilities exist.

3. Merge all the resulting pruned (small) sets of points into a large set of points, which is considered the final output of the pruning method. The merging is done by simple set union; points with conflicting labelings are discarded.

4.2.4 Skull removal

A skull removal procedure (also known as skull stripping, or intra-cranial cavity extraction) should be applied to the MRI data before feeding it to the pruning and classification algorithms; otherwise, TPM-selected "background" class training points might fall on the skull, and the pruning algorithm may be confused by the multi-cluster distribution of the background class. Many such automatic procedures exist [33, 36, 49, 65, 79, 80].

4.3 Final classification

As explained in section 3.2, it is desirable to use a non-parametric classifier (which is capable of modeling data distributions which are not multi-variate Gaussian,

or Normal) for the final supervised classification of all the voxels in the 3D MRI volume.

The supervised classifier proposed here is the classic k nearest-neighbor (kNN) classifier. Given a training set of samples (which is read and stored during initialization), for each data point to be classified it computes the point's closest k training samples (using a given distance measure in feature space); then, a classification decision is made by taking a vote among these k closest samples. Any ties are resolved by comparing the sums of the feature space distances of the two competing groups.

Previous work [42] showed the artificial neural network (ANN) classifier to perform well on MRI brain data. This neural network is trained using a so called error back-propagation iterative algorithm (a gradient descent technique). An experiment was done to compare the performance of ANN and kNN ($k = 45$, Euclidean distance) on real, multi-spectral (T1/T2/PD), MRI data; the results are shown in Figure 4.3. It can be seen that kNN overall is more robust (has less variability in performance), and is (on average) more accurate when false positives are present in the training set.

Moreover, the ANN classifier has many parameters that need to be tuned to the particular application: number of nodes in the hidden layer, and training-related parameters (learning rate, momentum, maximum number of training iterations, stopping criterion). The behavior of the training process, and the decision boundaries encoded in the trained network are not easily understood.

On the other hand, the kNN classifier has only two parameters: k, and n (the number of training samples). Their influence on classifier's performance (error probability) is more clear:

- n needs to be as large as is practical in order to get a good estimate of the true feature space class distributions (good data sampling).

- Intuitively, k should be "large" in order to use as much evidence as possible;

on the other hand, k should be no more than a small fraction of n in order to keep the k-nearest samples in a relatively small feature space neighborhood of the data point to classify [28]. Quantitatively, it was shown [25, 66] that if k satisfies both of the following two conditions:

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

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

then the kNN classifier (for $n \to \infty$) will give an optimal (minimum) error probability. In practice n is finite, so k has to be chosen more carefully; a commonly used value is $k = \sqrt{n}$ (Enas [29] suggested values of $n^{2/8} \ldots n^{3/8}$).

The computational complexity of the kNN classifier can be reduced by several techniques (see section 4.5.5 in [28]). The implementation used in this work relies on a fast nearest-neighbor lookup library developed by Mount and Arya [51], which pre-processes the training set using box-decomposition (BD) trees [3, 4]. This C++ library also supports other features (that are not used in this work), such as approximate nearest-neighbor searches and distance measures less computationally expensive than the Euclidean distance, so the kNN computational requirements could be further reduced.

Figure 4.3 also shows that sampling the TPM at a lower τ (0.5 instead of 0.9) improves the classification, assuming the pruning works well; this gave an additional motivation for experimentally exploring various τ values – the topic of Chapter 5. On the other hand, on raw training points (with false positives), a lower τ brings worse performance – which is to be expected.

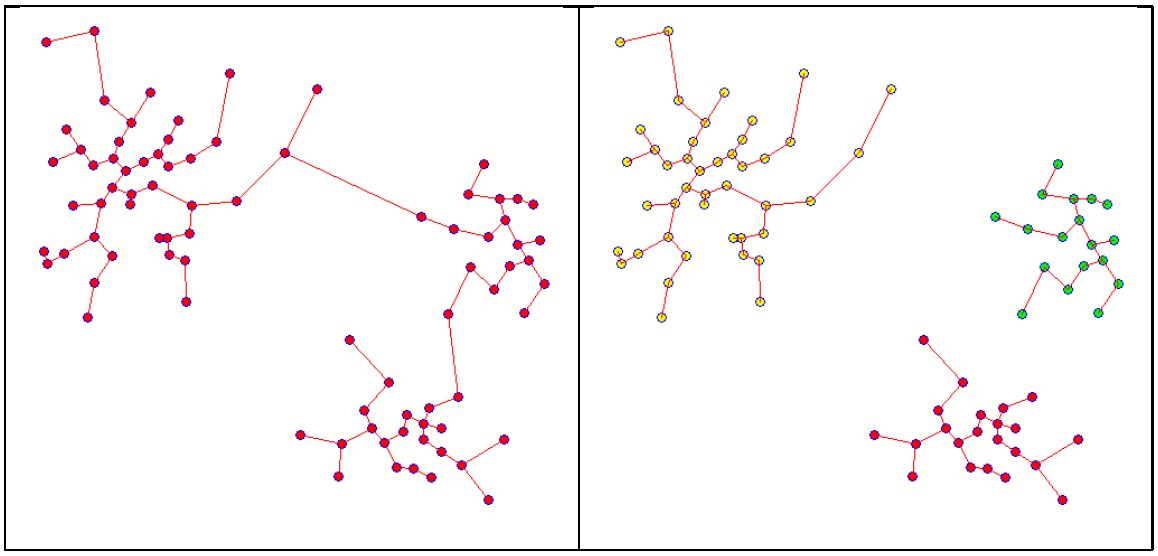


Figure 4.1: *Left*: minimum spanning tree (MST) of a set of points in the plane. *Right*: the result of cutting the "inconsistent" edges using method A (see page 26), at $T = 1.45$. Images produced with [56].

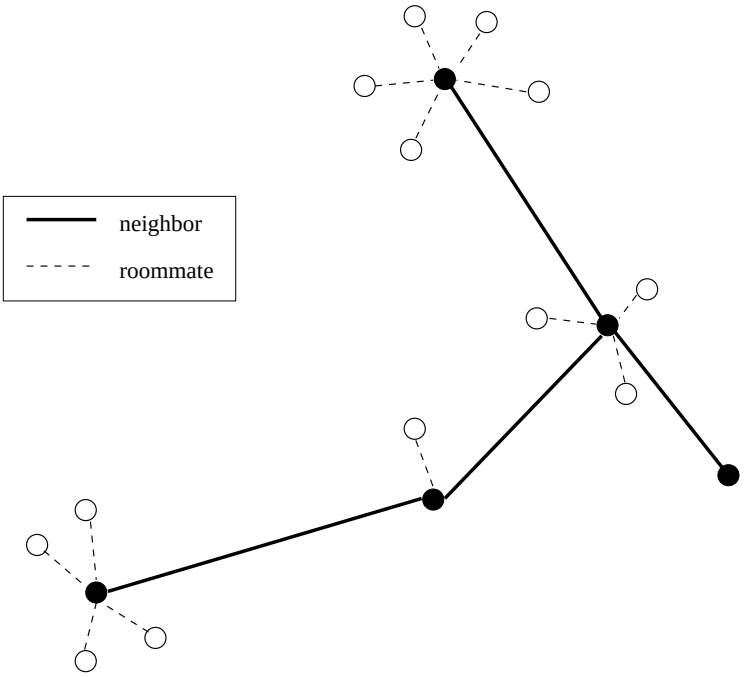


Figure 4.2: Graph with two kinds of relationships (represented as edges) between the nodes: "neighbors", and "roommates". The "roommate" relationship corresponds to a zero-length edge in the spanning tree.

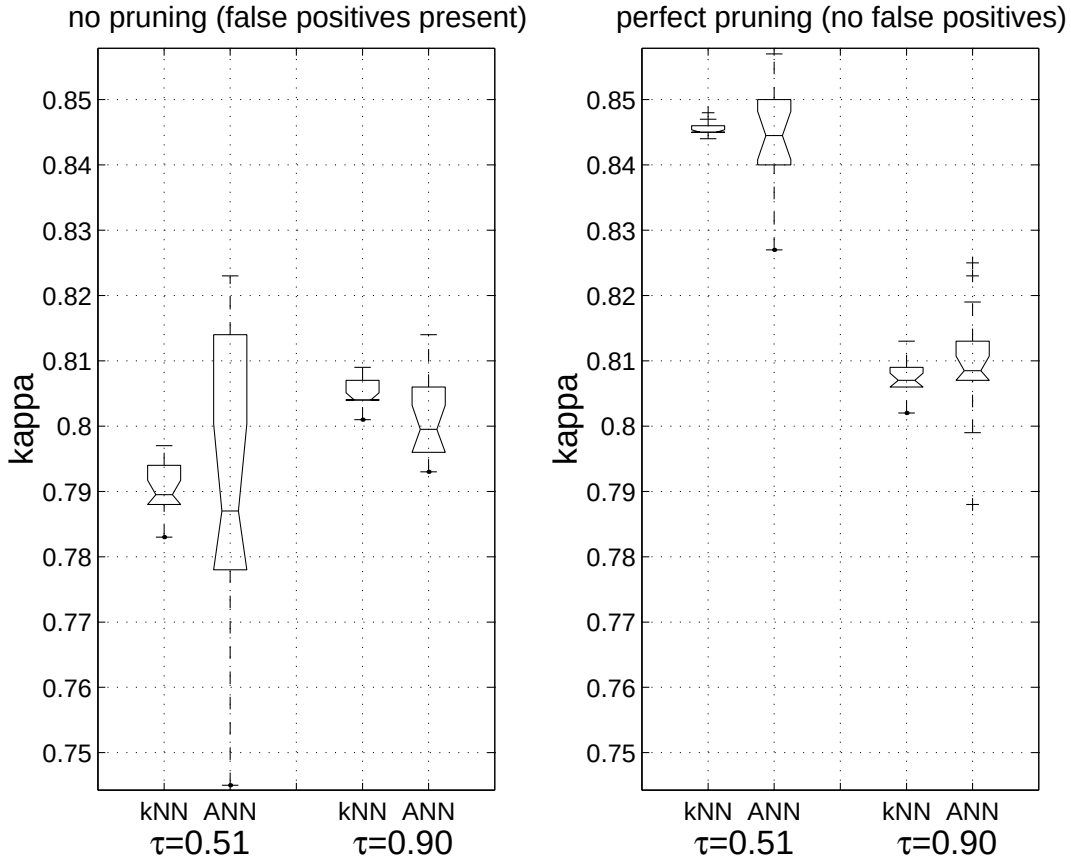


Figure 4.3: Comparison of kNN and ANN classifier performance, as measured by kappa (defined in section 2.2.3). Multi-spectral T1/T2/PD real data (section 2.2.2), 5000 training samples per class randomly extracted from the TPM at the indicated thresholds; 18 repetitions of the experiment with different sets of training samples. It can be seen that overall kNN has less variability in performance than ANN. Moreover, the right-hand plot shows that sampling the TPM at a lower τ (0.5 instead of 0.9) improves the classification, assuming correct pruning. (A description of box-plots is given in Appendix A.)

Chapter 5

Experiments and Results

Experiments were performed in order to validate the training set pruning method, and also the entire brain tissue classification scheme proposed here. The following were explored:

- influence of the TPM threshold τ for training samples extraction (the TPM-s used were the ones produced by Kollokian [42], from a group of young normal subjects, as described on page 15).
- performance on both subject brains similar to the TPM used, and on brains with significant morphological differences from the TPM.
- pruning method A versus method B.
- single-feature (T1 MRI only), as well as multi-feature (T1, T2, PD MRI-s) operation.

This chapter presents the experiments and their results; a discussion and conclusions follow in Chapter 6.

5.1 MRI datasets

Three MRI datasets were used in these experiments:

1. realistic simulations (section 2.2.1) driven by a new custom set of brain "phantoms"; these are described in more detail below (section 5.1.1).

2. real T1-T2-PD (multi-spectral) scans of a young normal individual (dataset described in section 2.2.2).

3. real multi-spectral scans of ischemia patients (who exhibit brain atrophy); only a qualitative evaluation was performed on these data.

For the quantitative measurements (on 1 and 2 above), the "gold standard" was the anatomical model (the "phantom") for the simulations, and the manual classification for the real dataset. For the latter, the cerebellum was left out when computing the various quantitative measures (since its classification was not available). However, the cerebellum was not masked out before the pruning process, since in a practical real image analysis situation the manual classification will not be available. For this work, instead of using an automatic skull removal procedure (section 4.2.4), the skull was simply removed using the known "gold standard" for the classification process.

5.1.1 Elderly brain simulations

Since the particular TPM-s used in this work represent the spatial probability distribution of a young normal population, a set of 10 realistic "phantoms" of elderly brains[1] were created as follows (based on the standard phantom – see section 2.2.1):

[1]Elderly subjects were chosen as their brains have a clear anatomical difference from the young population that generated the TPM-s (see Figure 3.3).

1. A set of 10 individual T1 scans were selected from a large database[2] available at the MNI. All are 60-70 years old, and from the same population. Half are males, half are females.

2. These individual scans were non-linearly registered ("warped") to the standard phantom using the ANIMAL method of Collins [17, 18].

3. The resulting deformation field was inverted and used for deforming the standard phantom, such that it looks similar to the source individual T1 scan.

The outcome is a set of 10 different "phantoms" – see Figure 5.1 for samples. While this procedure does not produce simulated scans which are identical to the original real T1 scans (since the non-linear registration procedure cannot match all the differences in the cortex folding pattern), the resulting anatomical models have known characteristics of aging brains [9]: increased atrophy, enlarged ventricles, and so on.

All the MRI-s (T1, T2, PD) were simulated as 1mm isotropic voxel acquisitions, with 3% noise, and with 20% INU (intensity non-uniformity). These are generally considered typical artifact severity[3] [14, 42].

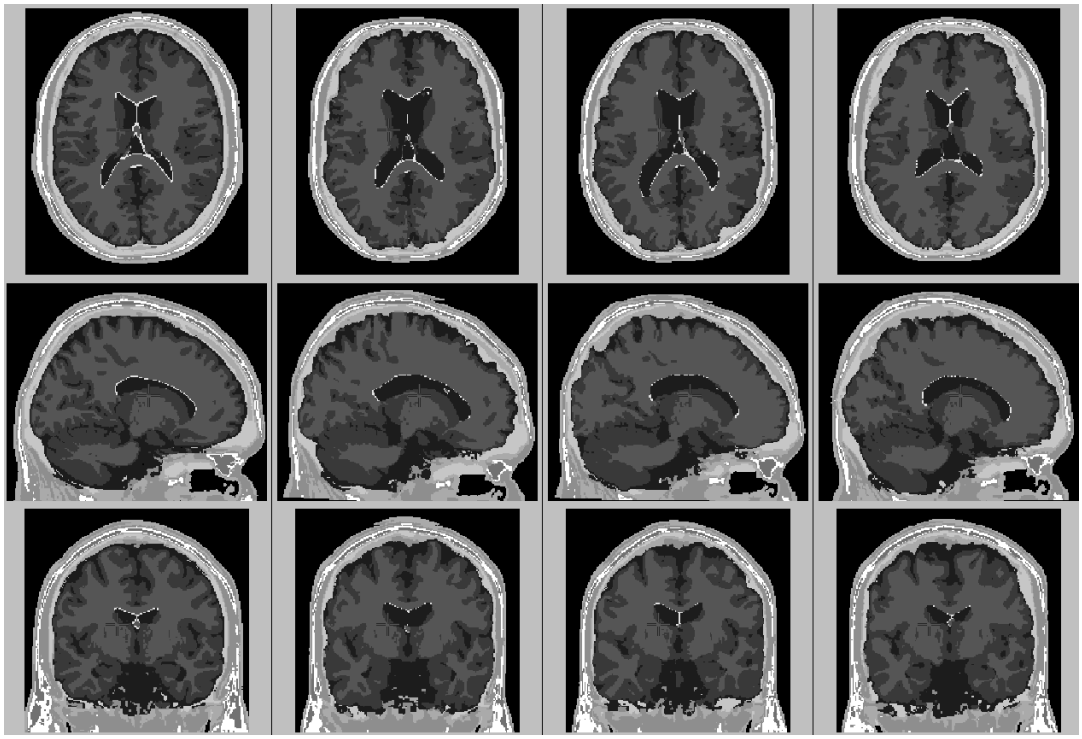
5.2 TPM threshold τ

Experiments were carried out in order to study how the TPM sample extraction threshold τ affects the performance of the pruning, and of the entire tissue classification scheme. The following experiments were performed, each with several repetitions (for assessing the statistical significance of each resulting data point):

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

[3]In a practical image analysis system, an INU correction pre-processing method (such as [63]) is typically employed, thus the INU level may be less than 20%.

A:



B:

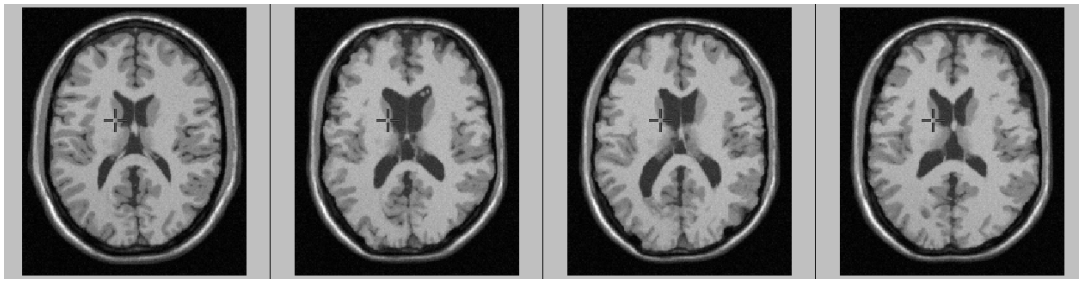


Figure 5.1: **A**: standard phantom (column 1), and elderly brain phantoms (columns 2-4), each with three orthogonal sections. Each tissue is represented as a different gray shade. Compared to the standard phantom (representing a young normal individual), the "elderly" phantoms (section 5.1.1) exhibit enlarged ventricles and overall brain atrophy (typical characteristics of aging brains).

B: T1-weighted simulated MRI images based on the phantoms above.

1. "elderly" brain phantom simulations, with a single-feature: T1-weighted simulated MRI with 3% noise and 20% INU; 7500 input points per class, except the $\tau = 0.99$ data points which had only 2000 input points[4]; 10 different source phantoms, each with 3 different MRI simulations (the difference was in the shape of the INU field[5], and in the noise) and 3 different sets of input points randomly sampled based on the TPM – for a total of 30 repetitions per data point.

2. "elderly" brain phantom simulations, with three features (T1/T2/PD); one set (triplet) of T1, T2, PD simulated MRI-s; experiment design same as for 1, but with only one experiment per phantom, for a total of 10 repetitions per data point (same set of input points for all repetitions).

3. the real dataset, with a single-feature: T1 (same MRI for all repetitions); 7500 input points per class (2000 points for $\tau = 0.99$); 10 repetitions per data point, each with a different sets of input points randomly sampled based on the TPM.

4. the real dataset, with three features (T1/T2/PD); experiment design same as for 3.

In the above, when an experiment is said to have been done on 7500 points per class, in fact it consisted of a set of 50 parallel prunings on chunks of 150 points for simulated data, and of 125 chunks of 60 points for the real dataset (as described in section 4.2.1, 150 points per class were chosen for the experiments on T1-only simulated data, and 60 points per class for the experiments on the T1-only real dataset). Although the pruning method can cope with larger sets in the multi-feature situation, for practical reasons (such as allowing a paired comparison of the

[4]Because the CSF TPM only has ≈ 2200 spatial locations with values ≥ 0.99.

[5]The three different INU field shapes were estimated from real MRI-s [14, 42], and then scaled for the required field severity.

single-feature and multi-feature operation) the same point set sizes were also used for the multi-feature experiments.

The following measures were used to quantify the number of false positives (FP) and true positives (TP) in the point sets:

- *false positives fraction (FPF)* :

$$\frac{FP}{FP+TP} \times 100\%$$

  (Note: $FP + TP =$ total number of points, before and after pruning)

- *percent (pct) reduction in FPF* :

$$\frac{FPF_{before} - FPF_{after}}{FPF_{before}} \times 100\%$$

Also of interest is the fraction of the original true positives preserved by the pruning algorithm. This fraction is important because, in practice, a large reduction in the FPF may not be useful if it is coupled with a severe loss of true positives – this could happen if, for example, the algorithm throws away most of the input points, regardless of them being false or true positives.

While a positive (and large) percent reduction in FPF is desirable (as an indicator of correct functioning of the pruning method), the real figure of merit in this application is FPF_{after} (the rate of false positives left in the point set after the pruning). A low FPF_{after} is desired, as it corresponds to a "mostly correct" training set for the final (supervised) tissue classification stage.

Statistics regarding FPF are shown in Figures 5.2 and 5.3 for the simulated data experiments, and in Figures 5.4 and 5.5 for the real dataset experiments. It can observed that for all experiments (with the sole exception of multi-spectral simulations with $\tau = 0.1$), the pruning reduces the FPF in the point set. Also, as τ decreases, the fraction of true positives preserved decreases.

A comparison of Figures 5.2 and 5.3 suggests that, for the elderly brain simulations, with method A, for $\tau > 0.1$, the reduction in FPF is slightly less in

the multi-feature (T1-T2-PD) than in the single-feature operation (T1 only). For $\tau \geq 0.3$, method B performs similarly on single-feature and multi-feature data, but erratically on multi-feature data for $\tau < 0.3$.

For the real dataset, a comparison of Figures 5.4 and 5.5 suggests that with method A, for $\tau > 0.1$, a slightly higher reduction in FPF is achieved in the single-feature (T1 only) than in the multi-feature (T1-T2-PD) operation. For $\tau \geq 0.5$, method B performs similarly on single-feature and multi-feature data.

Further measurements of practical interest derived from all these experiments are the topic of the next section.

5.3 Final tissue classification

It is important to study how pruning influences the results of the final supervised tissue classification stage. Based on the experimental observations of section 4.3, the kNN supervised classifier was used, with k=45. Sample classification results (images) are given in Figures 5.6 and 5.7.

For a quantitative figure of merit, the kappa similarity measure (defined in section 2.2.3) was computed against the gold standard; kappa was only computed over the brain area (including the CSF between the cortex and dura/skull), as indicated by the gold standard. For comparison, the plots also show the result for an experiment with $\tau = 0.99$ and no-pruning ("raw"): the supervised classifier was simply trained with the raw samples extracted from the TPM at $\tau = 0.99$ (this is the traditional method [42]).

The resulting values for the simulated data experiments are plotted in Figures 5.8 and 5.9. Kappa results for the real dataset experiments are given in Figures 5.10 and 5.11. For all experiments, the sharp decreases in kappa for some low τ values can be explained by the corresponding plots of "true positives preserved" (Figures 5.2-5.5): if the number of training points kept by the pruning method is too low, the

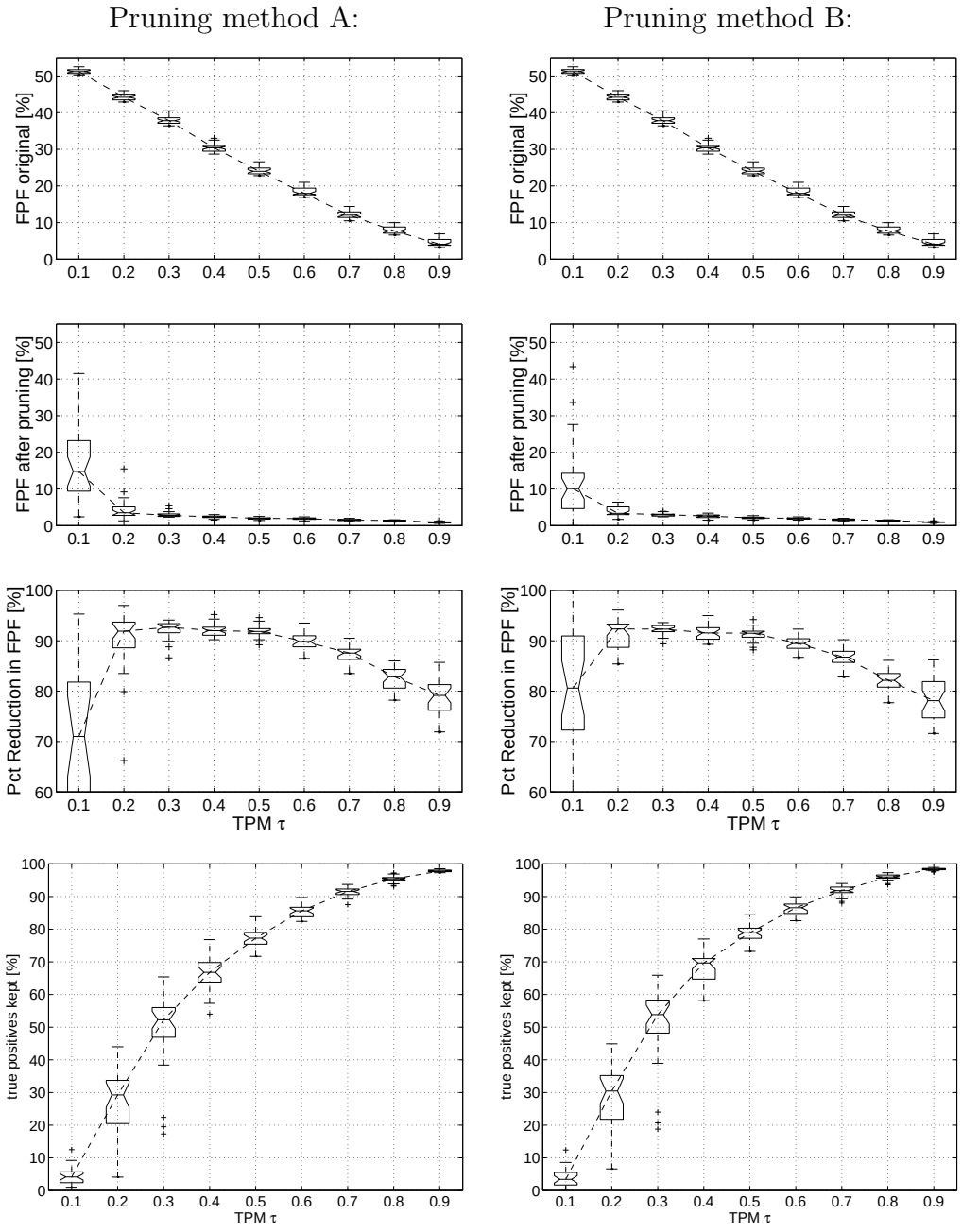


Figure 5.2: *Simulations (elderly brain phantoms), T1-only:* There is no significant difference between methods A and B. For $\tau \geq 0.3$, $FPF_{after} \leq 3\%$ (and does not vary significantly with τ), and $\geq 50\%$ of the true positives in the input are preserved. (see page 43 for the definition of the measures plotted above, and Appendix A for details on box-plots)

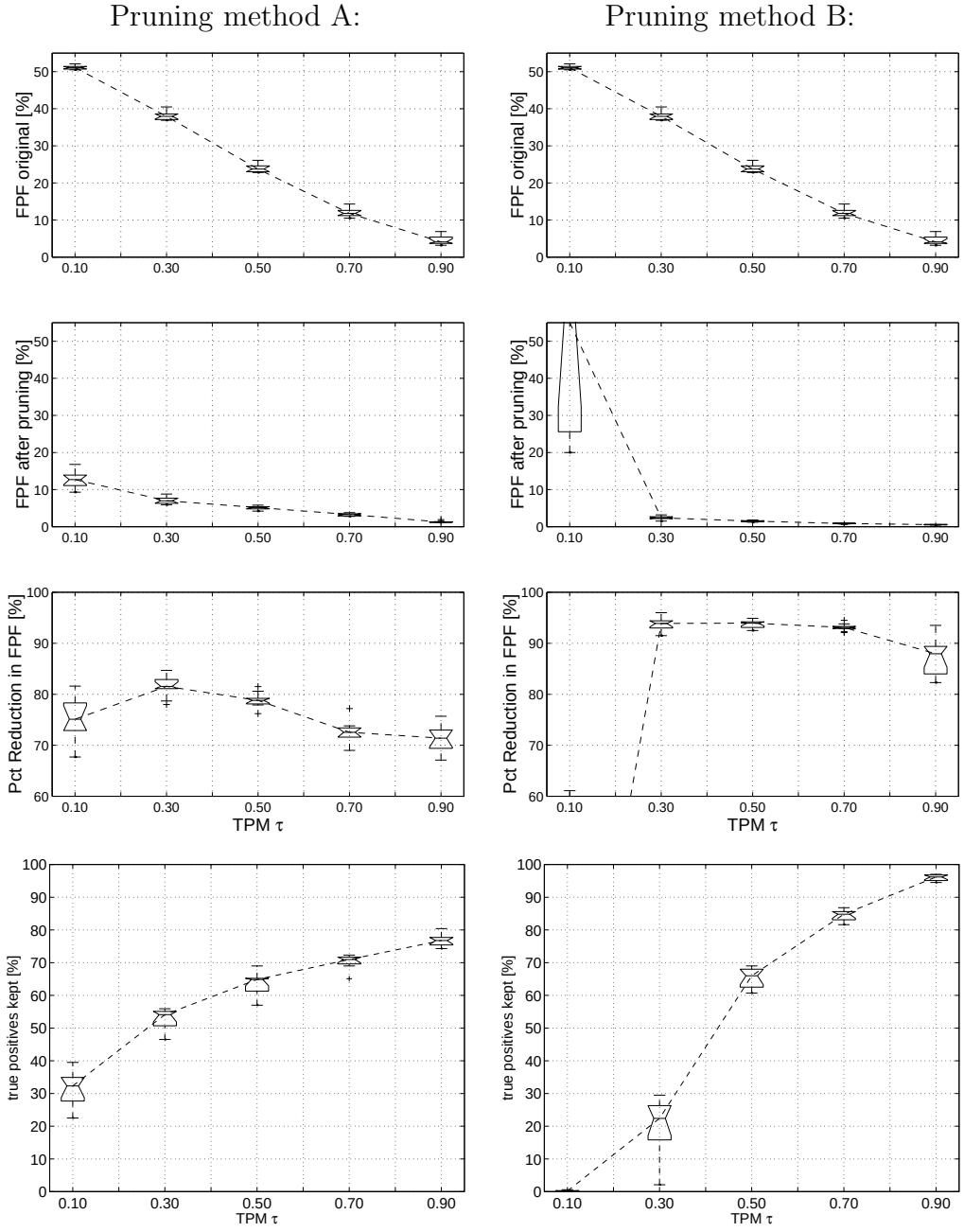


Figure 5.3: *Simulations (elderly brain phantoms), T1-T2-PD:* The plots suggest that method B performs better than A (higher reduction in FPF, thus lower FPF_{after}); however method B breaks down for $\tau < 0.3$ (unlike method A, which still performs satisfactorily at $\tau = 0.1$). With both methods, for $\tau \geq 0.4$, FPF_{after} remains low ($\leq 5\%$), and $\geq 50\%$ of the true positives in the input are preserved.

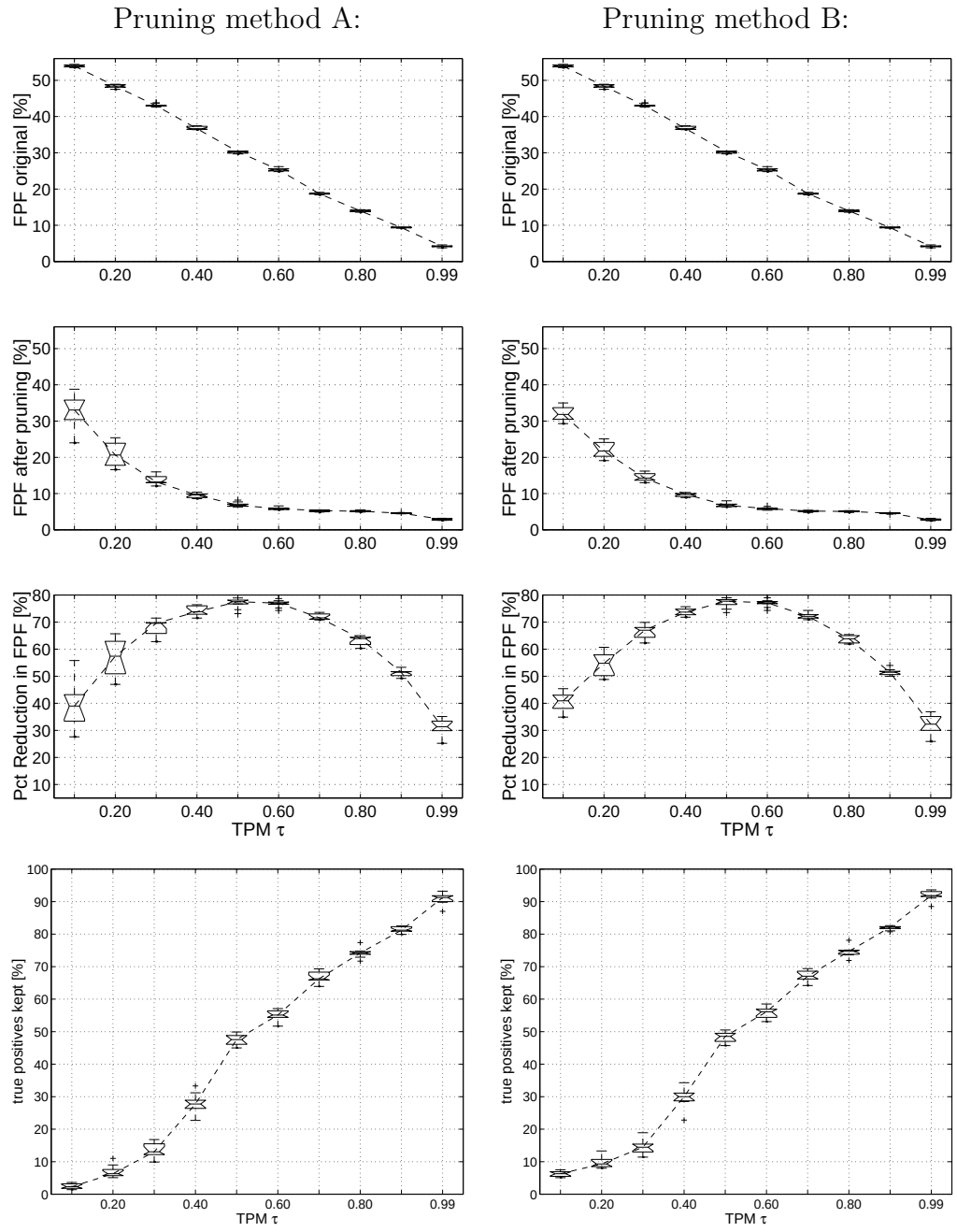


Figure 5.4: *Real dataset, T1-only:* There is no significant difference between methods A and B. For $\tau \geq 0.6$, $FPF_{after} \leq 6\%$ (and does not vary significantly with τ), and $> 50\%$ of the true positives in the input are preserved.

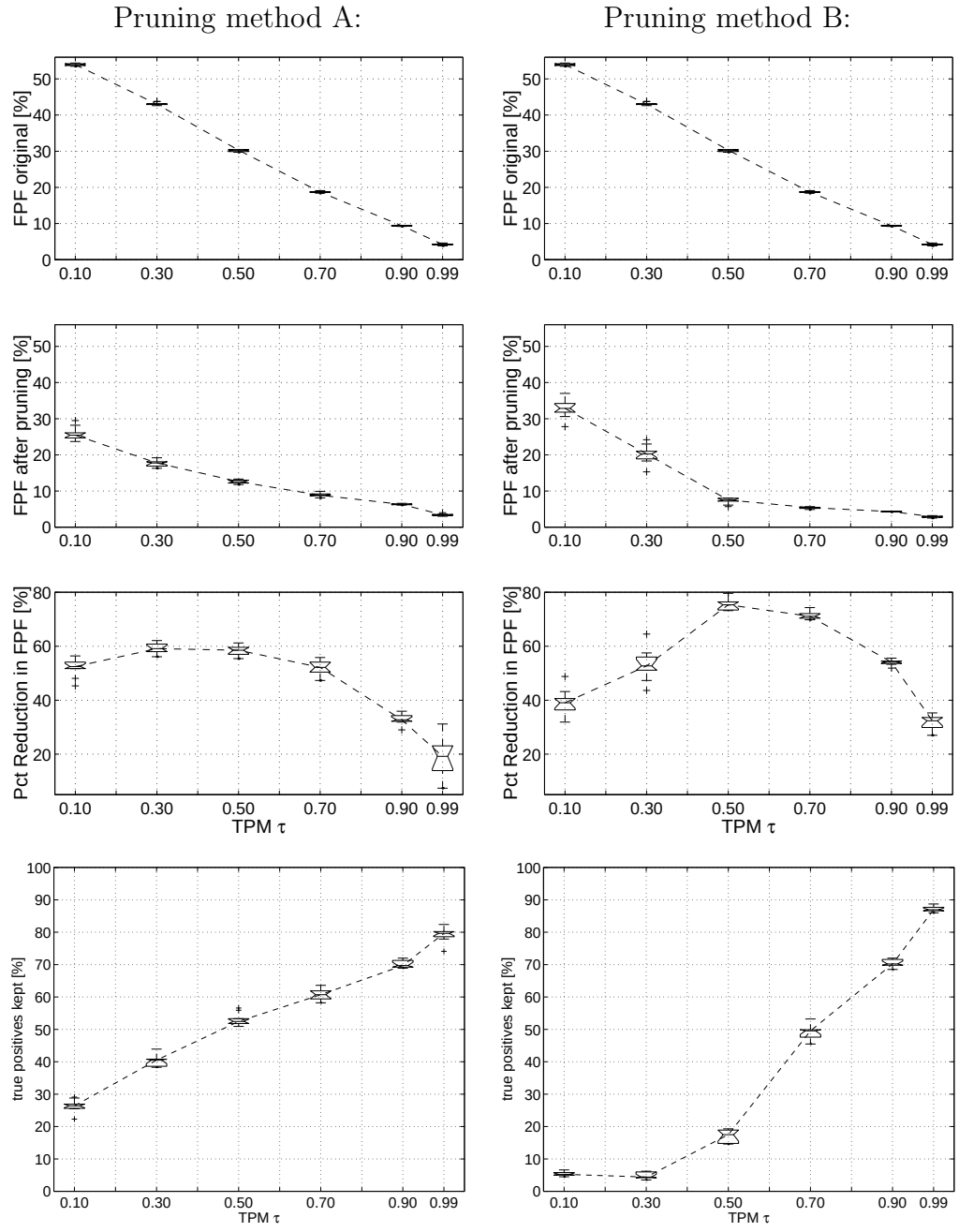


Figure 5.5: *Real dataset, T1-T2-PD:* The plots suggest that method B performs better (higher reduction in FPF) for $\tau \geq 0.5$, but performs poorly for $\tau < 0.5$ (very few true positives preserved); method A is superior for $\tau \leq 0.3$. With both methods, for $\tau \geq 0.7$, FPF_{after} is satisfactory ($\leq 9\%$), and $\geq 50\%$ of the true positives in the input are preserved.

performance of the final kNN classifier is poor.

5.3.1 Qualitative evaluation on additional real data

The pruning method (and the entire tissue classification scheme proposed here) was also validated on additional real multi-spectral MRI data[6] acquired as part of a clinical trial of patients diagnosed with ischemia. Since no "gold standard" was available for these data, only a qualitative evaluation of the resulting classification was possible. As before, 7500 raw training points per class were selected based on the TPM-s; the pruning was done on chunks of 150 points (per class) at a time. However, unlike in the previous experiments, the skull was not removed before the classification process.

The results for a sample dataset are shown in Figure 5.12. This particular dataset exhibits severe brain atrophy (i.e. significant morphological difference from the "young normal" TPM), which is the likely cause for the poor performance of the classification without pruning.

[6] T1: $1\,mm^3$ resolution, T2/PD: 1x1x3.5$\,mm$ resolution; same pre-processing as for the other real data (section 2.2.2).

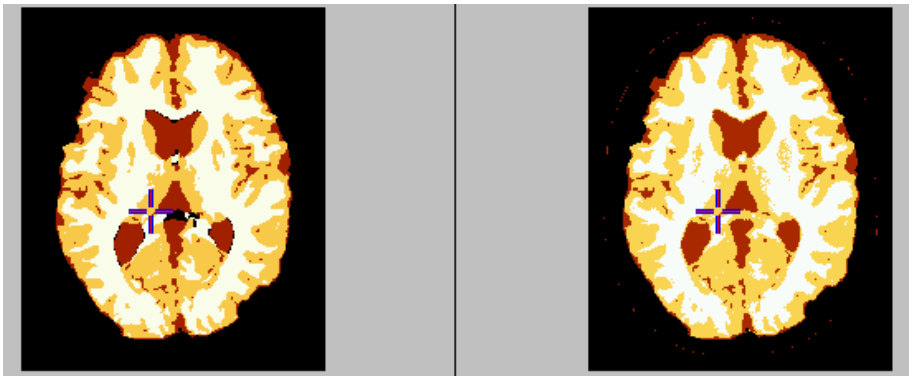


Figure 5.6: *Simulations (elderly brain phantom), T1-T2-PD:* Left: anatomical model ("gold standard"), right: final classification result for method A, $\tau = 0.50$, multi-feature (T1-T2-PD) operation. The different gray levels correspond to the different classes.

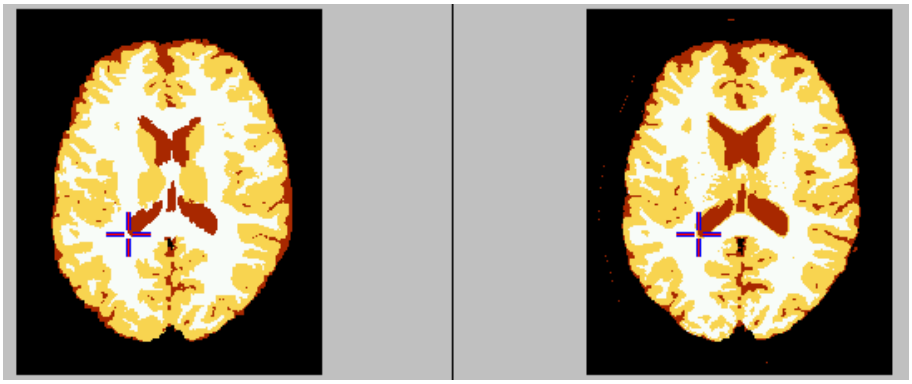


Figure 5.7: *Real dataset, T1-T2-PD:* Left: manual classification ("gold standard"), right: final classification result for method A, $\tau = 0.90$, multi-feature (T1-T2-PD) operation. The different gray levels correspond to the different classes.

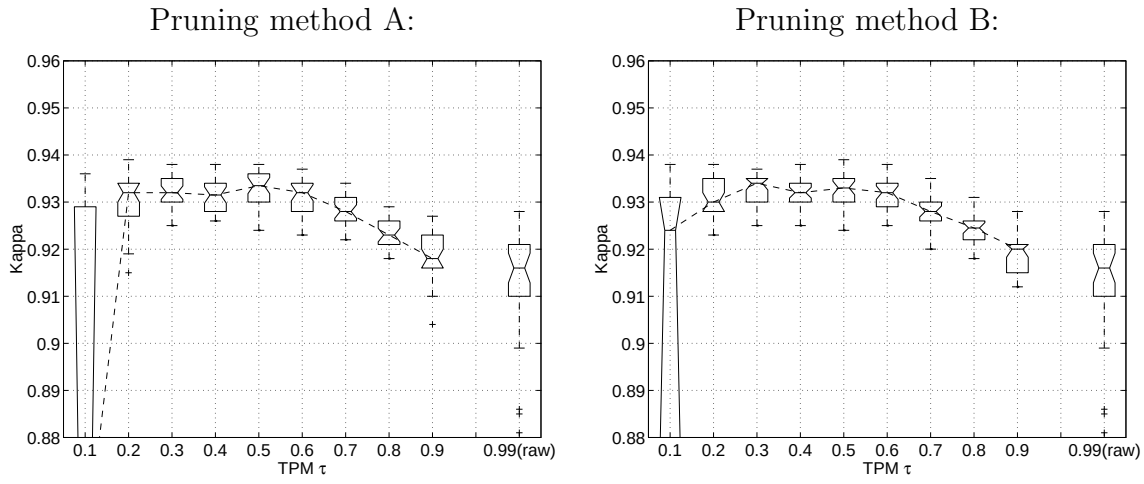


Figure 5.8: *Simulations (elderly brain phantoms), T1-only:* Final classification kappa (see section 5.3). For comparison, "0.99(raw)" shows the result for no pruning. There is no significant difference between methods A and B. For $\tau = 0.2...0.8$, the pruning gives a statistically significant improvement over "0.99(raw)" (box notches do not overlap), with the highest kappa-s for $\tau = 0.2...0.6$.

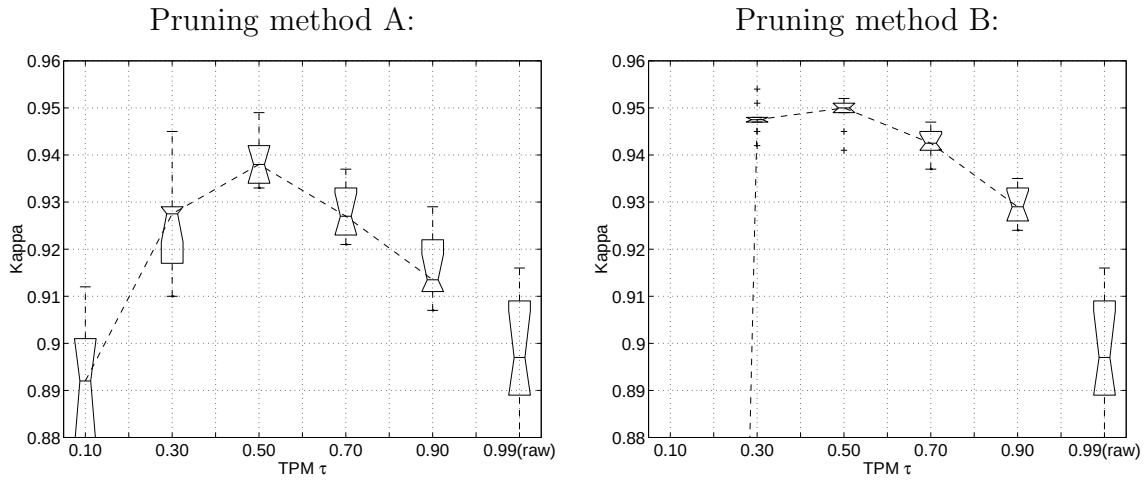


Figure 5.9: *Simulations (elderly brain phantoms), T1-T2-PD:* Final classification kappa. For comparison, "0.99(raw)" shows the result for no pruning. The plots suggest that method B performs overall better than A, and that the highest kappa is at $\tau = 0.5$ for both methods. For $\tau = 0.3...0.9$, the pruning gives a statistically significant improvement over the "raw" experiments (box notches do not overlap).

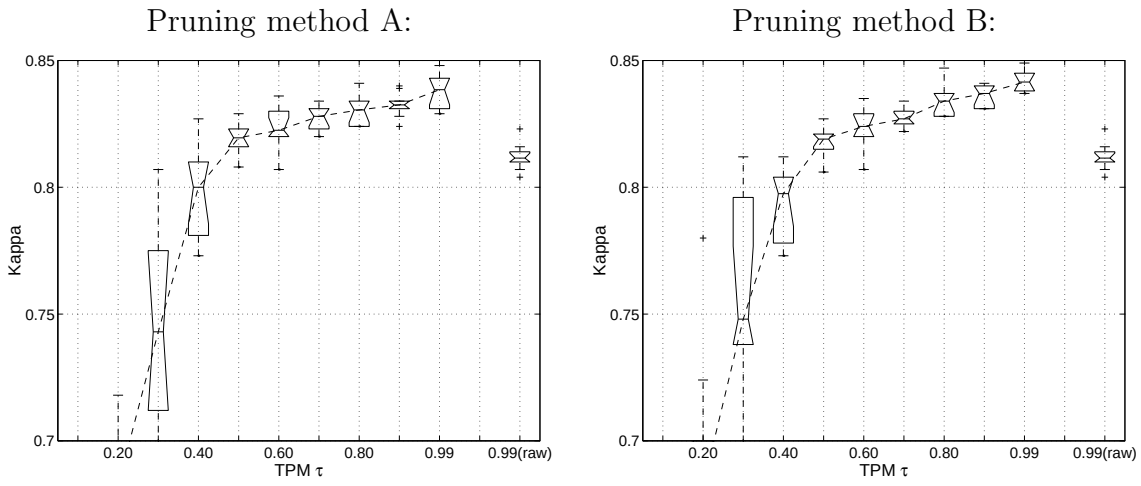


Figure 5.10: *Real dataset, T1-only:* Final classification kappa. For comparison, "0.99(raw)" shows the result for no pruning. The plots do not show a significant difference between the two methods. For $\tau \geq 0.5$, the pruning gives a statistically significant improvement in kappa over the "raw" experiments (box notches do not overlap), with the highest kappa-s obtained for $\tau = 0.99$.

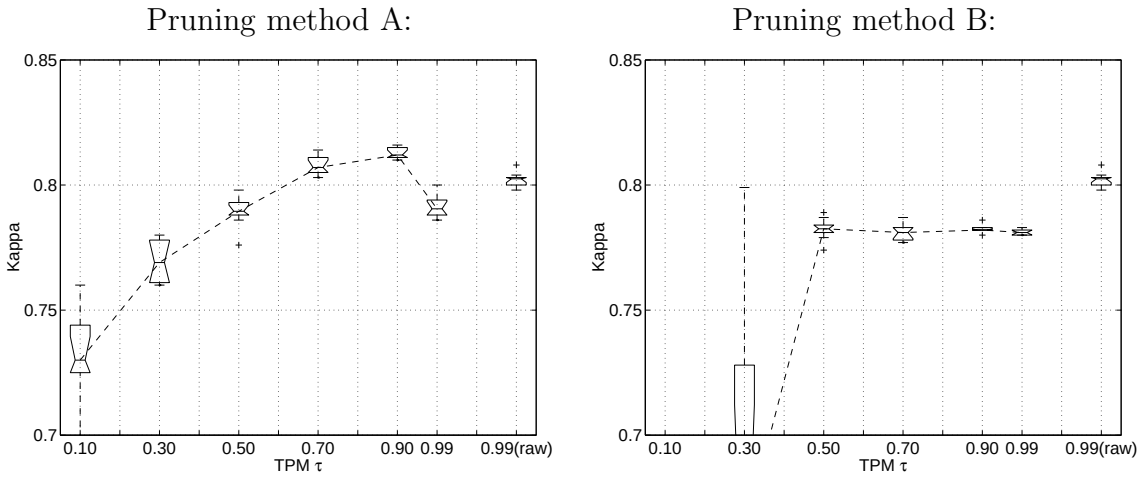


Figure 5.11: *Real dataset, T1-T2-PD:* Final classification kappa. For comparison, "0.99(raw)" shows the result for no pruning. Method A produces a statistically significant improvement for $\tau = 0.7...0.9$, with the highest kappa for $\tau = 0.9$. However, method B does not improve kappa over the no-pruning case, and breaks down for $\tau < 0.5$.

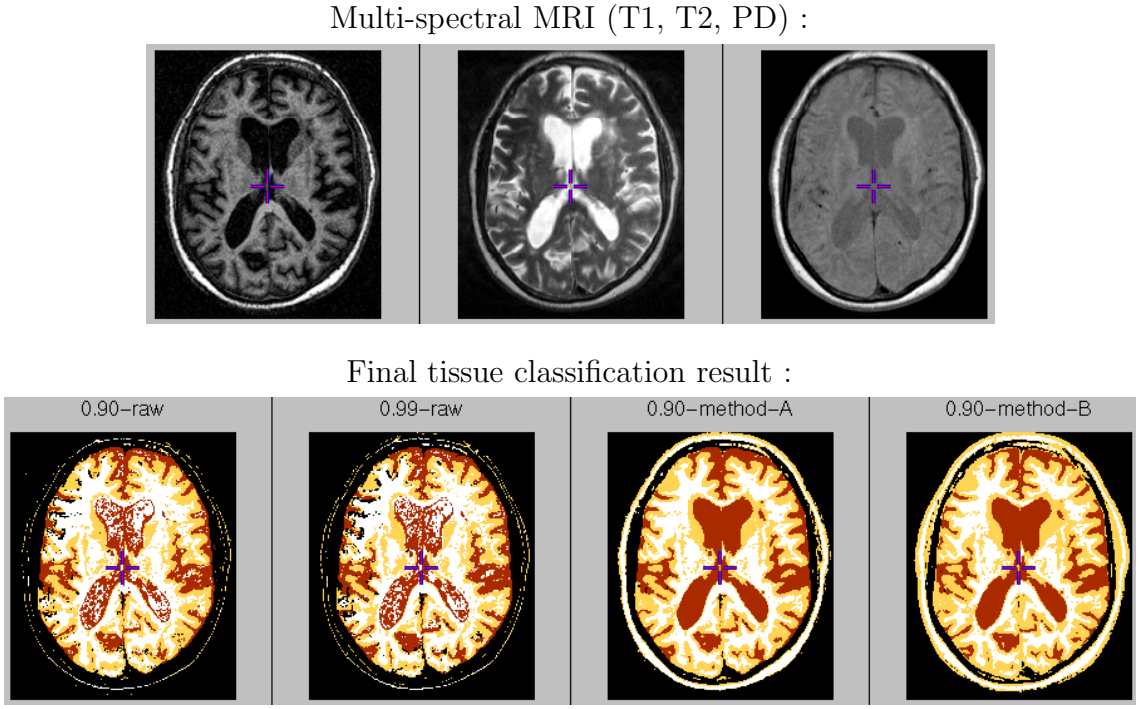


Figure 5.12: *Qualitative evaluation: real multi-spectral dataset 2 (ischemia patient):* The brain tissue classification with no pruning ("raw", $\tau = 0.90$ and $\tau = 0.99$) is poor – note that some voxels inside the ventricles were mis-classified as white matter. Both pruning methods A and B ($\tau = 0.90$) give significantly better tissue classification. (the different gray levels correspond to the different classes)

Chapter 6

Discussion and Conclusions

6.1 Results

Based on the experimental results presented in Chapter 5, it can be concluded that the MST-based pruning method achieves its goal of reducing the rate of false positives (mis-labeled samples) in the point set selected using the TPM-s. Moreover, the pruning improves the final tissue classification.

The pruning method does especially address the situation when the subject's brain anatomy is significantly different than the tissue spatial probability distribution represented by the model (TPM). In this situation, simply using the raw set of training samples extracted based on the prior spatial tissue probability (TPM) is even less adequate.

The TPM used for these experiments was produced from a "young normal" population. Thus, the results on elderly and diseased brains are more important than the results on the young normal dataset.

Here is a summary and discussion of the experimental results presented in Chapter 5.

False positives fraction (FPF): Most importantly, in all the quantitative experiments the pruning method reduces the FPF in the point set. However, as the TPM point selection threshold τ decreases, the fraction of true positives preserved by the pruning also decreases (i.e. the size of the pruned point set decreases), and the FPF_{after} increases. Nevertheless, for any $\tau \geq 0.4$ for the elderly brain simulations, or any $\tau \geq 0.7$ for the real young-normal dataset, more than half of the true positives are preserved, and the FPF_{after} does not vary significantly and remains low ($\leq 5\%$ for simulations, and $\leq 9\%$ for the real dataset).

Final classification result: The results of the complete tissue classification scheme proposed here were judged against the ones of the traditional "0.99(raw)" method (consisting of TPM $\tau = 0.99$ and no pruning, and the same final supervised kNN classifier). On real MRI scans from a database of ischemia patients, the pruning (both method A and B) produced a clear qualitative improvement over the raw method (see Figure 5.12).

In the quantitative experiments on elderly simulated scans, the pruning improved the kappa similarity measure for all $\tau \geq 0.3$; this improvement was statistically significant ($p < 0.05$) for $\tau \leq 0.8$ on single-spectral (T1 only) data, and for $\tau \leq 0.9$ on multi-spectral (T1-T2-PD) data. The improvement was from $kappa = 0.90$ to $kappa = 0.95$ on multi-spectral data, with pruning method B, and $\tau = 0.5$.

On the real scans of a young-normal subject, the kappa had a small improvement (statistically significant, $p < 0.05$) for $\tau \geq 0.5$, in T1-only operation. However, in the multi-spectral (T1-T2-PD) operation the pruning did not always improve the kappa (although the difference was less than 0.02 for all $\tau \geq 0.5$). Nevertheless, this is not a "failure" of the pruning method: since the subject was part of the same population as the one represented by the TPM-s, the no-pruning ("raw") classification works satisfactorily by itself.

Method A versus method B: These two MST breaking methods are defined on page 26. Intuitively, pruning method A should be able to adapt itself to the local conditions in the graph, while method B assumes that clusters have similar variance in feature space (which is not necessarily so, as shown in section 3.2). No significant difference between the two methods was observed in the single-feature experiments (T1-only).

However, in the multi-feature experiments method B dropped more of the original true positives (i.e. produced smaller output point sets) than method A. Moreover, method B exhibited a sharp break-down in performance for $\tau < 0.3$ on elderly simulated data (Figure 5.3), and for $\tau < 0.5$ on the young-normal real MRI dataset (Figure 5.5). In contrast, method A performed satisfactorily even at low values of τ.

In terms of the final classification results (on multi-spectral data), the experiments did not show a clear difference between the two methods (except for the low τ values where method B breaks down). The differences in kappa were ≤ 0.012 on elderly simulations (Figure 5.9), and ≤ 0.04 on the young-normal dataset (Figure 5.11). On the ischemia patient real data, no clear qualitative difference can observed between the two methods (Figure 5.12).

6.1.1 Limitations

The brain tissue classification approach proposed here has a few limitations:

- It requires a prior model: a tissue probability map in a stereotaxic space. However, for human brains this is generally available.

- The training set pruning method requires prior knowledge about the relative position of the class clusters in feature space. In the current implementation, this is specified as an ordering along one of the feature axes. Although the T1-weighted image intensity was used here, other MRI contrasts (such as T2)

can fulfill this role.

- Not all of the false positives are "pruned", and part of the true positives in the original ("raw") training set are discarded as well. This was observed to be due to the pruning method's occasional difficulty in separating the main clusters – the iterative graph breaking procedure (described at page 26) gets to low values of T, which leads to excessive fragmentation of the graph, and to smaller main clusters. The cause of this behavior is the (partial) overlap of the class distributions in feature space.

6.2 Summary of contributions

1. A completely automatic procedure for brain tissue classification from MR anatomical images was described. The procedure uses for initialization prior tissue spatial probability maps (an "anatomical atlas", or "model") in a standard, brain-based coordinate system ("stereotaxic space").

2. A novel method was developed for eliminating ("pruning") the mis-labeled samples from the set of points suggested by the model. This pruning method uses a minimum spanning tree graph-theoretic approach in feature space, together with domain information specific to this application.

3. The classification procedure is robust against morphological differences between the subject of the classification and the particular model used for initialization. Previously published methods (section 2.3) reportedly do not achieve this.

4. The classification procedure is robust against variations in the source MRI data — the use of the pruning method allows the classifier training (initialization) points to have a better spatial coverage of the brain than simple

model-based methods (that are restricted to brain areas with high prior tissue probability).

5. The performance of the procedure was demonstrated by quantitative and qualitative experiments on both real and simulated MRI data, and on subjects who are both similar and dissimilar to the model used for initialization. The quantitative experiments were repeated for achieving statistical significance. In particular, the pruning method gives a statistically significant improvement of classification versus a simple model-based initialization method.

6. The entire procedure is non-parametric in that it does not make any assumptions about data distribution in feature (image intensity) space. Many existing methods assume multi-variate Normal (Gaussian) class intensity distributions and, as discussed in section 3.2, this is can be a questionable assumption.

7. The only requirement of the pruning method is that the majority of the sample locations selected using the model are in fact correctly labeled (this condition should hold for each of the tissue classes in the classification process).

8. This procedure will provide better tissue classification than existing methods for research and clinical studies of the development, functioning, and pathology of the human brain.

6.3 Future work

Several directions for future research can be envisioned:

- The finite spatial resolution of MRI (partial volume effect) causes the blurring of the interface between tissue types in the image. Since the inter-tissue boundary voxels do not contain pure tissue but a mix of two (or more) tissue types, their MR signal intensity will lie between tissue clusters in feature space;

this can hamper the pruning method. It would be desirable for the 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 integrating the gradient information into the feature space is worth exploring.

- The presence of multiple points at the same location in feature space (caused by the limited numerical precision of the intensity data, see section 4.2.1) could be considered by the graph breaking method: intuitively, it should be more difficult to break an edge between two nodes that have many "roommates". A weighting could be applied to the edge lengths according to the true number of points at their ends.

- Because of the inherent limitations of the MRI simulations used in this work (section 2.2.1), it would be interesting to perform additional quantitative performance measurements on real MRI-s of elderly or diseased subjects, or on subjects with space-filling brain lesions (e.g. tumors, stroke), even if only a partial manual classification is available for these scans.

- Regarding the kappa measure of the classification performance, it would be interesting to try other similarity measures as well. For example, a new similarity measure based on information theory was recently proposed [7]. Also, it would be desirable to use a measure which weighs differently various misclassifications, according to their spatial position in the brain – this could be done by using something like "weighted kappa" [16], or by using the standard kappa measure but only on smaller, specific regions of the brain which are typically poorly classified.

Appendix A

Box Plots

This thesis presents experimental results using the so called "box and whisker" plots (produced using Matlab [68]). 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.
- lateral "notches" of the box represent a 95% confidence interval about the median of the sample; hence a side-by-side comparison of two notched boxes is the graphical equivalent of a t-test (if notches do not overlap, then $p \leq 0.05$).
- 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.

Appendix B

Glossary and Abbreviations

cerebellum A large structure at the lower back of the human brain. It has fine structures of intertwined gray and white matter.

cerebro-spinal fluid (CSF) Substance found surrounding the brain and within the ventricular system of the brain and spinal cord.

classification See **tissue classification**.

false positive See Table 2.1.

feature space A coordinate system, typically used by a classifier, where the coordinate along each axis is given by the value of a feature (a measurement).

gold standard The reference, assumed to be the "true answer", against which a segmentation or classification result is evaluated.

intensity non-uniformity (INU) An artifact inherent to the MR imaging process. It is usually observed as a smooth, low spatial frequency variation in the image intensity. See also section 2.1.

k nearest-neighbor (kNN) classifier See section 4.3.

kappa A chance-corrected similarity measure between two image segmentations. See section 2.2.3.

magnetic resonance imaging (MRI) Non-invasive medical imaging technique that can produce high-resolution images with good contrast of the different biological soft tissue types.

MNI Montreal Neurological Institute (McGill University).

partial volume effect Due to the finite resolution of the MR imaging process, the image voxels (typically about 1 mm^3 or larger) may contain a mixture of more than one tissue type, which all contribute to the measured signal. This leads to voxels of intermediate intensity at the boundary between tissues. See section 2.1.

phantom In this thesis, a brain phantom is a digital brain model that is used both as an input to an MRI simulator, and also as a "gold standard" for quantitative evaluation of the automatic classification of the simulated MR image. See section 2.2.1.

proton density (PD) image An MR image in which the intensity of a voxel is predominantly determined by the density of hydrogen atoms present within voxel's spatial extent.

pulse sequence The particular arrangement of control signals in an MRI scanner that produces an image. It controls image characteristics such as tissue contrast, noise, and spatial resolution.

registration In this work, linear spatial registration is the procedure that determines a linear (affine) transformation between two brain-based coordinate systems. If the registration is performed between an individual brain image and a standard atlas (such as Talairach's), the resulting transformation can

be used to resample the individual image to the stereotaxic space defined by the atlas [20, 67]. See also **stereotaxic space**.

resampling The technique of changing the sampling grid of a digital image.

Rician distribution The distribution of the magnitude of the sum of a constant and a complex Gaussian-distributed random variable.

segmentation In this thesis, a brain segmentation is the set of class (tissue type) labels assigned to each voxel in the image volume.

stereotaxic space A standard frame of reference (coordinate system) defined by anatomical landmarks of the human brain. It allows the removal of affine (translation, rotation, scale) differences between individual brains. The particular stereotaxic space used at the MNI (and in this work) is the one defined by the Talairach atlas [67]. See also **registration**.

supervised classifier A classifier that is trained (learns correct behaviour) in a supervised fashion: a sample set of data is supplied, along with the correct classes (the "true answer") for all these data.

$\mathbf{T_1}$ image An MR image in which the contrast between tissues is largely due to the difference in the intrinsic tissue property T_1.

$\mathbf{T_2}$ image An MR image in which the contrast between tissues is primarily given by the difference in the T_2 tissue property.

tissue classification In this context, tissue classification is the procedure of labeling each image voxel with a tissue type. Also called "tissue segmentation".

tissue probability map (TPM) A stereotaxic space TPM of a given tissue is a spatial probability distribution representing a certain subject population. See section 3.1.

training set The set of correctly labeled sample data used to train a supervised classifier.

true positive See Table 2.1.

unsupervised classifier A classifier that does not have access to samples of correctly labeled data. Sometimes, not even the number of classes is known.

ventricles Cavities deep inside the brain, which are part of the central nervous system's ventricular system (filled with cerebro-spinal fluid). The first and second ventricle are the large pair of cavities visible at the center of a brain image.

voxel A voxel is for a 3-dimensional (3D) digital image what a pixel is for a 2D image, i.e. an image element.

Bibliography

[1] Brainweb – simulated brain database. http://www.bic.mni.mcgill.ca/brainweb/.

[2] M. Albertson and J. Hutchinson. *Discrete Mathematics with Algorithms*. Wiley, New York, 1988.

[3] S. Arya and D. Mount. Algorithms for fast vector quantization. In J. Storer and M. Cohn, editors, *Proc. of DCC '93: Data Compression Conference*, pages 381–390. IEEE Press, 1993.

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

[5] J. Ashburner. *Computational Neuroanatomy*. PhD thesis, University College London, 2000.

[6] J. Ashburner and K. J. Friston. Voxel-based morphometry — the methods. *NeuroImage*, 11(6):805–821, june 2000.

[7] F. Bello and A. C. F. Colchester. Measuring global and local spatial correspondence using information theory. *Medical Image Computing and Computer-Assisted Intervention - MICCAI'98. First International Conference. Proceedings. Springer-Verlag.*, pages 964–73, 1998.

[8] J. C. Bezdek, L. O. Hall, and L. P. Clarke. Review of MR image segmentation techniques using pattern recognition. *Medical Physics*, 20(4):1033–1048, July/Aug. 1993.

[9] D. D. Blatter, E. D. Bigler, S. D. Gale, S. C. Johnson, C. V. Anderson, B. M. Burnett, N. Parker, S. Kurth, and S. D. Horn. Quantitative volumetric analysis of brain MR: normative database spanning 5 decades of life. *American Journal of Neuroradiology*, 16(2):241–51., Feb 1995.

[10] M. B. Carpenter. *Core Text of Neuroanatomy*. Williams & Wilkins, third edition, 1985.

[11] H. S. Choi, D. R. Haynor, and Y. Kim. Partial volume tissue classification of multichannel magnetic resonance images - a mixel model. *IEEE Transactions on Medical Imaging*, 10(3):395–407, Sept. 1991.

[12] L. P. Clarke, R. P. Velthuizen, M. A. Camacho, J. J. Heine, M. Vaidyanathan, L. O. Hall, R. W. Thatcher, and M. L. Silbiger. MRI segmentation: Methods and applications. *Magnetic Resonance Imaging*, 13(3):343, 1995.

[13] L. P. Clarke, R. P. Velthuizen, S. Phuphanich, J. D. Schellenberg, J. A. Arrington, and M. Silbiger. MRI: stability of three supervised segmentation techniques. *Magnetic Resonance Imaging*, 11(1):95–106, 1993.

[14] C. Cocosco, V. Kollokian, R.-S. Kwan, and A. Evans. Brainweb: Online interface to a 3D MRI simulated brain database. In *NeuroImage (Proceedings of 3-rd International Conference on Functional Mapping of the Human Brain)*, volume 5 (4, part2/4), page S425, Copenhagen, may 1997.

[15] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological measurements*, 20:37–46, 1960.

[16] J. Cohen. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, Oct. 1968.

[17] D. Collins. *3D model-based segmentation of individual brain structures from magnetic resonance imaging data*. PhD thesis, McGill University, Montreal, Canada, December 1994.

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

[19] D. Collins, J. Montagnat, A. Zijdenbos, A. Evans, and D. Arnold. Automated estimation of brain volume in multiple sclerosis with BICCR. In *Proceedings of 17th International Conference on Information Processing in Medical Imaging*, Davis, CA, USA, june 2001.

[20] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography*, 18(2):192–205, March/April 1994.

[21] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–8, Jun 1998.

[22] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[23] B. M. Dawant and A. P. Zijdenbos. *Handbook of Medical Imaging, Vol. 2: Medical Image Processing and Analysis (editors: Sonka, M. and Fitzpatrick, J.M.)*, volume PM80, chapter 2 ('Image Segmentation'). SPIE PRESS, Bellingham, WA, USA, june 2000.

[24] C. DeCarli, J. Maisog, D. G. M. Murphy, D. Teichberg, S. I. Rapoport, and B. Horwitz. Method for quantification of brain, ventricular and subarachnoid CSF volumes from MR images. *Journal of Computer Assisted Tomography*, 16(2):274–284, Mar./Apr. 1992.

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

[26] R. O. Duda. Feature selection and clustering for human computer interfaces (course notes). http://www-engr.sjsu.edu/knapp/HCIRDFSC/FSC_home.htm, 1997. web course notes.

[27] R. O. Duda. Pattern recognition for human computer interfaces (course notes). http://www-engr.sjsu.edu/knapp/HCIRODPR/PR_home.htm, 1997. web course notes.

[28] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2nd edition, 2001.

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

[30] G. Gerig, O. Kübler, R. Kikinis, and F. A. Jolesz. Nonlinear anisotropic filtering of MRI data. *IEEE Transactions on Medical Imaging*, 11(2):221–232, June 1992.

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

[32] E. Haack et al. *Magnetic Resonance Imaging, Physical Principles and Sequence Design*. Wieley-Liss, New York, 1999.

[33] H. K. Hahn and H.-O. Peitgen. The skull stripping problem in MRI solved by a single 3D watershed transform. In *Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, oct 2000.

[34] G. Harris, N. C. Andreasen, T. Cizadlo, J. M. Bailey, H. J. Bockholt, V. A. Magnotta, and S. Arndt. Improving tissue classification in MRI: a three-dimensional multispectral discriminant analysis method with automated training class selection. *Journal of Computer Assisted Tomography*, 23(1):144–54, Jan-Feb 1999.

[35] K. Held, E. R. Kops, B. J. Krause, W. M. Wells 3rd., R. Kikinis, and H. W. Muller-Gartner. Markov random field segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 16(6):878–86, Dec 1997.

[36] R. K. Justice, E. M. Stokeley, J. S. Strobel, R. E. Ideker, and W. M. Smith. Medical image segmentation using 3-D seeded region growing. In *Proceedings of SPIE: Image Processing*, volume 3034, pages 900–910, Newport Beach, Feb. 1997.

[37] N. Kabani, L. Collins, and A. Evans. Hemispheric differences in gray matter volume of adult human brain. In *Society for Neuroscience Annual meeting*, New Orleans-LA, USA, oct 1997.

[38] N. Kabani and A. Evans. A detailed atlas of the human brain using 3D MRI. *(journal tbd)*, 2001. (in preparation).

[39] N. Kabani, D. MacDonald, C. Holmes, and A. Evans. 3D atlas of the human brain. In *Neuroimage (Proceedings of Human Brain Mapping 1998 Meeting)*, Montreal, Canada, june 1998.

[40] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, and A. C. Evans. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. *IEEE Transactions on Medical Imaging*, 14(3):442–53, Sept. 1995.

[41] E. R. Kandel, J. H. Schwartz, and T. M. Jessel. *Principles of Neural Science*. McGraw Hill, fourth edition, 2000.

[42] V. Kollokian. 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, November 1996.

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

[44] R. K. Kwan, A. C. Evans, and G. B. Pike. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging*, 18(11):1085–97, Nov 1999.

[45] R. K.-S. Kwan, A. C. Evans, and G. B. Pike. An extensible MRI simulator for post-processing evaluation. In *Proceedings of the Fourth International Conference on Visualization in Biomedical Computing (VBC)*, Hamburg, Germany, 1996.

[46] D. Laidlaw. *Data Visualization Techniques (editor Bajaj, C.)*, chapter 'Continuous Bayesian tissue classification', pages 107–129. J. Wiley & Sons, Chichester, UK, 1999.

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

[48] D. H. Laidlaw, K. W. Fleischer, and A. H. Barr. *Handbook of Medical Imaging*, chapter 13, pages 195–211. Academic Press, San Diego, 2000. Chapter title: Partial volume segmentation with voxel histograms.

[49] J. D. MacDonald. *A Method for Identifying Geometrically Simple Surfaces from Three Dimensional Images*. PhD thesis, McGill University, Montreal, Quebec, Canada, 1998.

[50] J. H. Martin. *Neuroanatomy Text and Atlas*. Appleton & Lange, second edition, 1996.

[51] D. Mount and S. Arya. ANN: Library for approximate nearest neighbor searching. http://www.cs.umd.edu/∼mount/ANN/.

[52] W. J. Niessen, K. L. Vincken, J. Weickert, B. M. T. Romeny, and M. A. Viergever. Multiscale segmentation of three-dimensional MR brain images. *International Journal of Computer Vision.*, 31(2-3):185–202, Apr 1999.

[53] J. O'Rourke. *Computational Geometry in C*. Cambridge University Press, second edition, 1998.

[54] D. L. Pham and J. L. Prince. Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Transactions on Medical Imaging*, 18(9):737–52, Sep 1999.

[55] J. C. Rajapakse, J. N. Giedd, and J. L. Rapoport. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Transactions on Medical Imaging*, 16(2):176–186, Apr. 1997.

[56] C. Razafimahefa and M. Soss. Clustering by the use of minimal spanning trees (an interactive java applet). http://cgm.cs.mcgill.ca/∼soss/clustering/.

[57] J. D. Schellenberg, W. C. Naylor, and L. P. Clarke. Application of artificial neural networks for tissue classification from multispectral magnetic resonance images of the head. In *Proceedings of the Third Annual IEEE Conference on Computer-Based Medical Systems*, pages 350–357, Chapel Hill, North Carolina, June 1990.

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

[59] J. G. Sled. A non-parametric method for automatic correction of intensity non-uniformity in MRI data. Master's thesis, McGill University, Montreal, QC, May 1997.

[60] J. G. Sled and G. B. Pike. Standing-wave and RF penetration artifacts caused by elliptic geometry: an electrodynamic analysis of MRI. *IEEE Transactions on Medical Imaging*, 17(4):653–662, Aug. 1998.

[61] J. G. Sled and G. B. Pike. Understanding intensity non-uniformity in MRI. In W. M. Wells, A. Colchester, and S. Delp, editors, *1st International Conference on Medical Computing and Computer-Assisted Intervention*, number 1496 in Lecture Notes in Computer Science, pages 614–622. Springer, 1998.

[62] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A comparison of retrospective intensity non-uniformity correction methods for MRI. In *Proceedings of the 15th International Conference on Information Processing in Medical Imaging (IPMI)*, pages 459–464, Poultney, VT, USA, June 1997.

[63] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, Feb. 1998.

[64] M. Sonka, S. Tadikonda, and S. Collins. Knowledge-based interpretation of MR brain images. *IEEE Transactions on Medical Imaging*, 15(4):443 – 452, aug 1996.

[65] R. Stokking, K. L. Vincken, and M. A. Viergever. Automatic morphology-based brain segmentation (mbrase) from MRI-T1 data. *Neuroimage*, 12(6):726–38, Dec 2000.

[66] C. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977. (with discussion).

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

[68] The MathWorks Inc. Matlab. http://www.mathworks.com.

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

[70] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–96, Oct 1999.

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

[72] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. A statistical framework for partial volume segmentation. In *Quantitative Analysis of signal abnormalities in MR imaging for multiple sclerosis and Creutzfeldt-Jakob disease (PhD thesis)*. Katholieke Universiteit Leuven, Leuven, Belgium, may 2001.

[73] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis. Adaptive template moderated spatially varying statistical classification. In *Lecture Notes in Computer Science – Proceedings of Medical Image Computing and Computer-Assisted Intervention - MICCAI'98*, volume 1496, pages 431–438, oct 1998.

[74] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis. Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analysis*, 4(1):43–55., Mar 2000.

[75] W. M. Wells III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, Aug. 1996.

[76] M. X. H. Yan and J. S. Karp. An adaptive bayesian approach to three-dimensional MR brain segmentation. In Y. Bizais, C. Barillot, and R. D. Paola, editors, *Information Processing in Medical Imaging (IPMI)*, pages 201–213. Kluwer, June 1995.

[77] A. Zijdenbos, R. Forghani, and A. Evans. Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of insect. In *Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 439–448, Cambridge MA, USA, Oct. 1998.

[78] A. P. Zijdenbos and B. M. Dawant. Brain segmentation and white matter lesion detection in MR images. *Critical Reviews in Biomedical Engineering*, 22(5-6):401–65, 1994. 192 refs, Review.

[79] A. P. Zijdenbos, B. M. Dawant, and R. A. Margolin. Automatic extraction of the intracranial cavity on transverse MR images. In *Proceedings of the 11th International Conference on Pattern Recognition*, pages 430–433, The Hague, The Netherlands, Aug./Sept. 1992. IAPR.

[80] A. P. Zijdenbos, B. M. Dawant, and R. A. Margolin. Automatic detection of intracranial contours in MR images. *Computerized Medical Imaging and Graphics*, 18(1):11–23, Jan./Feb. 1994.