INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road. Ann Arbor. MI 48106-1346 USA
313/761-4700 800/521-0600

Performance Analysis of Automatic Techniques for Tissue Classification in Magnetic Resonance Images of the Human Brain

Vasken Kollokian

A Thesis
in
The Department
of
Computer Science

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Computer Science at
Concordia University
Montreal, Quebec, Canada

November 1996

CONCORDIA UNIVERSITY

Schoole of Graduate Studies

This is to certify that the thesis prepared

By : **Vasken Kollokian**

Entitled : **Performance Analysis of Automatic Techniques for Tissue Classification in Magnetic Resonance Images of the Human Brain**

and submitted in partial fulfilment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee :

_____ Chair

_____ Examiner

_____ Examiner

_____ Thesis Supervisor (Concordia)

Dr. Rajjan Shinghal

_____ Thesis Supervisor (McGill)

Dr. Alan Evans

Approved by _____

Graduate Program Director

_____ 1996 _____

Dean of Faculty

(Engineering and Computer Science)

Abstract

Classification of Magnetic Resonance (MR) images of the human brain into anatomically meaningful tissue labels is an important processing step in many research and clinical studies in neurology. The medical imaging research community is presented with a wide choice of classification algorithms from artificial intelligence and pattern recognition. This thesis describes the development of a controlled test environment, where different classification algorithms were implemented and their performance evaluated in a brain imaging context.

Furthermore, a mechanism for automating supervised classification algorithms is proposed through the use of *a priori* knowledge of neuro-anatomy, presented in the form of brain tissue probability maps. The results obtained through the automated methods compared favorably to those obtained through human supervision.

The performance of five supervised (Artificial Neural Networks, Bayesian, k-Nearest Neighbors, C4.5 decision tree, Minimum Distance) and two unsupervised (Hard C Means, Fuzzy C Means) classification algorithms is compared under varying conditions of MR imaging artifacts. The Artificial Neural Networks classifier was observed to be the best overall performer.

Résumé

La classification d' images obtenues par l'entremise de l'Imagerie par Résonance Magnétique (IRM) du cerveau humain est une étape importante en recherche de base et clinique en neurologie. Une grande variété d'algorithmes de classification basés sur l'intelligence artificielle et reconnaissance de forme sont offerts à la communauté scientifique en imagerie médicale. Cette thèse décrit le développement d'un environnement côntrolé pour comparer la performance des différents algorithmes de classification dans un contexte d'imagerie cérébrale.

Un mécanisme est proposé pour automatiser les algorithmes supervisés par l'utilisation d'une connaissance *a priori* de l'anatomie cérébrale, sous forme de modèle de probabilité. Les résultats obtenus avec ces méthodes automatisées se comparent favorablement à ceux obtenus sous supervision humaine.

Les performances de cinq méthodes supervisées (Réseaux de Neurones Artificiels, Bayesien, k-Plus-Proche Voisins, C4.5 Arbre de décision, and Distance Minimale) et de deux méthodes non-supervisées ("Hard C Means", "Fuzzy C Means") sont comparées dans des conditions de qualité d'images IRM variables. La méthode de Réseaux de Neurones Artificiels s'est montrée la plus performante.

Dedication

To my wife Seta, and to my parents for true love, support, understanding and tons of patience...

Acknowledgments

The work presented in this thesis would not have been possible without the priceless guidance, encouragement and patience of my thesis supervisors Drs. Alan Evans and Rajjan Shinghal. Their perseverance made the difficult task of writing this thesis, while holding a demanding job, more manageable. I would like to especially thank Dr. Evans (the boss), for his insistence on granting me a leave of absence from my duties as the systems administrator of McConnell Brain Imaging Center, without which this thesis would not have been completed.

I wish to thank Dr. Alex Zijdenbos for providing extraordinary help and insight in every aspect of my research; Dr. Louis Collins for great ideas and making sure that I did things right; Micheline Kamber for providing so much to build on; Dr. Noor Kabani for being very tolerant and helpful with all my neuro-anatomical requests, especially with the digital phantom and the training samples. I am grateful to my colleague Peter Neelin for providing so much software support, and brilliant answers to the most puzzling of questions; to Jong Lee for providing additional neuro-anatomical support and training samples; John Sled for expertise in all aspects of MR physics (especially the artifacts); David MacDonald for the wonderful software and libraries; Remi Kwan for a state-of-the-art MR simulator; Dr. Colin Holmes for being scanned 27 times; Jean-François Malouin for keeping the lab together during my absence; Greg Ward for always knowing the way to the answers for the International Consortium for Brain Mapping (ICBM) data; Mark Wolforth for Linux and LaTeX wizardry and friendly competition; Dr. Bruce Pike, Dr. Terry Peters and Rick Hoge for MR expertise; Sylvain Milot whose help and general understand-

ing of the issues has been great; Kate Hanratty for moral support and proof reading; Sean Marrett, Dr. David Reutens, Dr. Keith Worsley, Dr. Gabriel Léger, Paule-Joanne Toussaint, Dr. Michaela Sanielevici, Roch Comeau, Olivier Rousset and Yilong Ma for valuable suggestions and support.

I would like to acknowledge the enormous computational resources provided by the McConnell Brain Imaging Center of the Montreal Neurological Institute, the ICBM and AutoImmune projects for providing ample data and disk space for the experiments.

I would like to thank my parents, my sister and her family for their support and understanding, without which this thesis would not be.

Finally, I would like to express my deepest gratitude to my wife, Seta, for love, support and caring. Without her patience, optimism, encouragement and psychology (and insisting on locking me up in my room, time after time), I would have never been able to finish.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Magnetic Resonance Imaging (MRI) is a medical imaging technology that provides anatomical, diagnostic and functional *in vivo* information about the human body in a non-invasive manner. MRI has been useful in the field of neurology, especially in imaging the brain and the spinal cord, where its unparalleled soft tissue contrast has been instrumental in providing information about brain structure both in diagnosis and in a wide variety of research and clinical studies.

1.1 What is tissue classification?

MR images of the brain are acquired in a three dimensional (3D) volumetric data set, where each brain image volume is a stack of two dimensional (2D) slices, and each slice is a set of picture elements or pixels (called voxels in 3D), of image intensities representing underlying anatomy. The process by which an image is delineated into distinct regions is known as **segmentation**, and the assignment of anatomically meaningful labels to each region is known as **classification**. Often in MR image processing literature, these terms have been used interchangeably. In this thesis, the term *classification* will mean the process of assigning each voxel in a brain image volume to one of three major tissue types in a healthy brain: gray matter (GM), white matter (WM), and cerebro-spinal fluid

(CSF). The latter is a fluid that surrounds the brain, and fills the *ventricles*, which are cavities deep inside brain structures [Carpenter, 1985].

Although the cerebro-spinal fluid is not a tissue type, it is considered a constituent of the brain. Normally, an image volume covering an entire head, contains more tissue types than mentioned above. These include muscle, fat, skin, skull and other connective tissues. Moreover, large areas in an image volume contain air, especially outside the head and in the sinuses. Since air and the above mentioned tissue types fall outside the brain, they are masked out in pre-processing stages that will be further described in coming chapters. Figure 1.1 shows a sample MR image, with the corresponding classified image.

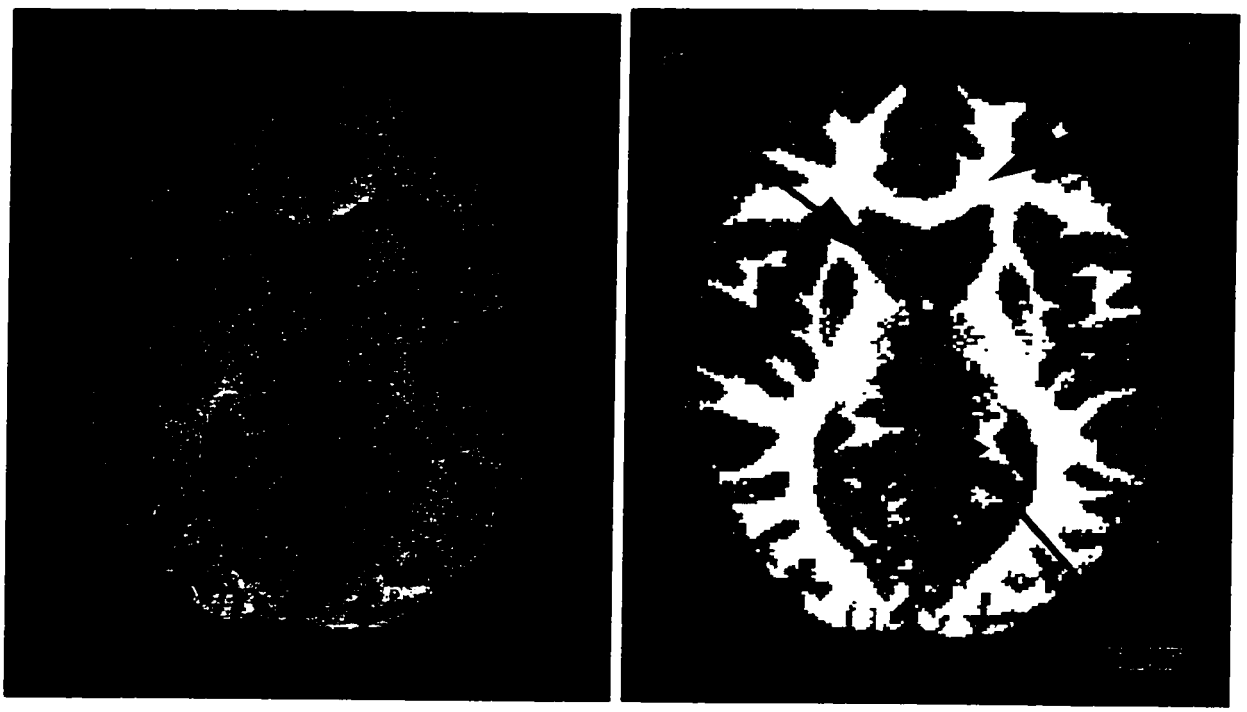


Figure 1.1: a) A sample MR image, b) corresponding classified MR image.

In the case where pathologic brains need to be considered, respective diseased tissues, such as lesions and tumors, could be identified and added to the above tissue types. Since this thesis will be dealing with only healthy brain volumes, the tissue types being considered will be limited to CSF, GM, and WM. Whenever voxels in the brain volume cannot be assigned to any of the 3 mentioned classes, they will be labeled as the background

(BCK).

1.2 Importance of tissue classification

Classification of brain image volumes into their constituent tissue classes is a necessary step in many brain-research areas. The calculation of specific tissue volume in particular, and brain size in general, is part of basic and clinical neuroscience research. For example, classification is used to:

- Quantify brain atrophy, which is a reduction in gray matter volume in certain types of patients; where it is necessary to calculate gray matter volume in the temporal lobes, [Jack et al., 1988; Jack et al., 1990; Cendes et al., 1993; Lee et al., 1995; Kabani et al., 1996].
- Help with the evaluation of certain diseases like Multiple Sclerosis [Kamber et al., 1992; Kamber et al., 1995].
- Evaluate effects of drug therapy on lesion or tumor load [Cline et al., 1987].
- Build tissue (healthy or otherwise) probability maps [Evans et al., 1994; Kamber et al., 1995].
- Generate phantoms for MRI and Positron Emission Tomography (PET) simulation studies [Ma et al., 1993; Rousset et al., 1993; Kwan et al., 1996].

Classification is also an important pre-processing step in validating the registration of multi-spectral and multi-modal brain volumes [Neelin et al., 1993]. For all these applications, it is vital that the classification process be accurate, reliable and reproducible.

The field of pattern recognition has provided the research community with numerous classification algorithms. Most of these require user intervention (supervised classifiers), while others function autonomously (unsupervised classifiers), each algorithm having its

respective advantages and disadvantages. The automation of the classification process eliminates the need for user intervention, making it more reproducible. This thesis describes a mechanism for automating classification algorithms in a brain imaging context, and quantitatively evaluates their comparative performance under varying conditions of image degradation.

1.3 Issues facing tissue classification

There is a large amount of literature in the field of MR image classification and Chapter 3 provides an extensive literature review. However, it is important to mention that there are a number of problems facing medical image processing in general and brain tissue classification in particular. This thesis addresses some of the issues introduced below in greater detail in the coming chapters.

1.3.1 Variable image intensities

MR imaging produces different levels of contrast when using different MR scanners or the same MR scanner at different times, affecting the quality of the acquired images. Intra-subject (within the same brain volume) intensity normalization is needed to obtain consistent classification results across brain slices. Inter-subject (between different brain volumes) intensity normalization is required when standardized intensity values need to be used.

1.3.2 Subject or patient motion

MRI is a multi-spectral imaging technology (Section 2.3) that can provide images with different contrasts of the same underlying anatomy, thereby increasing the discriminating power of tissue classification algorithms. To obtain these multi-spectral images, subjects are scanned with different imaging parameters. Motion artifacts are therefore introduced,

causing mis-alignment of images. In order to fully exploit the discriminating information provided by the multi-spectral images, these motion artifacts need to be eliminated.

1.3.3 Image degradation due to noise

All current MR scanners have numerous deficiencies that affect image quality in a variety of ways (Section 2.4). Noise in MR images results primarily from two sources: **1)** due to magnetic field inhomogeneities of the MR scanner and thermal fluctuations in the electronic circuitry of the imaging hardware, and **2)** imaging parameters like acquisition time, total scan time and spatial resolution [Nishimura, 1993]. Chapter 2 discusses the principles of MR imaging and the sources of image degradation.

1.3.4 Radio-Frequency (RF) inhomogeneity

Imperfections in the imaging hardware like the receiver and gradient coils (Section 2.2.2) results in local variations in intensity of similar tissue types across different parts of the brain. For example, white matter tissue appears much brighter in the anterior part of the brain than it does in the posterior part. Since algorithms rely on intensity information to classify unknown voxels, variations in intensity impede accurate classification as discussed in greater detail in Section 2.4.2.

1.3.5 Partial volume effect

Owing to the finite resolution of the MR scanner, signals from different adjacent tissue types are mixed together in a single voxel. This results in blurred boundaries between the tissues, reducing the accuracy of the classification algorithms. For example, an algorithm may incorrectly classify a voxel as gray matter, when in fact the signal intensities from both cerebro-spinal fluid and white matter tissue types are mixed together to produce a voxel intensity similar to that of gray matter.

1.3.6 Neuro-anatomical variability

Classification algorithms need training samples, which are examples of a specific tissue's MRI intensity obtained from representative voxels in each tissue. Traditionally, these training samples are provided manually by experts and are specific to individual brains. Since it may be necessary to classify a large number of brain volumes, manual training sample selection becomes a tedious and time consuming task. Because of neuro-anatomical variability (caused by individuals having different sized and shaped brains), image intensity and quality variations, automating this training becomes a difficult process. This thesis describes a model-based approach to automate tissue classification, which takes into consideration neuro-anatomical variability in a probabilistic sense (Chapter 4).

1.3.7 Manual classification

Numerous studies seek to establish statistical significance in testing certain hypotheses by processing a very large number of brain volumes. Manual tissue classification by neuro-anatomists is a tedious, difficult, time consuming, subjective and not very reproducible task because of degradation in image quality. Furthermore, poor eye-hand coordination, fatigue and less than optimal man-machine interface (e.g. mouse-driven operation) makes it impractical to resort to manual means to achieve accuracy, reliability and reproducibility. This thesis describes the development of a completely automated classification process that provides accurate, reliable and consistent results without the need for human intervention.

1.3.8 Quantitative vs qualitative validation

In order to compare the performance of different classifiers, numerous researchers rely on qualitative expert opinion as a validation mechanism. However, such visual assessment lacks the objectivity of quantitative methods needed to perform critical assessment of different algorithms. Validation methods will be discussed in greater detail in Chapter 5.

1.3.9 Wide choice of classification algorithms

Extensive pattern recognition and image processing literature exists in the field of medical imaging in general, and in MR imaging in particular. As research advances are made, more sophisticated algorithms become available, which raises the question of optimal functionality (which one is better?). Chapter 3 presents a more detailed literature review of current medical imaging research and describes the classifiers implemented in this thesis, pointing to their strengths and weaknesses.

1.4 Objectives and scope of the thesis

Problem : Since there are numerous issues facing tissue classification and a wide choice of classification algorithms, it is necessary to understand how the varying conditions of some of the above mentioned MR imaging parameters affect the performance of different classifiers: which parameters, and under what conditions, are least affected and perform optimally, and which ones are most affected and perform less optimally. Moreover, the need to classify a large number of brain volumes requires automation of the training sample selection process, while ensuring results similar to those obtained through expert manual training.

1.4.1 Objectives

The above requirements lead to the following specific objectives:

- Create a versatile evaluation environment (black box), where different classifiers can be incorporated and interchanged easily.
- Create tissue probability maps to automate the training sample selection process and provide *a priori* 3D spatial information to certain classifiers.

- create a controlled environment using an MRI simulator (Section 5.3.5), where systematic search can be conducted to investigate the issues facing tissue classification (Section 1.3).
- Systematically compare the performance of the following classifiers: Minimum Distance, Bayesian, k-Nearest Neighbors, Artificial Neural Networks, C4.5 decision tree, Hard C-Means and Fuzzy C-Means.
- Provide an answer to the question "which of the above mentioned classifiers performs best?"

The **scope** of this thesis will be limited to the following constraints:

- Deal only with the principal tissue classes in a normal brain volume: gray matter, white matter and cerebro-spinal fluid.
- Limit the MR search parameters to Noise and RF inhomogeneity
- Limit the search space of voxel resolution (partial volume) to slice thickness, i.e. all 2D images have the same pixel dimensions (181x217).

1.4.2 Desirable characteristics of the classification environment

- The training and classification method should be accurate, reproducible, reliable, simple to use, and require no user-interaction.
- The method should be modular, where each section of the classification process can be improved independently.
- The classification environment should be versatile enough to incorporate new classifiers with simplicity and ease.
- The method should operate in *Standard Brain Space*, where tissue probability maps can be used to automate the classification process.

- The method should have an unbiased objective validation scheme to verify the classification results.

- The method should be general, allowing for the seamless inclusion of additional features from other imaging modalities like PET and Computed Tomography (CT).

1.5 Thesis overview

Chapter 2 introduces the principles of Magnetic Resonance Imaging, and the sources of MR image degradation (artifacts). Chapter 3 then presents the classifier methodologies and surveys current literature about brain tissue segmentation from a large number of researchers in this field, emphasizing and explaining the theoretical basis of classifiers tested in this thesis. Brain tissue probability maps and the automation of training samples selection are discussed in Chapter 4. Chapter 5 examines the issues of validating classified MRI volumes. Chapter 6 describes the motivation and the creation of the *Simulated Brain Database*. Experimental results and discussion are presented in Chapter 7. Finally, concluding remarks and future work in this field are given in Chapter 8.

Chapter 2

Magnetic Resonance Imaging

2.1 Introduction

Since the discovery of X-rays by Roentgen in 1895, there have been significant developments in medical imaging technologies which permit the visualization of the interior of the human body without the need for surgical intervention. Today, medical imaging methodologies such as X-ray, Computed Tomography, Positron Emission Tomography, and Magnetic Resonance Imaging provide images of the interior of human body from both anatomical and metabolic perspectives. Because of its flexibility in measuring a wide range of physical parameters, MRI has played an important role in imaging the human body, especially the central nervous system, where its spatial resolution, superior soft tissue contrast and anatomical detail has placed it at the forefront of available imaging modalities. This chapter will review the basic physics of MRI, emphasizing its multi-spectral nature. Thereafter, it will briefly discuss sources of MR image quality degradation (artifacts), and show sample MR images of different modalities that have been used for this thesis.

2.2 Principles of MRI

2.2.1 Basic theory

Atomic nuclei with an odd number of protons and/or neutrons possess nuclear momentum. These charged nuclei, which in the MRI literature are referred to as *spins*, have small magnetic moments. Magnetic resonance imaging involves the manipulation of these spins under the influence of several different externally applied magnetic fields [Nishimura, 1993; Philips, 1984].

Since biological tissues are rich in water, the hydrogen atom (in the water molecules) with its single proton, is the body's most abundant nuclide, and hence makes an excellent candidate for MR imaging. In this section, the physical phenomenon upon which MRI is based, will be briefly described. For more in depth discussion, other texts by [Nishimura, 1993; Sprawls, 1992; Plewes and Bishop, 1992; Allen, 1992] should be consulted.

In the absence of any external magnetic field, the hydrogen atoms are oriented randomly with a net magnetic moment of zero magnitude (Figure 2.1a). When placed in a strong 1.5 Tesla (Tesla = 10000 Gauss) static magnetic field, denoted B_0, two effects arise immediately:

- Two groups of spins arise. The first group aligns itself in a parallel direction as the static magnetic field B_0, and the second aligns itself in an anti-parallel direction. The ratio between the first and second groups is temperature dependent, and after equilibrium is reached, is usually 0.999993. The excess of nuclei in the first group will produce a net magnetization, M, oriented in the same direction as the main field (Figure 2.1b).

- The spins precess at a well-defined frequency called the **Larmor Frequency** ω (Figure 2.2) defined as:

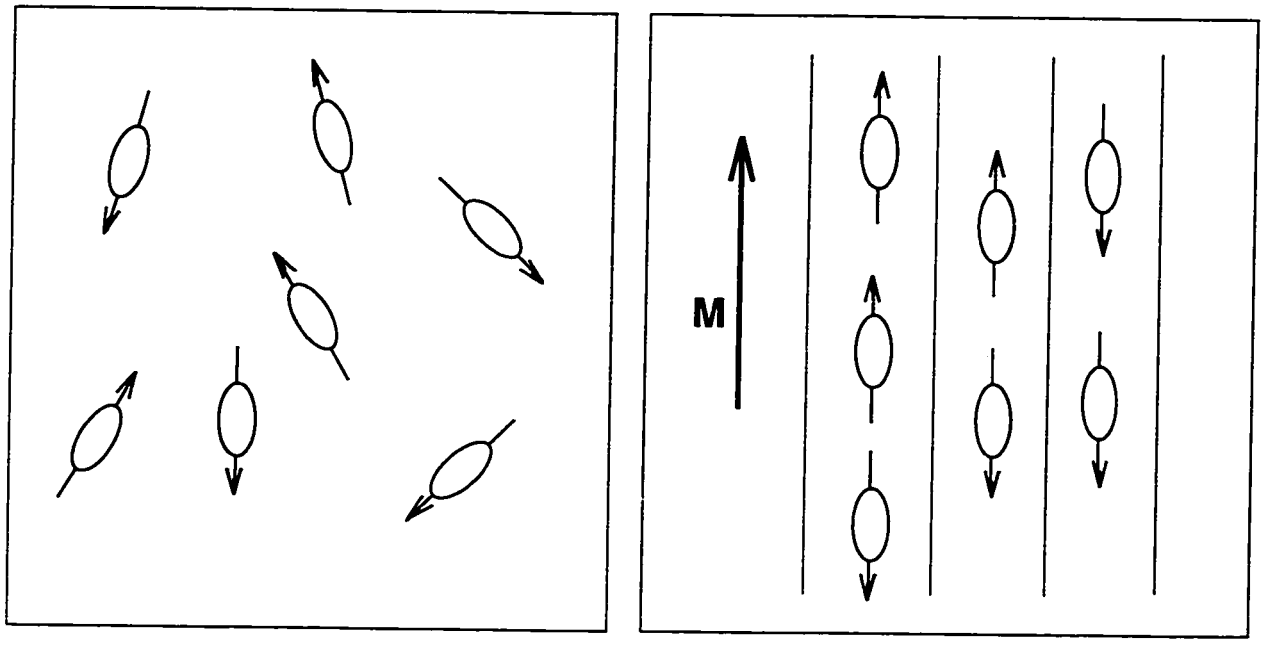


Figure 2.1: (a) Thermal equilibrium resulting in zero net magnetization, (b) Net magnetization resulting from the influence of a static field B_0.

$$\omega = \gamma B_0 \tag{2.1}$$

where γ, the gyro-magnetic ratio, is dependent on the type of nuclei. For example, in a 1.0 Tesla magnetic field, hydrogen nuclei have a Larmor frequency of 42.6 MHz, and at 1.5 Tesla the Larmor frequency becomes 63.9 MHz.

An analogy can be made between the Larmor frequency of specific nuclei and identical tuning forks. When a pair of identical tuning forks are placed side by side, and when the first tuning fork is struck, the vibrations produced make the second fork vibrate at the frequency of the first. Similarly, when a pulse of electro magnetic radiation at the Larmor frequency (typically in the radio frequency range) is imparted to a collection of nuclei in equilibrium in a strong magnetic field, they absorb this energy. Depending on the orientation and the duration of the Radio Frequency (RF) pulse, the nuclei deviate from

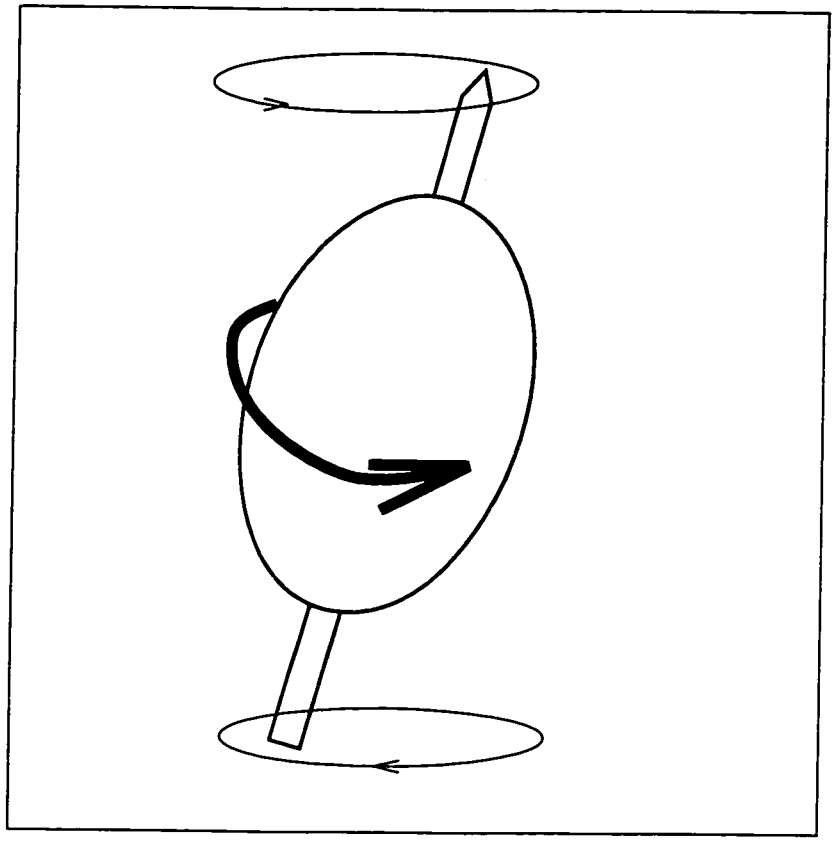
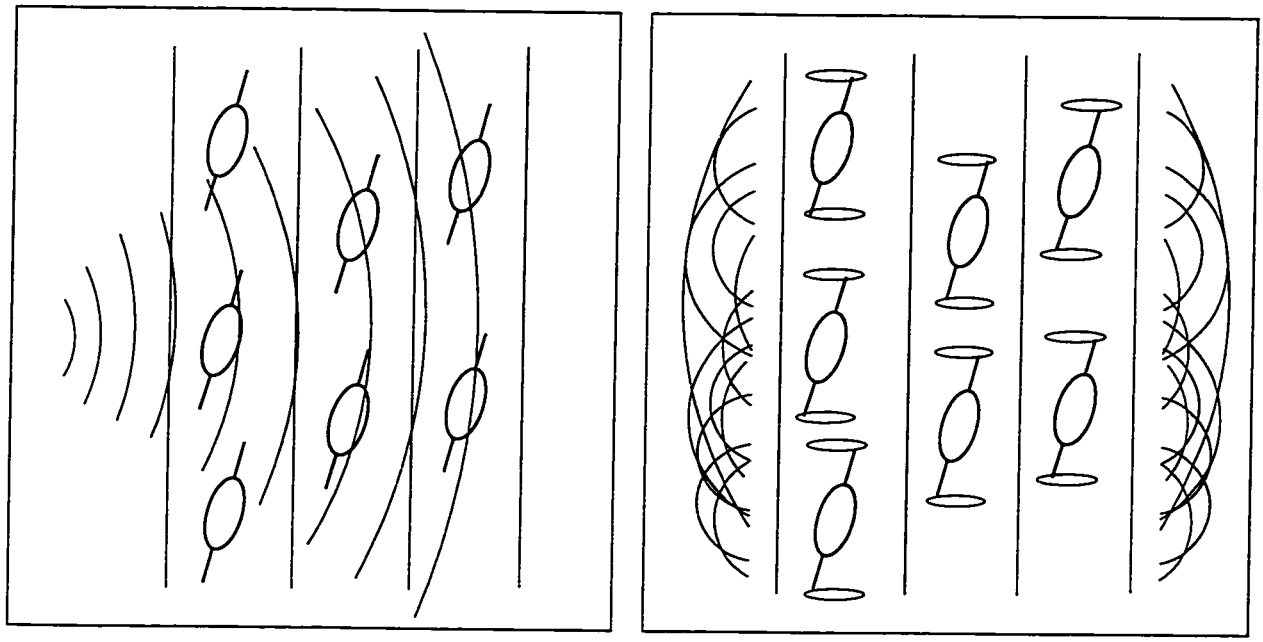


Figure 2.2: Precession of nuclei at Larmor frequencies when excited out of equilibrium.

their equilibrium state (Figure 2.3a). As soon as the energy pulse stops, the nuclei start to relax and reach their original equilibrium state, while emitting RF energy at the same frequency that it was absorbed (Figure 2.3b). Measuring this signal at each specific location in a regular pattern and reconstructing the data into an image is the principle behind Magnetic Resonance Imaging, as detailed below.

2.2.2 Imaging hardware

Figure 2.4 shows the block diagram of an MR imaging system. It consists primarily of a large super-conducting magnet (cooled to 4° Kelvin to eliminate electrical resistance), which is used to impose a strong and uniform main magnetic field B_0. Also, in order to deliver RF energy pulses and measure the emitted signal, RF transmission and reception



(a) (b)

Figure 2.3: (a) RF energy pulse tilting nuclei out of equilibrium, (b) Tilted nuclei returning to equilibrium, emitting RF signals in the process.

coils are used, which in some cases are the same set of coils. After the coils transmit an RF energy pulse by delivering voltage through the RF amplifiers, they switch to receiving mode and act as antennae to measure the electro magnetic signal emitted by the nuclei under study.

Since RF coils encompass the entire region of interest, it is not usually possible to differentiate signals from different spatial locations. In order to spatially encode RF signals, linear gradient fields in all three orthogonal axes G_x, G_y, G_z are applied in addition to the main magnetic field B_0, so that the frequency of the spins become a function of location in x, y, z (remember Larmor frequency is dependent on the strength of the main magnetic field as shown in Equation 2.1), thus continuous variations in the applied field, generated by the x, y, z gradient fields, provide the necessary spatial encoding. The gradient fields are created through a set of gradient coils and amplifiers. These subsystems are connected to several computer systems that control events in the imaging sequence,

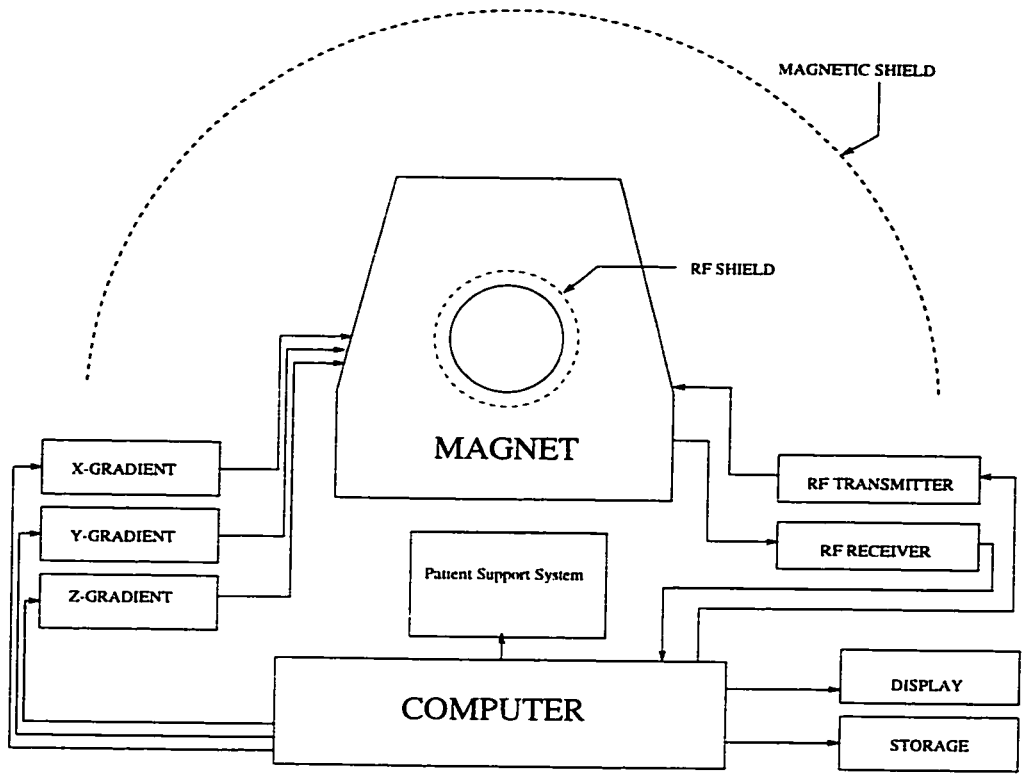


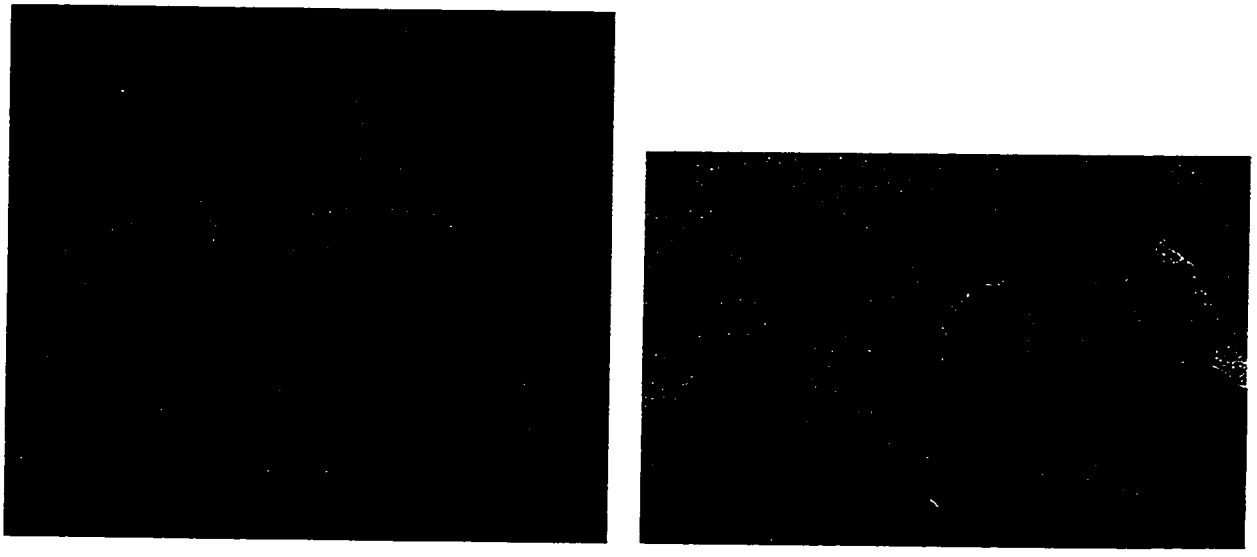
Figure 2.4: Block diagram showing the main components of an MR imaging system.

acquire and process data and, finally, display results. Figure 2.5a shows a photo of an MR scanner and its control console, and Figure 2.5b shows a patient being prepared for a scan.

2.2.3 Equilibrium, excitation, and relaxation

At any time, the net magnetization vector M can be decomposed into two component vectors, M_z in the direction of the main magnetic field, also known as the longitudinal component, and M_{xy} in the plane perpendicular to the main magnetic field, also known as the transverse component (Figure 2.6). It is important to note that in an equilibrium state, the magnetization vector M, has no transverse component M_{xy}, and the magnitude of M is the same as M_z (Figure 2.7).

By delivering an RF pulse at the proper Larmor frequency, this magnetization vector



(a) (b)

Figure 2.5: (a) Photograph of an MR scanner (b) Patient being prepared for a scan.

M can be tilted away from equilibrium. The degree of tilt depends on the intensity and duration of the RF pulse. The spins are said to be excited out of their equilibrium state. In effect, the transverse magnetization M_{xy}, increases in magnitude, and starts precessing around the main magnetic field B_0, while the longitudinal magnetization M_z decreases in magnitude. After the RF pulse stops, the excited nuclei return to their equilibrium state. This process is known as *relaxation* during which they emit electro-magnetic radiation, which is the source of the MR signal, recorded through the RF receiver coils.

The longitudinal magnetization vector M_z returns to its equilibrium state at an exponentially decreasing rate to assume its original value M. The time it takes to reach equilibrium is known as the *Longitudinal Relaxation Time* T_1, depicted in Figure 2.8. On the other hand, the transverse magnetization vector M_{xy} decays exponentially due to interactions among spins, to reach equilibrium value of zero in a much shorter time period. The time it takes for M_{xy} to decay completely is known as the *Transverse Relaxation Time* T_2, depicted in Figure 2.9. The longitudinal and transverse relaxation times are shown graphically in Figure 2.10. The values of these relaxation times T_1 and T_2 depend

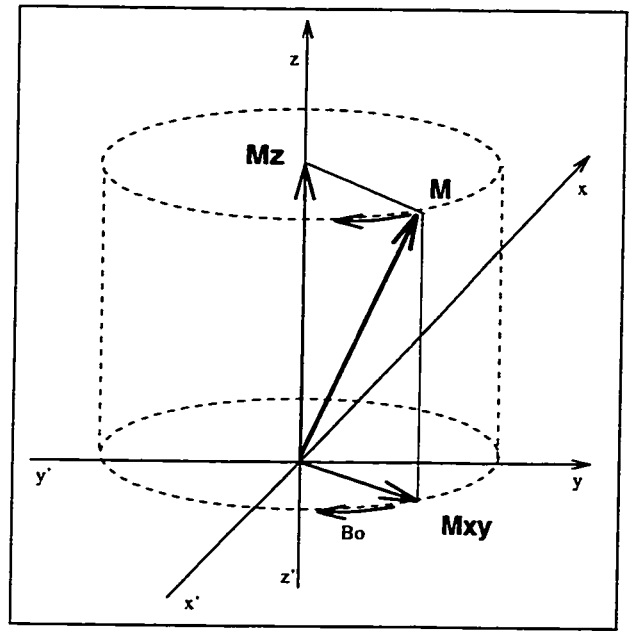


Figure 2.6: Diagram showing longitudinal (M_z) and transverse (M_{xy}) vector components of net magnetization vector M.

on molecular structures of the substances being imaged and their physical states (solid or liquid). For example, T_1 in biological tissue ranges from 50 milliseconds to a few seconds; it is shorter in liquids than in solids. T_2 time is in the range of 40 milliseconds to one second; it is much shorter in solids (microseconds) than in liquids (milliseconds). Therefore different tissue components having distinct molecular structures and composition will, in effect, have different relaxation times, thus making it possible to distinguish them using MRI.

2.3 Multi-spectral nature of MRI

As mentioned earlier, to measure MR signals, nuclei are excited out of equilibrium through a sequence of RF pulses of varying intensity and duration. As they return to their equilibrium state, they emit an RF signal, which is the measured MR signal. This sequence of events (equilibrium, excitation and measurement), is repeated several times to resolve

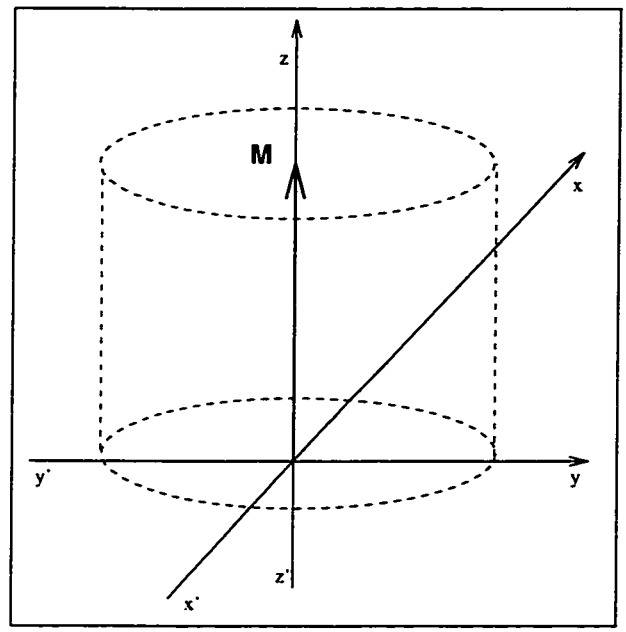


Figure 2.7: Diagram showing no transverse component of net magnetization vector M in magnetic equilibrium.

different parts of the brain volume and to increase the Signal to Noise Ratio (SNR) in the presence of random noise.

Different types of RF pulses can tilt the magnetization vector M by different angles and different directions. The degree of tilt depends on the duration of the RF pulse, and can be set to any arbitrary value. For example, an RF pulse that tilts the magnetization vector by 90 degrees is known as a 90° pulse. The interval at which a sequence of RF pulses is repeated is known as the Repetition Time (TR), and the time at which MR signals are measured relative to the start of the sequence is known the Echo Time (TE). By using various tilt angles (commonly referred to as flip angles) in sequence at specific TE and TR values, numerous pulse sequences can be designed to maximize the contrast between different tissues. Figure 2.11 shows a schematic representation of TE and TR in one such pulse sequence frequently used in MR imaging.

Since contrast between different tissues in an MR image is due to the difference in their tissue specific parameters (like T_1 and T_2), the relative "weight" of these parameters

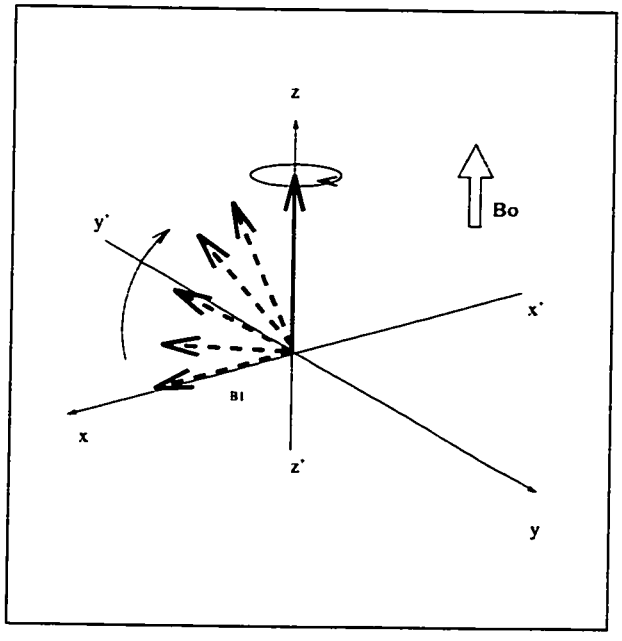


Figure 2.8: Diagram showing sequence of dotted vectors, representing stages of longitudinal relaxation.

is controlled with different pulse sequences. These sequences will produce T_1- or $T2-$ weighted images, indicating that the tissue contrast in the image is mainly due to T_1 or T_2 tissue parameters respectively, since signal intensity is rarely due to one tissue parameter [McVeigh and Atalar, 1992]. This makes MR imaging multi-spectral in nature, where several different gray-scale images, of the same underlying anatomy, are obtained by different imaging sequences.

2.4 Sources of MR image degradation

All current MR imaging systems have shortcomings that result in image degradation in one form or another. In this thesis, the three main sources of image degradation (artifacts) will be considered: image noise, RF inhomogeneity and partial volume.

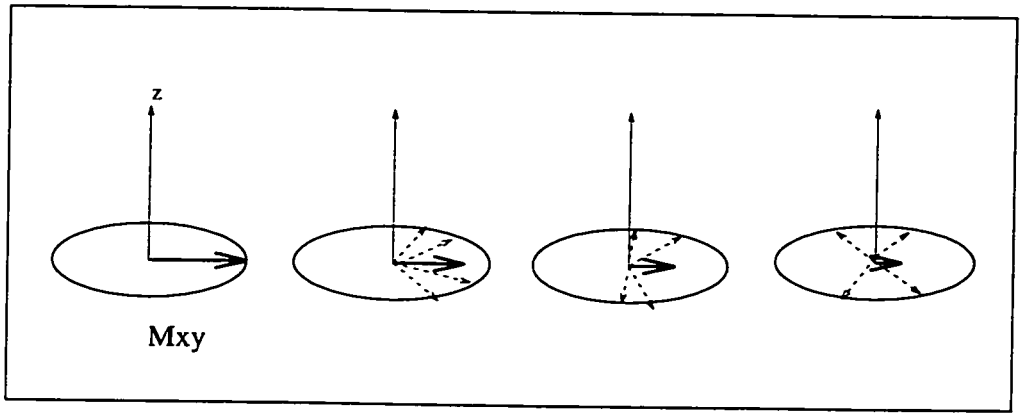


Figure 2.9: Sequence of diagrams showing stages of transverse relaxation.

2.4.1 Image noise

The noise in MR images stems from several sources. One source is thermal fluctuations in the electronic circuitry of the imaging hardware. It contributes to variations in the signal being measured. Also, thermal activity within the tissue appearing as random RF energy contributes to the signal of the body being imaged, thus producing variations in voxel intensities. The main source of noise in MR is Gaussian distributed and additive [Nishimura, 1993]. The constant presence of noise is one of the limitations on MR image quality.

2.4.2 Radio-Frequency inhomogeneity

Owing to design imperfections in the RF coil, the pulses applied to excite the spins out of their equilibrium may not be uniform across the entire field of view of the object being imaged. Non-uniformity also results from subject induced field variations [Simmons et al., 1994]. These inhomogeneities ultimately result in variations of intensity of similar tissue types across the field of view of the volume. Thus, a particular tissue appears brighter in some locations and darker in others and this variation can be modeled as a slowly-varying multiplicative field. Although slowly-varying intensity changes do not influence image analysis by inspection, they may pose drawbacks on automated image analysis.

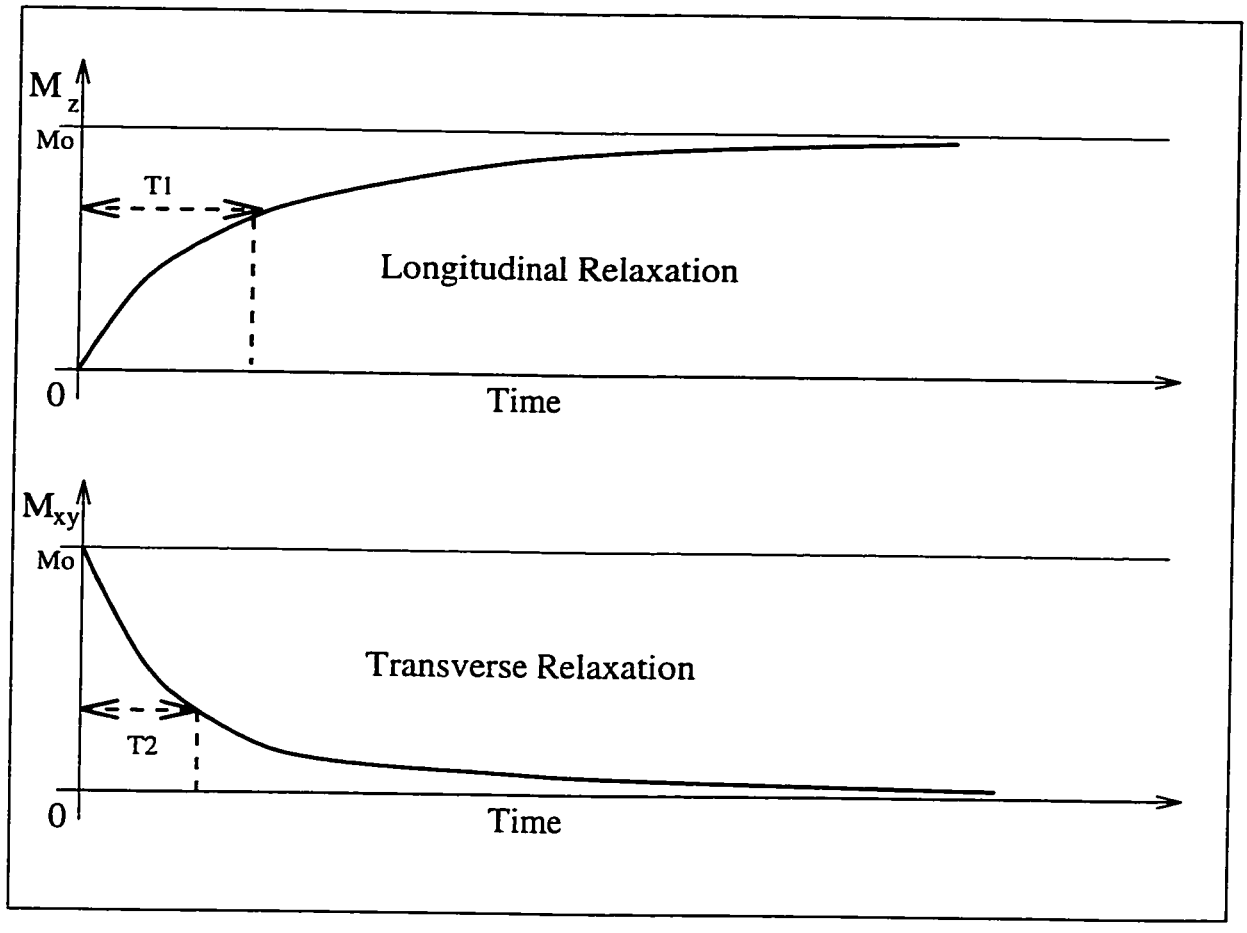


Figure 2.10: Graphic representation of longitudinal and transverse relaxation. M_0 is the magnetization at equilibrium.

2.4.3 Partial volume effect

Image voxels represent discrete tissue samples that are being imaged by an MR system. The size of the voxel determines the resolution of the MR image. The smaller the voxel, the higher the resolution of the acquired image, thus providing greater anatomical detail at a cost of longer acquisition time. There are practical limitations to the size of the voxels that can be imaged in an MR system. Furthermore, owing to the mix of different cellular densities at a microscopic level, the transition from gray matter to white matter tissue types is smooth and gradual, rather than abrupt.

Partial volume refers to the situation where, depending on the size of the voxels, more

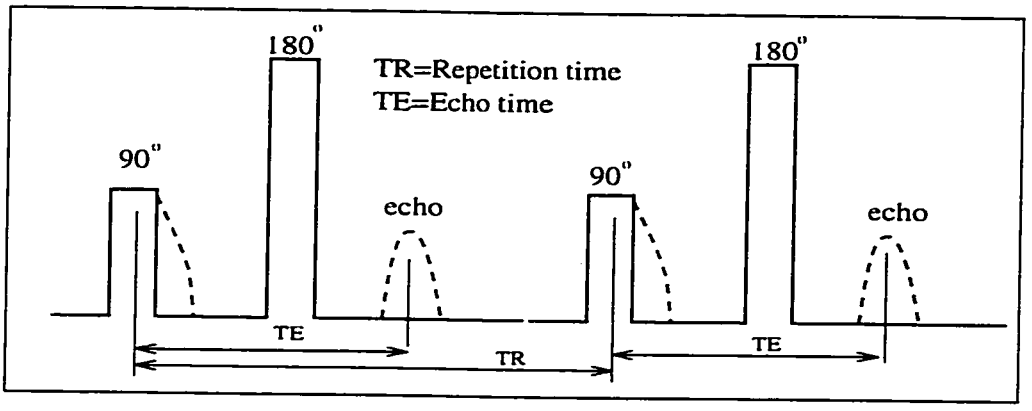


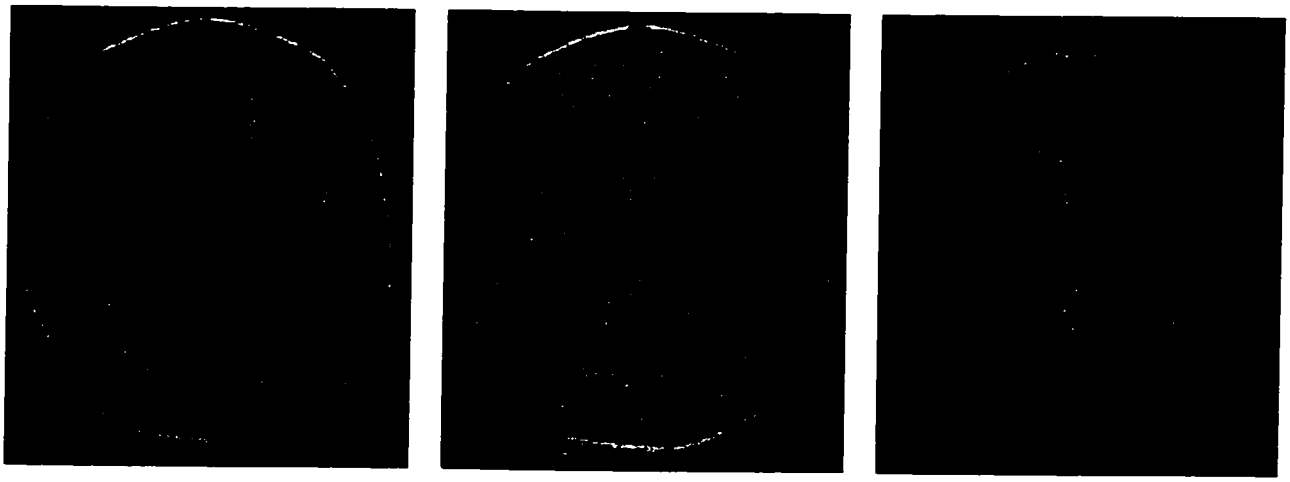
Figure 2.11: Diagram showing flip angles, repetition time (TR), and echo time (TE) of a typical MR imaging pulse sequence.

than one tissue type are mixed together in a single voxel, the intensity of which is a mix of the different tissue types. It is common practice in clinical settings to acquire image slices ranging in thickness over several millimeters. Increasing slice thickness effectively increases the voxel size (in one dimension), thus decreasing resolution and creating partial volume effects. The extent of this partial volume is directly proportional to the slice thickness.

2.5 Sample images

In this thesis, three different MR imaging modalities will be used as multi-spectral features in tissue classification. These are T_1-weighted images, T_2-weighted images and Proton Density (PD) weighted images, which reflect the amount of nuclei present in the body being imaged. All the images were acquired on a Philips Gyroscan ACS II scanner with a 1.5 Tesla super-conducting magnet.

Figure 2.12a shows an example of T_1-weighted image, where TE and TR values were 10 and 18 milliseconds respectively, and the flip angle was 30°. Figures 2.12b 2.12c are examples of dual-echo (TE values of 35 and 120 milliseconds) PD- and T_2-weighted images respectively, acquired at a TR value of 3300 milliseconds and flip angle of 90°.



(a) (b) (c)

Figure 2.12: (a) Sample T_1-weighted image acquired at TE of 10ms, TR of 18ms, and flip angle of 30° (2) Sample PD-weighted image acquired at TE of 35ms, TR of 3300ms, and flip angle of 90° (3) Sample T_2-weighted image acquired at TE of 120ms, TR of 3300ms, and flip angle of 90°.

2.6 Concluding remarks

Understanding the basic physics of MR imaging, plays an important role in appreciating the flexibility it provides for producing multi-spectral images. This is especially true for brain tissue classification, where new pulse sequences designed to produce different tissue contrasts could be used to add to the feature dimensionality of the images being classified, thus potentially augmenting the discriminating power of the classification algorithms to detect healthy and pathological tissue types for basic and clinical research.

Chapter 3

Review of Classifier Theory

3.1 Introduction

This chapter reviews some of the research carried out in MR image processing and pattern recognition, giving a survey of different classification techniques used in MRI. Emphasis is placed on classifiers implemented and tested in this thesis, explaining the general principle and theory behind each one, followed by a literature survey of other work that relies on these classifiers.

3.2 Literature survey

There is a large body of information already published in the area of image classification in MRI, especially of the central nervous system. Considerable effort has been put in examining different classification techniques in normal and pathological cases, and extensive comparative studies have been carried out by several researchers [Hall et al., 1992; Clarke et al., 1992; Clarke et al., 1993; Clarke et al., 1995; Vaidyanathan et al., 1995]. Since it is beyond the scope of this thesis to do an all encompassing survey of MR image classification techniques, the primary focus of this review chapter will be limited to classifiers

implemented in this thesis. For a comprehensive coverage, the reader is referred to other critical review papers [Zijdenbos and Dawant, 1994; Clarke et al., 1995].

In order to classify MR images, a certain number of "features" representing anatomy have to be used. In most cases, voxel intensities have been considered to represent features of underlying anatomy [Hall et al., 1992; Cline et al., 1990; Clarke et al., 1993; Taxt et al., 1992; Gerig et al., 1992; Vannier et al., 1991]. Some others have calculated additional features from voxel intensities, such as textures [Ehricke, 1990].

Since voxel intensities can come from a single image or from several images, MR image features can be divided broadly into two groups: *gray scale*, also known as single contrast, where a single gray scale image is used in the classification process, and *multi-spectral*, where several different gray scale images of the same anatomy having different tissue contrasts are considered. There are numerous studies that have used single gray scale images for classification, and the techniques used are quite varied. The most intuitive and simplest is *intensity thresholding*. In this method, a global threshold value is determined, usually by interactive methods [Lim and Pfefferbaum, 1989], where voxels are assigned to different classes based on their gray scale intensity. The advantage of this system is that it provides an easy and fast method of classifying image data. However, to function optimally, a global threshold value has to be determined that will hold across the whole image volume.

Other single contrast methods include *edge-detection*, which involves a gradient operator that detects sharp changes in image intensity, denoting an anatomical boundary or edge. Another technique is *region growing*, which involves planting a seed, often interactively by an operator, in the image data, thereby grouping voxels of similar intensities together in a region. The fundamental drawback that affects all of the above methods is not only their operator dependency for optimum functionality, but the assumption of uniform tissue intensity distribution throughout the entire volume. Given that MR imaging suffers from a number of artifacts (Section 2.4), like RF inhomogeneity, this assumption is rarely valid.

As mentioned previously in Chapter 2, MR imaging is *multi-spectral*: numerous images of varying contrast levels of the same anatomy are obtained by different pulse sequences. These are modifiable acquisition parameters of the MR scanner that yield different frequency images at each spatial location [Vannier et al., 1985; Vannier et al., 1987; Vannier et al., 1988; Vannier et al., 1991]. These frequencies can be grouped together in D-dimensional feature vectors that can be used by pattern recognition methods to produce better classification results compared to those obtained from single gray scale images [Bezdek et al., 1993; Hall et al., 1992; Just and Thelen, 1988; Vannier et al., 1985; Vannier et al., 1987; Vannier et al., 1988; Vannier et al., 1991]. The D-dimensional feature vectors yield a multi-dimensional measurement space called *feature space*. Since the features are image intensities represented by real numbers, the feature space is $\Re^D$, and classification is accomplished by partitioning this feature space into distinct regions representing the classes of interest. [Schalkoff, 1992].

The multi-dimensionality of MR images due to longitudinal and transverse relaxation times and proton density provides tissue characterization that can be very helpful for the classification of tissues. As new pulse sequences are developed, these can be added to the already existing D features to yield better discriminating ability to tissue classification. Different pulse sequences are also helpful in enhancing pathological cases [Taxt et al., 1992]. Of course all this comes at a price, since adding more spectral signatures translates into longer scanning time, thus causing more discomfort to the patient, or subject. In addition, the possible head movement between scans results in mis-aligned data sets, more noise and image artifacts.

Throughout the literature, great emphasis is placed in differentiating **supervised** and **unsupervised** classification methods. Since supervised methods require operator intervention for selecting a training set, they are considered less reproducible than unsupervised methods, which do not involve operator intervention except at the final stages of labeling clustered data. Both methods have been discussed extensively in the literature.

3.3 Supervised classification methods

Supervised classifiers need to be supplied with a training set produced by an expert, who selects training samples from image data sets. This training set takes the form of samples on a voxel or region basis, chosen from the image, or outlines of clusters in scatter plots in feature space. Scatter plots are multi-dimensional histograms, where each axis represents an intensity feature. In multi-spectral imaging systems like MRI, a number of features can be used to construct 2D or 3D scatter plots. Figure 3.1 shows such a scatter plot of T_1-, T_2- and PD-weighted images.

These scatter plots reveal the characteristic signatures of different tissues in feature space. Clusters found in these plots help mark the decision boundaries that allow algorithms to classify the multi-contrast image data. Scatter plots help evaluate the quality of image data [Gerig et al., 1992]. Ideally, these clusters would be well-separated, and classification would simply require the labeling of each cluster. In practice, however, they overlap and the challenge of classifiers is to form proper decision boundaries in feature space, by minimizing the number of overlapped voxels that are misclassified into the wrong boundary or cluster.

MR imaging artifacts affect the distribution of clusters in feature space. As images get noisier, clusters tend to spread out and overlap. RF inhomogeneity tends to elongate clusters in feature space. Figure 3.2 shows three scatter plots based on simulated T_1- and T_2-weighted MR images calculated over the entire volumes as histogram images, demonstrating the effects of noise and RF inhomogeneity on clusters in feature space.

In general, supervised classification methods can be categorized into statistical (parametric, non-parametric), decision-tree (rule-based), neural, and genetic. In this thesis, the performance of the following supervised methods has been compared:

- Minimum Distance (MD)
- Bayesian (BAYES)

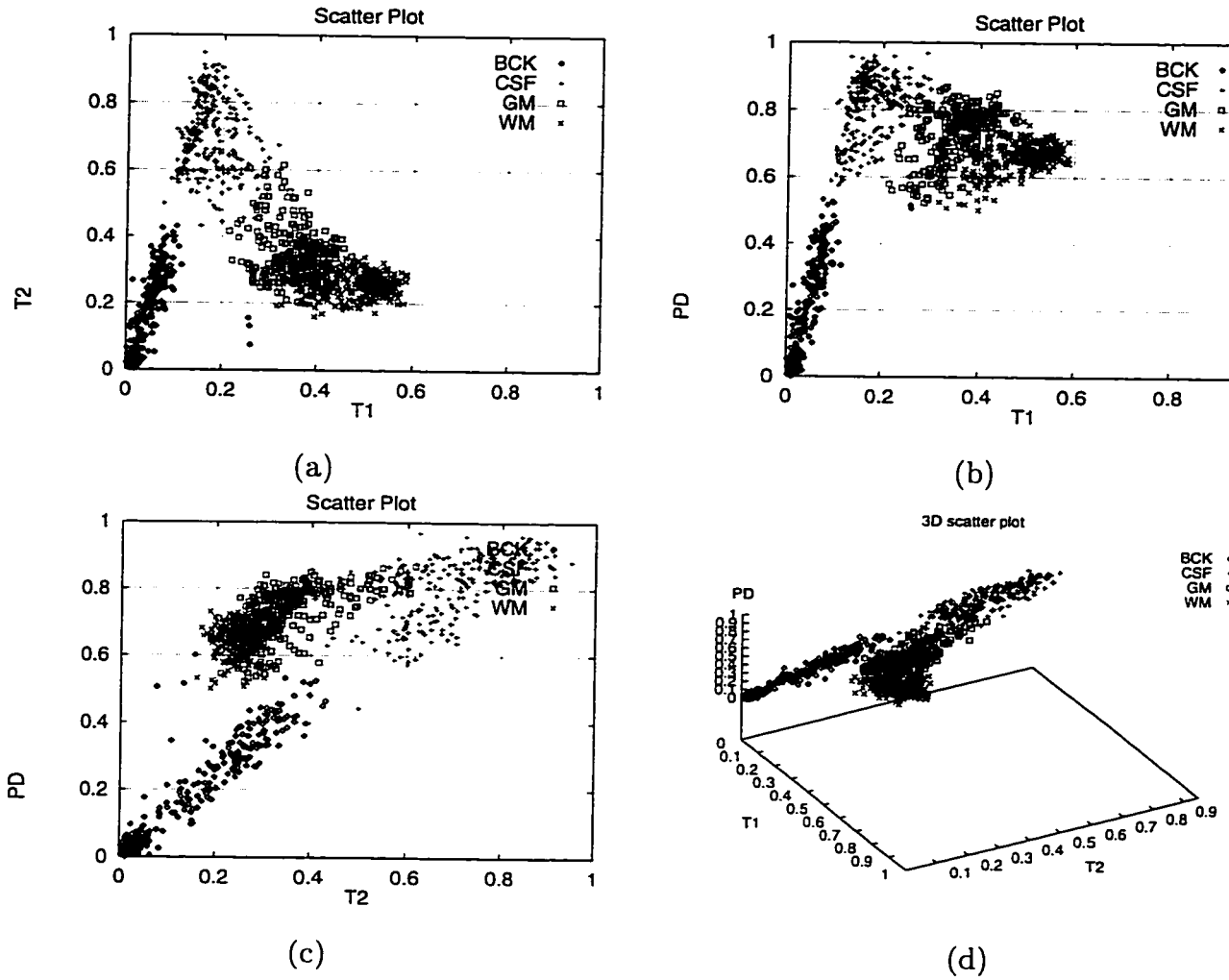
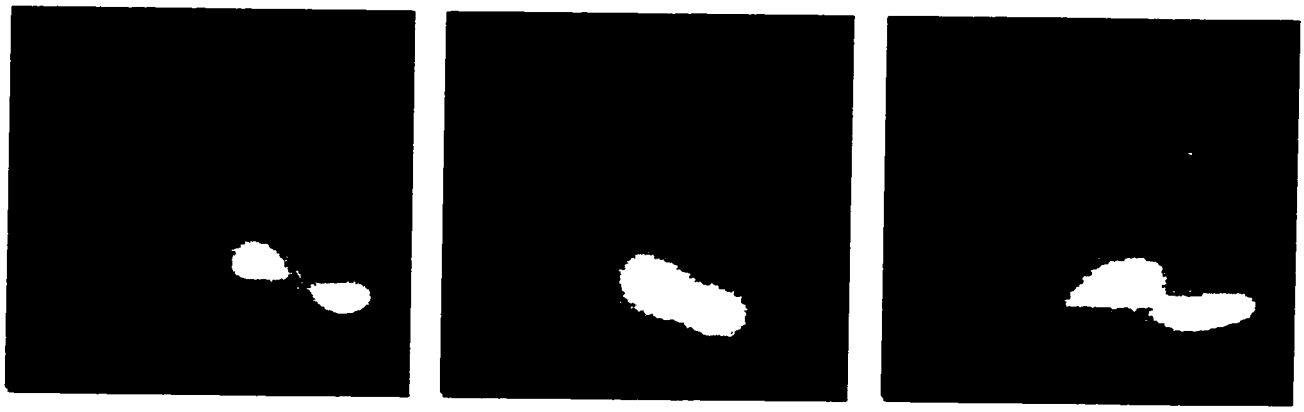


Figure 3.1: Scatter plots from (a) $T_1 - T_2$, (b) $T_1 - PD$ and (c) $T_2 - PD$ weighted images and (d) $T_1 - T_2 - PD$ weighted images. The dense clusters represent the different tissue types in the image. Note that ideally these clusters would not overlap.



(a) (b) (c)

Figure 3.2: Histogram images based on simulated T_1-weighted (abscissa) and T_2-weighted (ordinate) MR images containing (a) Normal levels of noise (3%) and RF inhomogeneity (20%), (b) increased level of noise (9%) (c) increased levels of RF inhomogeneity (50%). Note the position and size of the clusters in each histogram. (Sections 2.4.1 and 2.4.2 discuss the issues surrounding different levels of noise and RF inhomogeneity, respectively).

- k Nearest-Neighbor (kNN)
- C4.5 Decision Tree (C4.5)
- Artificial Neural Nets (ANN)

The following subsections explain the theory of each classifier used in this thesis. In order to help understand the mathematical formalism, it is necessary to establish a common set of symbols and definitions used in the description of the classifiers. The features of an image voxel is represented by a D-dimensional feature vector in $\Re^D$, denoted by $\mathbf{x}$:

$$\mathbf{x} = (x_1, x_2, \ldots, x_D) \tag{3.1}$$

where each of the x_k is the kth feature of the voxel sample ($k = 1, \cdots, D$) , and $\Re^D$ is the space of D-tuples of real numbers.

The training set comprises of N voxel samples, where each sample belongs to one of c classes, denoted by C_j ($j = 1, \cdots, c$). Furthermore, each class has N_j training samples such that:

$$N = \sum_{j=1}^{c} N_j \tag{3.2}$$

Let $\vec{\mu}_j$ denote the class prototype, also known as sample mean or cluster centroid for each class C_j, calculated as follows:

$$\vec{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{x}_i \quad \forall_j \tag{3.3}$$

where $\mathbf{x}_i$ the ith sample of class C_j.

Let d denote a Euclidean "distance" measure in $\Re^D$, between two vectors $\vec{v}_i$ and $\vec{v}_k$ defined as:

$$d(\vec{v}_i, \vec{v}_k) = \sqrt{(\vec{v}_i - \vec{v}_k)^t (\vec{v}_i - \vec{v}_k)} \tag{3.4}$$

t stands for transposed vector. The distance d satisfies the following three conditions:

$$\begin{cases} d(\vec{v}_i, \vec{v}_k) & \geq & 0 \\ d(\vec{v}_i, \vec{v}_k) & = & 0 \Leftrightarrow \vec{v}_i = \vec{v}_k \\ d(\vec{v}_i, \vec{v}_k) & = & d(\vec{v}_k, \vec{v}_i) \end{cases} \tag{3.5}$$

The Euclidean distance d, is often called a measure of *dissimilarity* between two samples: $d(\mathbf{x}_i, \mathbf{x}_k)$, or between a sample and a class centroid: $d(\mathbf{x}_i, \vec{\mu}_j)$.

3.3.1 Minimum Distance classifier

The Minimum Distance (MD) classifier, also called 'Nearest Prototype' [Duda and Hart, 1973; Bezdek, 1981] is based on distance measurements in feature space, from unknown samples to class prototypes. These prototypes are estimated from a given set of training samples according to Equation 3.3. The MD classifier decides on class membership of an unknown sample $\mathbf{x}$ based on the following relation:

$$\mathbf{x} \in C_j \Leftrightarrow d(\mathbf{x}, \vec{\mu}_j) = \min_{1 \leq k \leq c} \{d(\mathbf{x}, \vec{\mu}_k)\} \tag{3.6}$$

The Minimum Distance classifier is a simple and efficient classifier that has been used by numerous researchers [Vannier et al., 1985; Vannier et al., 1987; Vannier et al., 1988; Vannier et al., 1991; Kamber et al., 1992; Kamber et al., 1995] as baseline classifier against which others have been compared. The disadvantage of this classifier is the fact that it is not sensitive to inter-class variances of the training samples. Furthermore, the Euclidean distance tends to favor hyper-spherical shaped clusters in feature space, which may not be sufficient to discriminate properly certain non-linearities found in data being classified, especially if the data contains severe artifacts that tend to disrupt the distribution of clusters in feature space.

3.3.2 Bayesian classifier

The Bayesian classifier (BAYES), also called Maximum Likelihood classifier [Duda and Hart, 1973; Schowengerdt, 1983; James, 1985] is based on the supposition that the distribution of each tissue type in the brain is multivariate Gaussian, represented by $p(\mathbf{x})$, the probability density function defined as:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} \mid \Sigma \mid^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \vec{\mu})^t \Sigma^{-1} (\mathbf{x} - \vec{\mu})] \tag{3.7}$$

where $\vec{\mu}$ is the centroid, and $\sum$ is the D-by-D *covariance matrix* for each class.

In order to estimate the probability density function $p(\mathbf{x})$, the centroid and the covariance matrix need to be evaluated for each class C_j. Equation 3.3 is used to calculate the class centroid $\vec{\mu}_j$, and $\sum_j$ is calculated according to the following equation:

$$\sum_j = \frac{1}{(N_j - 1)} \sum_{i=1}^{N_j} (\mathbf{x}_i - \vec{\mu}_j)(\mathbf{x}_i - \vec{\mu}_j)^t \tag{3.8}$$

Thereafter, given an unknown sample $\mathbf{x}$, it is assigned to class C_j if the *a posteriori* probability of $P(C_j \mid \mathbf{x}) > P(C_i \mid \mathbf{x}), \quad \forall_{i \neq j}$, that is, we maximize $P(C_j \mid \mathbf{x})$ over $C_1, \ldots, C_c$. By Bayes' theorem:

$$P(C_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_j) P(C_j)}{p(\mathbf{x})} \tag{3.9}$$

since $p(\mathbf{x})$ is present as a fixed amount for all classes, maximizing $P(C_j \mid \mathbf{x})$ entails $p(\mathbf{x} \mid C_j)P(C_j)$, which is equivalent to maximizing:

$$\log p(\mathbf{x} \mid C_j) + \log P(C_j) \tag{3.10}$$

Since the probability densities are assumed to be multivariate Gaussian, $p(\mathbf{x} \mid C_j) \sim \mathcal{N}(\vec{\mu}_j, \Sigma_j)$, then from Equation 3.7 the following discriminant function can be evaluated for each class [Duda and Hart, 1973]:

$$g_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \vec{\mu}_j)^t \Sigma_j^{-1} (\mathbf{x} - \vec{\mu}_j) - \frac{D}{2} \log 2\pi - \frac{1}{2} \log \mid \Sigma_j \mid + \log P(C_j) \tag{3.11}$$

The term $\frac{D}{2}\log 2\pi$ can be dropped from the above equation for the purpose of evaluating Equation 3.10. Usually, the probabilities $P(C_j)$ are considered to be equal for all classes, and hence eliminated from above equation. However, if *a priori* information about class distribution is available (Chapter 4), this information can be used in Equation 3.11.

The Bayesian classifier has been used extensively by a very large number of researchers [Vannier et al., 1985; Vannier et al., 1987; Vannier et al., 1988; Cline et al., 1990; Gutfinger et al., 1991; Vannier et al., 1991; Clarke et al., 1992; Gerig et al., 1992; Kamber et al., 1992; Clarke et al., 1993; Kamber et al., 1995]. It makes use of inter-class variances in the training samples and has the added advantage of making use of *a priori* probabilities, if available. Furthermore, it is based on the Mahalanobis distance, $d_M(\mathbf{x}, \vec{\mu}_j) = \sqrt{(\mathbf{x} - \vec{\mu}_j)^t \sum_j^{-1} (\mathbf{x} - \vec{\mu}_j)}$, (see Equation 3.11). Unlike Euclidean distance, it tends to favor hyper-ellipsoid shaped clusters in feature space. Compared to the Minimum Distance classifier, this tends to make the Bayesian classifier well suited to handle certain artifacts like RF inhomogeneity, which elongates cluster in feature space. However, this in itself also may not be sufficient to discriminate properly certain non-linearities found in data being classified. Moreover, the Bayesian classifier assumes that tissue distribution is multivariate Gaussian, which some researchers [Clarke et al., 1993] state that is not the case for MR data.

3.3.3 k Nearest-Neighbor classifier

Parametric classification methods such as MD and BAYES, function under the assumption that tissue class distributions follow some statistical model, characterized by a number of parameters (for instance, a multivariate Gaussian distribution is characterized by calculating the mean and variance of tissue intensities). This assumption is not necessarily true for tissue classes in MRI, and as a result this type of classifier can perform suboptimally [Bezdek et al., 1993]. Studies have shown that parametric classifiers are the least stable among several different classification methods [Clarke et al., 1993]. The k-Nearest Neighbor (kNN) classifier does not assume any underlying parametric distribution

of data [Kulkarni, 1994]. The kNN algorithm is the following:

- Step 1: Choose a set X_L containing m training samples, $X_L = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$.
- Step 2: Choose $1 \leq k \leq m$, the number of nearest neighbors, for which the algorithm will be executed.
- Step 3: Choose a distance measure, usually Euclidean (Equation 3.4), between pairs of D-dimensional feature vectors.
- Step 4: For each unknown sample $\mathbf{x}$, calculate and rank order of distances $d(\mathbf{x}, \mathbf{x}_i)$ where d is the distance measure chosen in step 3:

$$d_1 \leq d_2 \leq \cdots \leq d_k \leq d_{k+1} \leq \cdots \leq d_m \quad \forall_i \tag{3.12}$$

- Step 5: If the majority of $d_1, d_2, \ldots, d_k$ belong to class C_j, then assign $\mathbf{x}$ to class C_j.
- Step 6: In case of a tie, calculate the sum of distances in each tied class, and assign $\mathbf{x}$ to the class with the minimum sum.

Researchers [Cline et al., 1991; Gutfinger et al., 1991; Clarke et al., 1992; Clarke et al., 1993] have reported favorable results with the kNN classifier. The major disadvantage of kNN is the fact that, in order to properly estimate the tissue density functionals, a large number of training samples have to be chosen which dramatically increases the time it takes to classify the data.

3.3.4 C4.5 Decision tree classifier

The C4.5 classifier generates a decision tree from a set of training samples based on the principle of *divide and conquer*. If ... then ..., rules that can be generated from decision trees provide insight concerning the nature of data being classified.

Tests at each node of the tree divide the training set into at least two subsets in a non-trivial fashion. Thereafter, continuing this principle recursively at each node, until all the training set samples end up in leaf nodes which represent the different classes.

Consider a training set G with samples belonging to c classes. The construction of the decision tree proceeds as follows [Quinlan, 1993; Quinlan, 1996]:

- If G contains samples all belonging to class C_j, then the tree is a leaf node with class C_j.
- If G contains no classes, then the tree is also a leaf node, but to associate this leaf node to a class, the information has to come from a source other than G.
- If G contains two or more classes. a test has to be chosen on a single attribute that will yield one or more mutually exclusive outcomes, thereby partitioning G into several subsets $(G_1, G_2, \cdots, G_q)$, based on the outcome of the test. At this point, the decision tree will be a node from which a number of branches will emerge corresponding to the outcome of the test.

The information available to construct the test at a node is the distribution of classes in G at that particular node. Therefore, one of the most crucial aspects of any decision tree classifier is the proper selection of the attribute to be tested. Since there may be several attributes available for testing at each test node, a measure of evaluating *test appropriateness* is necessary to determine the most discriminating aspect of the test function, that will ultimately subdivide the training samples at the given test node.

ID3, which is a decision tree classifier developed by Quinlan [Quinlan, 1986], has made use of an entropy function defined as follows [Quinlan, 1993; Quinlan, 1996]:

$$\text{info}(G) = -\sum_{j=1}^{c} \frac{\text{freq}(C_j, G)}{|G|} \times \log_2 \left(\frac{\text{freq}(C_j, G)}{|G|} \right) \text{ bits} \tag{3.13}$$

where $\text{freq}(C_j, G)$ is the number of samples in G that belong to class C_j, and $|G|$ is the number of samples in G.

Let us assume that a test T on a given attribute results in splitting G into q subsets. The information provided by this test is the weighted sum of the subsets:

$$\text{info}_T(G) = \sum_{i=1}^{q} \frac{|G_i|}{|G|} \times \text{info}(G_i) \tag{3.14}$$

$$\text{gain}(G, T) = \text{info}(G) - \text{info}_T(G) \tag{3.15}$$

In these equations, *gain* indicates the information gained by dividing G according to test T, and $\text{info}_T(G)$ is the average information required to identify the class of a sample in G. We select that attribute for testing which maximizes $\text{gain}(G, T)$

The gain criterion has been criticized to favor tests that results in large number of subsets q [Quinlan, 1993]. The C4.5 classifier addresses this problem by normalizing this apparent gain due to large number of subsets q. This modification is accomplished by defining two terms *split* and *gain ratio*:

$$\text{split}(G, T) = -\sum_{i=1}^{q} \frac{|G_i|}{|G|} \times \log_2 \left(\frac{|G_i|}{|G|} \right) \tag{3.16}$$

$$\text{gain ratio}(T) = \text{gain}(G, T)/\text{split}(G, T) \tag{3.17}$$

where *split* describes the potential information gained by splitting G into q subsets, and *gain ratio* tells how useful that split is. However, when it comes to constructing tests involving continuous attributes, several researchers have reported superior results to that of C4.5 [Auer et al., 1995; Dougherty et al., 1995]. This has led to two recent modifications by Quinlan to rectify a weakness in C4.5, which prompted a newer version (Release 8)

that has demonstrated superior results [Quinlan, 1996]. The first modification involves accounting for increased cost associated with the test on continuous attributes, and is carried out as follows: Assume a particular continuous attribute $\mathcal{A}$ has the sorted values $(o_1, o_2, \cdots, o_r)$. There are only $(r-1)$ possible splits on $\mathcal{A}$, thus $(r-1)$ thresholds of $\mathcal{A}$. This requires $\log_2(r-1)$ additional bits of information, the value of the gain(G,T) has to be adjusted accordingly :

$$\text{gain}(G,T) = \text{info}(G) - \text{info}_T(G) - \log_2(r-1)/|G| \tag{3.18}$$

So a test on continuous attributes with numerous outcomes (large values of q) will be less likely to have a maximum value of the splitting criterion, among possible tests, thereby less likely of being selected. The second modification involves Equation 3.17. Usually the information obtained from the gain ratio is used to select an appropriate threshold value in order to maximize discrimination among continuous attributes. Given the modification above, the value of the split in Equation 3.17 will vary as a function of the threshold. Quinlan thinks that this is an unnecessary complication, and hence chooses to use the value of the threshold itself, rather than the split ratio, to maximize the gain.

The advantage a decision tree classifier is that insight can be gained in analyzing the way the tree classifies the data. Hence a set of *if* ... *then* ... rules can be constructed that can help understand the nature of data being classified. The disadvantage of this method, on the other hand, becomes evident when the number of training samples increases. It usually results in huge decision trees that are difficult to interpret and decipher, and thus have to be pruned in order to reduce them to manageable sizes. Depending on the efficiency of the pruning algorithms [Quinlan, 1993], these costly operations may take a long time with no apparent gain in discriminating power.

3.3.5 Artificial Neural Network classifier

Artificial Neural Networks (ANNs) are parallel information processing systems that are derived from mathematical abstractions of the human nervous system. ANN's comprise interconnected processors called *neurodes* [Caudill and Butler, 1992]. There are numerous kinds of neural networks (Hopfield, Kohonen, Hamming, Carpenter, Layered) [Lippmann, 1987] of which the most widely used architecture is the layered neural network. Figure 3.3 shows a diagram depicting an ANN with an input layer, two hidden layers (this number can vary depending on the architecture designed) and out output layer.

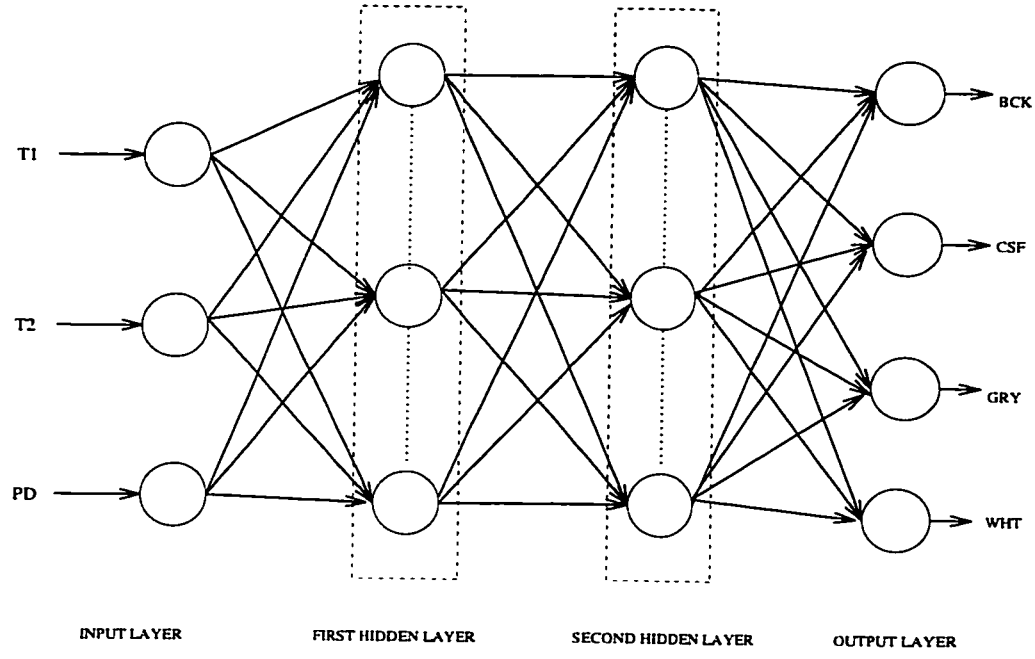


Figure 3.3: Diagram showing the architecture of a layered Artificial Neural Network.

To train a layered network, an error back-propagation (BP) algorithm is used, which minimizes a gradient descent function while training samples are presented to the network, samples for which the desired output is known. Thereafter a set of weights on the neurode connections are adjusted to produce the proper output. The BP-ANN algorithm is as follows:

- Step 1: Choose a learning rate $0 \leq \beta \leq 1$, momentum rate $0 \leq \alpha \leq 1$ and stopping

criterion ϵ.

- Step 2: Initialize the set of weights w_{ij} on the neurode connection to small random values.

- Step 3: Present a training sample's D-dimensional input feature vector $\mathbf{x}_{in}$, where the desired target output vector is $\mathbf{xc}_{tar}$.

- Step 4: Calculate the actual output vector $\mathbf{xc}_{obs}$ of the network for input vector $\mathbf{x}_{in}$, using the following sigmoid logistic function:

$$\mathbf{xc}_{obs} = \frac{1}{1 + e^{-(\mathbf{x}_{in})}} \tag{3.19}$$

- Step 5: Starting at the output nodes, recursively adjust the weights by:

$$\Delta w_{ij} = \beta E_j \mathbf{xc}_i + \alpha \Delta w_{ij}^{previous} \tag{3.20}$$

where Δw_{ij} is change in the weight from hidden node i, or from the input node j, $\mathbf{xc}_i$ is the input to this neurode in this connection, and E_j is an error term for node j such that:

$$\begin{cases} E_j^{output} = \mathbf{xc}_{tar} - \mathbf{xc}_{obs} & \text{for an output layer} \\ E_j^{hidden} = \mathbf{xc}_{hidden}(1 - \mathbf{xc}_{hidden}) \sum_{j=1}^{s} w_{ij} E_j^{output} & \text{for hidden layer} \end{cases} \tag{3.21}$$

where s is the number of neurodes in the hidden layer, and the superscripts "output" and "hidden" indicate the layer to which the neurode belongs.

- Step 6: If $E_j \leq \epsilon$ for the output node, stop; otherwise goto to step 3.

Artificial neural nets have been extensively used in brain image classification by several researchers [Hall et al., 1992; Clarke et al., 1992; Clarke et al., 1993; Özkan et al., 1993; Zijdenbos et al., 1993]. Their main advantage is their tremendous adaptability to non-linearities that may be present in image data. On the other hand, it is difficult to gain insight in the "reasoning" encoded in a trained ANN, because the discriminating power of the algorithm is represented as a set of weights associated with the connections between neurodes. Also, depending the size of the training samples and the complexity of the network, training time can be considerable.

3.4 Unsupervised classification methods

In unsupervised classification methods, the entire image data is represented as a set X, comprising of n D-dimensional feature vectors denoted by $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$.

Unsupervised learning, also called clustering, involves the assignment of hard $\{0,1\}$ or fuzzy $[0,1]$ cluster labels u to data elements in X. The partitioning of X into c-partitions indicating class membership, is represented by a $(c \times n)$ matrix $U = [u_{ik}]$ satisfying the following conditions [Bezdek et al., 1993]:

$$\begin{cases} 0 \leq u_{ik} \leq 1 & \forall_{i,k,} \\ 0 < \sum_{k=1}^{n} u_{ik} < n & \forall_{i,} \\ \sum_{i=1}^{c} u_{ik} = 1 & \forall_{k} \end{cases} \tag{3.22}$$

The following are examples of different U matrices; a $(c \times n)$ generic matrix U_0, a 2-by-3 hard-partitioned matrix U_1, and a 3-by-3 fuzzy-partitioned matrix U_2, where each column represents hard or fuzzy membership of each feature vector to each cluster.

$$U_0 = \begin{bmatrix} u_{11} & \cdots & u_{1k} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ u_{i1} & \cdots & u_{ik} & \cdots & u_{in} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ u_{c1} & \cdots & u_{ck} & \cdots & u_{cn} \end{bmatrix} \tag{3.23}$$

$$U_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{3.24}$$

$$U_2 = \begin{bmatrix} 1.00 & 0.25 & 0.15 \\ 0.00 & 0.15 & 0.10 \\ 0.00 & 0.65 & 0.75 \end{bmatrix} \tag{3.25}$$

The entire image volume is classified using clustering algorithms, since unsupervised methods do not rely on training sets. It is however necessary to specify c, the number of clusters (tissue classes) to search for. Once determined, they require human intervention in order to label clusters with the appropriate tissue classes, since the clustering mechanisms have no way to tell what cluster should belong to what specific tissue class.

Two well known clustering algorithms are described in this thesis : Hard C-Means (HCM), also known as k-Means [Duda and Hart, 1973; Hartigan, 1975; Bezdek, 1981] and Fuzzy C-Means (FCM) [Bezdek, 1981]. These methods, as well as most unsupervised methods, make use of a *criterion* or *objective function* that tells the quality of clusters of image data in feature space, and which is typically minimized in an iterative fashion. The HCM and FCM criterion functions are described next, followed by a common algorithm used in their respective minimizations.

3.4.1 Hard C-Means classifier

A commonly used objective function is the *sum-of-squared-errors* function [Duda and Hart, 1973] defined as:

$$J_1(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} d(\mathbf{x}_k - \mathbf{v}_i)^2 \tag{3.26}$$

where

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} u_{ik}\mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}}, \ \forall i \tag{3.27}$$

here d is the Euclidean distance as defined in Equation 3.4, $V = (\mathbf{v}_1, \cdots, \mathbf{v}_c)$ is the vector representing cluster centroids, and U is the hard partition of X into c-partitions. (U, V) may minimize J_1 if:

$$u_{ik} = \begin{cases} 1, & d(\mathbf{x}_k - \mathbf{v}_i) < d(\mathbf{x}_k - \mathbf{v}_j), \quad j = 1, \cdots, c, \quad j \neq i \\ 0, & otherwise \end{cases} \quad \forall i, k. \tag{3.28}$$

HCM will produce a matrix U of hard clusters where each feature vector receives a unique cluster membership. It is considered as a multi-pass Minimum Distance classifier with the advantage of requiring no user input to train the algorithm; however, it does require user input to decide on the number of clusters, and assign different clusters to various tissue classes. A few researchers [Gutfinger et al., 1991] have used HCM in classifying MR brain images. One of the main disadvantages of this, and other unsupervised methods is the very long processing time, and when classifying 3D image volumes, the large computer memory requirements.

3.4.2 Fuzzy C-Means classifier

For the FCM clustering algorithm, the objective function defined in Equation 3.26 can be generalized as follows [Bezdek et al., 1993]:

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d(\mathbf{x}_k - \mathbf{v}_i)^2 \tag{3.29}$$

where

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n}(u_{ik})^m}, \ \forall i \tag{3.30}$$

and $m \in [1,\infty]$, V and $\mathbf{v}_i$ are the same as in HCM, defined earlier. Similarly (U,V) is the fuzzy partition of X that may minimize J_m for $m > 1$ if:

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d_A(\mathbf{x}_k - \mathbf{v}_i)}{d_A(\mathbf{x}_k - \mathbf{v}_j)}\right)^{\frac{2}{(m-1)}}\right]^{-1}, \quad \forall i,k \tag{3.31}$$

where

$$d_A(\mathbf{x}_k - \mathbf{v}_i) = \sqrt{(\mathbf{x}_k - \mathbf{v}_i)^t A(\mathbf{x}_k - \mathbf{v}_i)} \tag{3.32}$$

A is any positive definite matrix. FCM produces a matrix U of fuzzy clusters where each data element will receive a fuzzy class membership over the c-partitions. Note that the cluster centroids $\mathbf{v}_i$ in case of FCM should not be confused with μ_i (Equation 3.3) as they are evaluated differently.

The following algorithm describes both FCM and HCM classifiers, appropriate differences are indicated in the processing steps:

- Step 1: Let the set X represent the entire image data set containing n samples, $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$.

- Step 2: Choose the number of classes c, the error term $\epsilon > 0$, the number of iterations W, the distance measure d. For FCM, additionally choose weighting exponent m and positive definite D-by-D matrix A.

- Step 3: Initialize the first matrix U_0 randomly to satisfy the conditions of Equation 3.4, such that $u_{ik} \in \{0,1\}$ (HCM), or $u_{ik} \in [0,1]$ (FCM).

- Step 4: Calculate initial cluster centroids $\{\mathbf{v}_i, 0\}$, using Equation 3.27 (HCM) or Equation 3.30 (FCM), for $i = 1, \ldots, c$.

- Step 5: For $l = 1, \ldots, W$ do iteratively.

  a. Compute $\{u_{ik,l}\}$ with Equation 3.28 (HCM) or Equation 3.31 (FCM), for $k = 1, \ldots, n$.

  b. Compute $E_l = \| U_l - U_{l-1} \| = \sqrt{\sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik,l} - u_{ik,l-1})^2}$;

  c. If $E_l \leq \epsilon$ stop; Otherwise calculate $\{\mathbf{v}_i, l\}$ using Equation 3.27 (HCM) or Equation 3.30 (FCM);

Researchers [Hall et al., 1992; Clarke et al., 1992; Clarke et al., 1993] have used FCM in MR image classification. Like HCM, the advantage of FCM is that it is an automated classification method, and it has got the added advantage of assigning fuzzy class membership. However, besides suffering from the same disadvantages of HCM clustering algorithm, because of its fuzzy nature, it has larger time and memory requirements. Another disadvantage of FCM is the fact that it has a tendency of favoring solutions of equal cluster populations. In order to compensate for this shortcoming, Bensaid [Bensaid et al., 1996] proposed a semi-supervised Fuzzy C-Means algorithm (SFCM), that relies on operator supplied sample training points to guide the initial cluster centroids in the right direction. This method has demonstrated promising result, and is tested in this thesis.

3.5 Concluding remarks

This chapter has reviewed the MR image classification literature. It has introduced the theoretical aspect of the classification algorithms tested in this thesis, while discussing the advantages and disadvantages of each one. The next chapter will introduce the concept of *tissue probability maps*, as a mechanism of automatic supervised classification algorithms.

Chapter 4

Brain Tissue Probability Maps

4.1 Introduction

In order to classify a brain volume using supervised methods, a particular classification algorithm needs to be trained on a set of voxel (or regional) samples from that volume (Chapter 3). The training set is created manually by experts who pick voxels or draw regions in a brain image volume from areas representative of the different tissue types to be classified by the algorithm.

However, it has been observed that experts are likely to choose samples that are sure to belong to a particular class, ignoring parts of the image volume where image quality degradation, such as partial volume effects (Section 2.4.3), is intense. This leads the algorithm to train only on the most certain voxel intensities, resulting in inconsistencies and poor classification. Furthermore, depending on the number of brain volumes to be classified, manual selection of training samples becomes a very tedious task with poor reproducibility. A mechanism that automates the training sample selection process in a robust and reproducible fashion is therefore highly desirable.

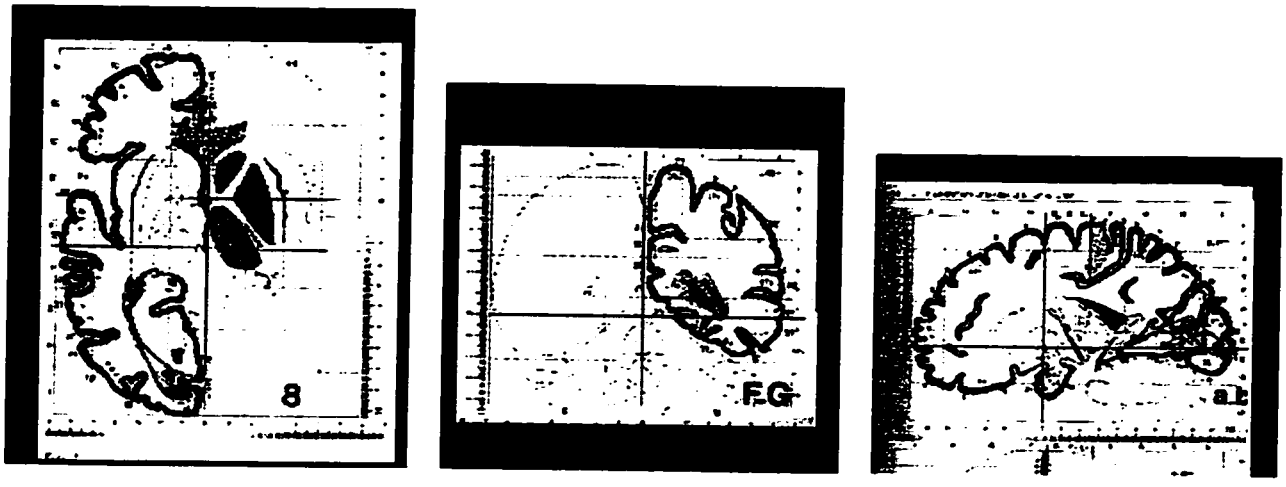
4.2 Stereotaxic space

The concept of **stereotaxic space**, developed by Talairach and Tournoux [Talairach et al., 1967; Talairach and Tournoux, 1988], defines a standard anatomically-based frame of reference, where brains of different sizes and shapes can be directly compared after removal of size and orientation differences. This external anatomical reference stems from the need to predict the positions of anatomical structures within the human brain while performing neuro-surgical procedures like inserting electrodes deep in the brain. By using a standardized proportional coordinate system, brain volumes transformed into this coordinate space, can be compared on a voxel-by-voxel basis. Each voxel has an associated "world coordinate," expressed as its Cartesian distance in mm from the stereotaxic origin. Figure 4.1 shows digitized sample slices of Talairach's stereotaxic brain atlas in three orthogonal views. The brain mapping research community has adopted this stereotaxic space as a common 3D coordinate system for mapping physiological, anatomical and functional information across many subjects [Fox et al., 1985; Evans et al., 1992b; Fox et al., 1994].

Stereotaxic brain space was considered by Kamber *et al.* [Kamber et al., 1992; Kamber et al., 1995] as a basis for creating 3D stereotaxic tissue probability maps used for providing *a posteriori* probabilities for white matter lesion classifications. Since white matter lesions predominantly occur in white matter, the white matter tissue probability map was used to disallow the detection of lesions at unlikely locations. The use of stereotaxic space is adopted in this thesis to serve as the basis for building new, higher resolution tissue probability maps, where each stereotaxic voxel provides geometric information that can be utilized as follows:

- Automatically select training samples representing various tissue types in a particular brain volume that has been transformed into stereotaxic space, therefore making it robust, reliable, and more importantly, reproducible.
- Provide *a priori* and *a posteriori* probability information to classifiers (Section 3.3.2).

- Impose a 3D spatial constraint on brain classification, in addition to intensity constraints that are provided by the gray-scale levels of different MR image sequences.



(a) (b) (c)

Figure 4.1: Sample images from (a) transverse, (b) coronal and (c) sagittal slice views of the Talairach atlas demonstrating stereotaxic space.

4.3 Tissue probability maps

All supervised classification methods found in the literature rely on operator intervention in selecting training samples. Because of this operator dependency, supervised methods are not considered to be very reproducible. Therefore, several researchers [Gutfinger et al., 1991; Gerig et al., 1992; Clarke et al., 1993; Bensaid et al., 1996] have adopted unsupervised methods.

Stereotaxic space makes it possible to consider the concept of Tissue Probability Maps (TPMs), which are tissue specific probability values denoting the level of certainty with which a particular voxel in 3D stereotaxic space belongs to one of the tissue classes (CSF, GM and WM). Such maps can be used in the selection of training samples for

classifying new MR image data sets, thus creating a mechanism for automating supervised classification algorithms.

4.3.1 The use of tissue probability maps in automating training sample selection

The original tissue probability maps created by Kamber *et al.* [Kamber et al., 1992; Kamber et al., 1995] (Figure 4.2) denote the probability (0% - 100%) with which each voxel in stereotaxic space belongs to any of the CSF, GM and WM tissue classes. Looking at individual probability maps in Figure 4.2, which are themselves image volumes, one sees images of varying intensity. The brighter a particular location, the more probable it is to belong to that specific tissue class.

The automatic selection of training samples proceeds by choosing a particular probability threshold (say 90%). Thereafter, the tissue probability maps are used to generate a list of all the stereotaxic voxel coordinates in a brain volume meeting the above mentioned probability threshold. From this list (which could exceed several thousand voxels - depending on the probability threshold and the resolution of the maps), a predetermined number of coordinates are chosen at fixed or random intervals as representative populations for each tissue type. For example, suppose a particular probability threshold resulted in generating 100,000 white matter voxels. A subset of 25 training voxels can be obtained by sub-sampling the large training set at regular intervals (step size is fixed at 4000) or at random intervals (step size is a random number between 1 and 4000). This sub-sampling method has the following advantages:

- The entire brain from the base to the top is traversed, ensuring proper sub-sampling of training voxels across the whole brain.
- Fixed intervals generate unique training sets, while random intervals generate varied training sets. This can be useful if one is interested in establishing statistical significance through the use of numerous similar training sets, generated by the same set

of parameters (see Section 8.2).

Two parameters control the automatic generation of training samples from a given set of tissue probability maps. **(a)** the threshold at which probabilistic voxel positions are generated, and **(b)** the number of voxels chosen from the total complement in **(a)**. By varying these two parameters, many training sample sets can be obtained. Table 4.1 shows a sample list of six such coordinates (two per class), and the tissue classes to which they belong. These coordinates are 3D voxel positions in millimeters from the *Anterior Commissure*, defined as the origin in the stereotaxic atlas.

Once intensity values for each of the 3D voxel samples are obtained, intensity ranges for each tissue type can be determined. This list can be used as a training sample set, indicating the most probable locations in stereotaxic space of specific tissue types. The coordinate list for these voxels can be used on any brain volume that has been transformed into stereotaxic space, thus eliminating the need to choose training samples for each brain volume individually.

In this thesis, the term **training set** refers to a list of labeled 3D probabilistic coordinates (Table 4.1) for each tissue type in stereotaxic space, as opposed to the MR image intensities at specific voxel locations which depend on the subject and the acquisition parameters. This method has the advantage of training the classifier on intensities extracted from probabilistic locations specific to individual brain volumes, independent of the resolution, intensity and neuro-anatomical variations across many volumes (Section 1.3).

4.3.2 Rationale for creating new TPMs

The tissue probability maps created by Kamber *et al.* [Kamber et al., 1992; Kamber et al., 1995] have certain shortcomings, namely:

- They were based on only 12 healthy volunteers.

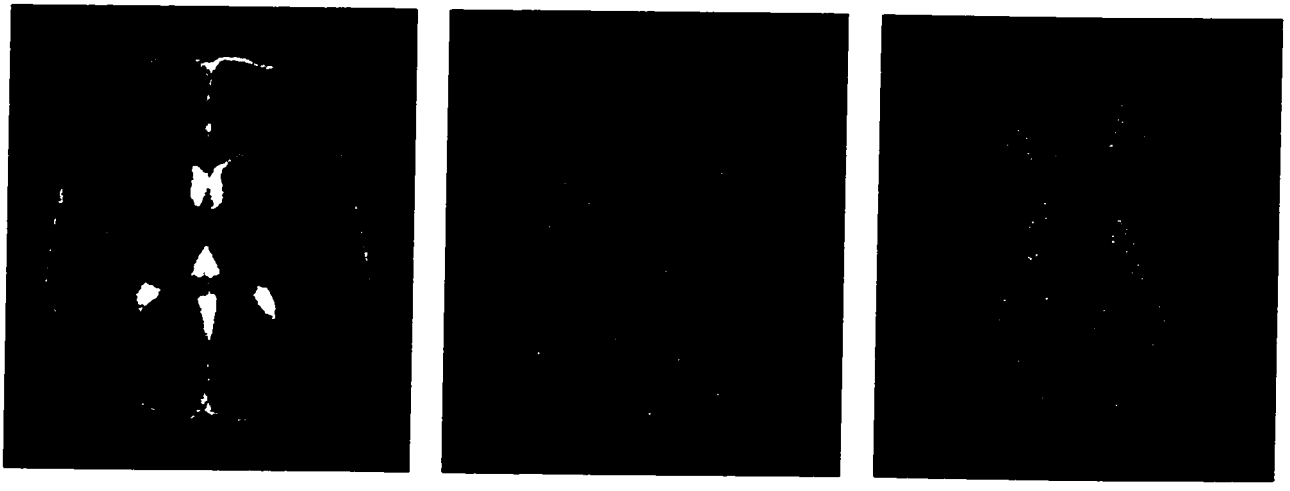
Stereotaxic coordinates			Class
X	Y	Z	*label*
−9.38	−40.08	−4.5	CSF
2.68	−45.24	−1.5	CSF
20.10	−32.34	−22.5	GM
−0.67	−69.32	−21	GM
21.44	−32.34	40.5	WM
−15.41	−2.23	39	WM

Table 4.1: An example of a training set; a list of stereotaxic coordinates and their respective tissue classes. Stereotaxic coordinates refer to millimetric coordinates relative to the *Anterior Commissure* which is the origin of stereotaxic space. Voxel coordinates on the other hand, are indices into the image volume of a given voxel.

- The brain volume of each volunteer was acquired at 2mm slice thickness, using an older Philips Gyroscan MR scanner at the Montréal Neurological Institute (MNI).
- The classification from which the TPMs were derived are based only on T_1-weighted and T_2-weighted imaging sequences.
- The transformation of individual brain volumes into stereotaxic space was done manually [Evans et al., 1992a].
- Only one classification algorithm (Bayesian) was used to construct the TPMs.

Figure 4.2 shows sample slices from three tissue probability maps created by Kamber *et al.* [Kamber et al., 1992; Kamber et al., 1995]. Although these maps have certain deficiencies, they provide the basis for creating new tissue probability maps with the following advantages:

- The new TPMs are based on 53 healthy volunteers, reflecting greater statistical power.



(a) (b) (c)

Figure 4.2: a) Cerebro-spinal fluid, b) gray matter and c) white tissue probability maps created by Kamber *et al.*

- The brain volume of each volunteer was acquired at 1mm slice thickness, using a recently installed Philips Gyroscan ACS II 1.5 T MR scanner (also at the MNI), which gives higher resolution and better quality images.
- The classification of the individual volumes for the new TPMs are based on *PD*-weighted as well as on T_1-weighted and T_2-weighted imaging sequences, to produce better classification.
- Transformation into stereotaxic space is done using an automated method, thus assuring consistency [Collins et al., 1994].
- Three separate algorithms, MD, ANN, and C4.5 (Chapter 3) were used to examine similarities and differences in the resulting tissue probability maps among the different algorithms.

The following subsections describe the steps taken to create the new tissue probability maps that resulted in superior spatial resolution and neuro-anatomical accuracy.

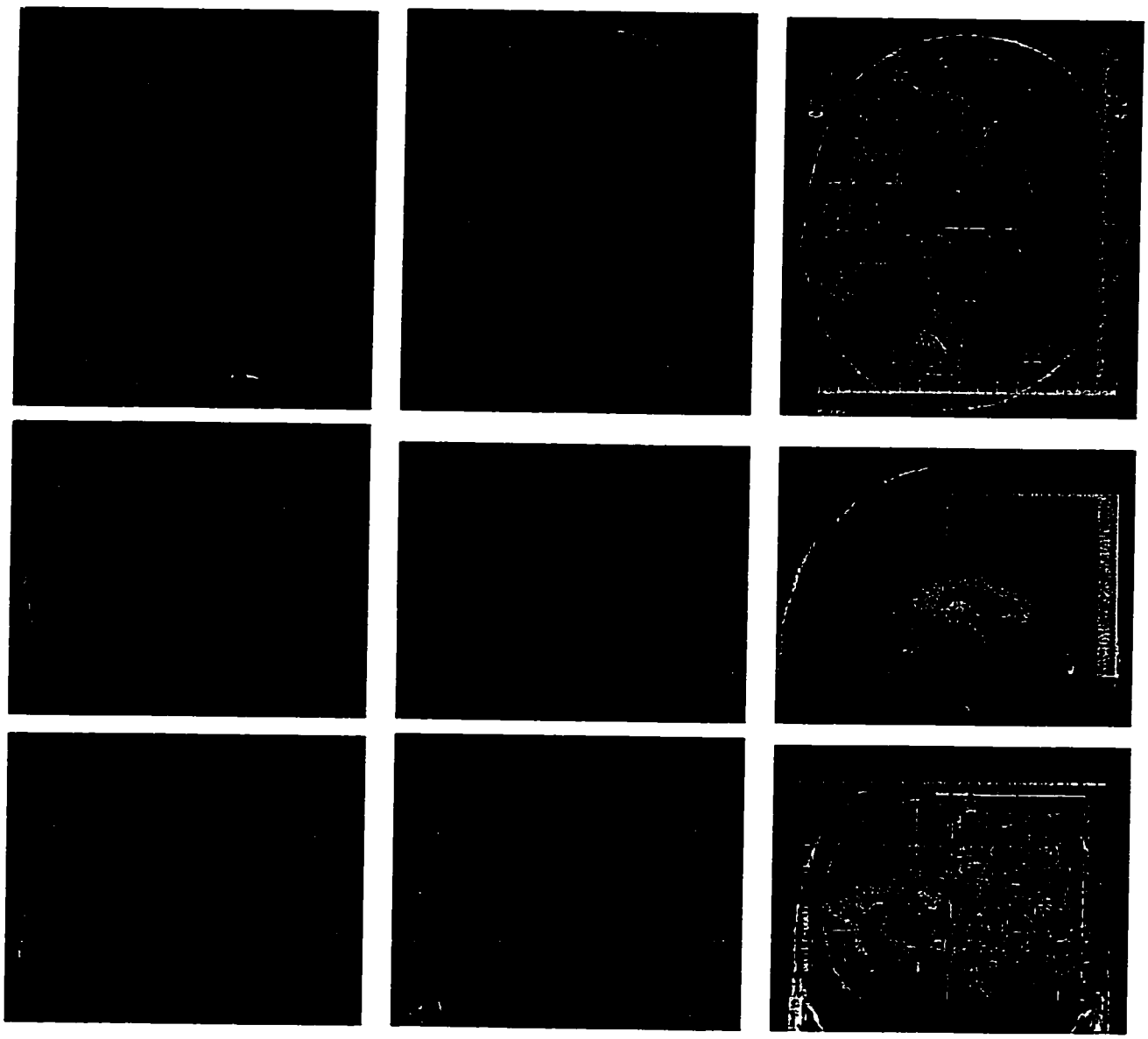
4.3.3 Pre-processing of data

Data Acquisition

The International Consortium for Brain Mapping (ICBM) is a collaboration between McGill University in Montréal, The University of California, Los Angeles, and The University of Texas Health Science Center, San Antonio, to produce a probabilistic map of human neuro-anatomy using MRI [Mazziotta et al., 1995]. The ICBM provided the MRI brain scans from 53 normal volunteers (31 males and 22 females with mean age of 23.6 years) to construct the new TPM. The MR scans were carried out on a Philips Gyroscan ACS II scanner with a 1.5 Tesla super-conducting magnet using gradient echo T_1-weighted, (TE=10ms; TR=18ms; flip angle=30°), and a double spin echo (TE=35,120ms; TR=3300ms; flip angle=90°) for the PD- and T_2-weighted images acquired at 1mm isotropic resolution.

Intra-subject registration and stereotaxic transformation

Since two separate sequences were used to obtain the above scans (T_1, T_2/PD), subject movement between scans can cause considerable mis-alignment of images. Therefore, the images obtained from the two sequences were first registered to each other to eliminate any such mis-alignment and then transformed into stereotaxic space, both procedures using a method developed at the MNI by Collins [Collins et al., 1994]. This automated method uses a nine parameter (3 translations, 3 rotations and 3 scaling) transformation process to convert brain volumes from their original space into stereotaxic space. Figure 4.3 demonstrates the effects of transformation of brain images from their original (native or scanner) space into stereotaxic space. Figure 4.3a shows an MR image volume in 3 orthogonal views in scanner space, and Figure 4.3b shows the same image volume after being transformed into stereotaxic space (cross-hairs in all images are situated at the *Anterior Commissure* - the origin of stereotaxic space).



(a) (b) (c)

Figure 4.3: Three orthogonal views demonstrating the effects of stereotaxic transformation. (a) Scanner (native) space (b) Stereotaxic space (c) Digitized stereotaxic atlas super-imposed. Cross-hairs in each image denote the position of the *Anterior Commissure*, which is the origin of stereotaxic space.

RF correction

The volumes were corrected for RF inhomogeneity by the method described in Section 6.3.4.

4.3.4 Post-processing of data

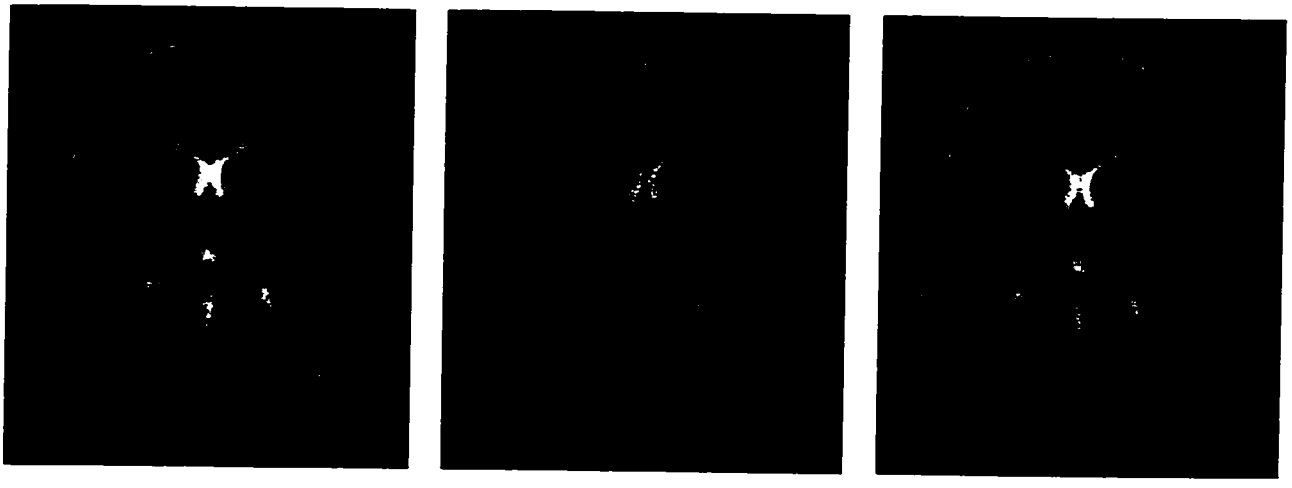
Training, classification and averaging

The original TPMs were used to automatically generate a training set that contained 50 sample voxels extracted at 100% probability threshold for each tissue class. The training set was then used to train and classify each of the 53 ICBM brain volumes by 3 different methods of classification: ANN, MD and C4.5. Fifty three individual tissue maps from the resulting classifications were averaged to form the new tissue probability maps. Figures 4.4, 4.5 and 4.6 show the tissue probability maps for CSF, GM and WM classes respectively, created by the three separate methods. The rationale for using three different methods was to determine whether each of the three methods produced similar or different results, and at the same to time find out which of the three method conveyed the most accurate neuro-anatomical information. The choice of algorithms was simply dictated by their availability (implementation) at the time this experiment was carried out.

Expert neuro-anatomical examination

A neuro-anatomist (NK) at the McConnell Brain Imaging Center carefully examined each of the tissue probability maps produced by the three different methods. Overall, she observed that the maps were very similar except for very few differences in the basal ganglia, gray matter structures deep inside the brain, which the ANN algorithm seemed to have portrayed best. Also, there were some differences among the methods outside of the brain, but these differences are not important, as they are masked out in post processing

stages. She concluded that the probability maps produced by the ANN method were the most representative of normal neuro-anatomy. Based on these recommendations, the brain tissue probability maps created by ANN were used in subsequent experiments. Figure 4.7 demonstrates the differences between the first and second generation tissue probability maps. Notice that the superior resolution of the new TPMs results in greater neuro-anatomical detail.



(a) (b) (c)

Figure 4.4: Cerebro-spinal fluid tissue probability map created by (a) ANN, (b) MD and (c) C4.5 algorithms.

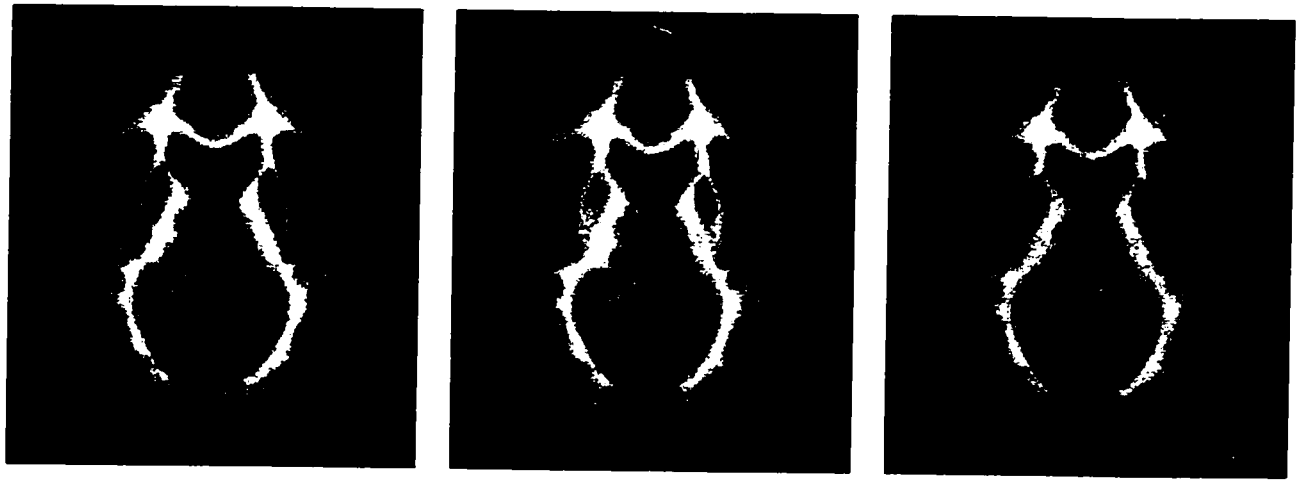
4.4 Concluding remarks

A mechanism of automating the training sample selection process for supervised classification algorithms has been discussed. Previously constructed tissue probability maps were used to build new TPMs that were based on 53 individuals and three different classification methods. The resulting maps were found to have superior spatial resolution. They were examined for neuro-anatomical accuracy by an expert, who chose the tissue probability maps created by the ANN algorithm to be the most accurate. Chapter 7 dis-

(a) (b) (c)

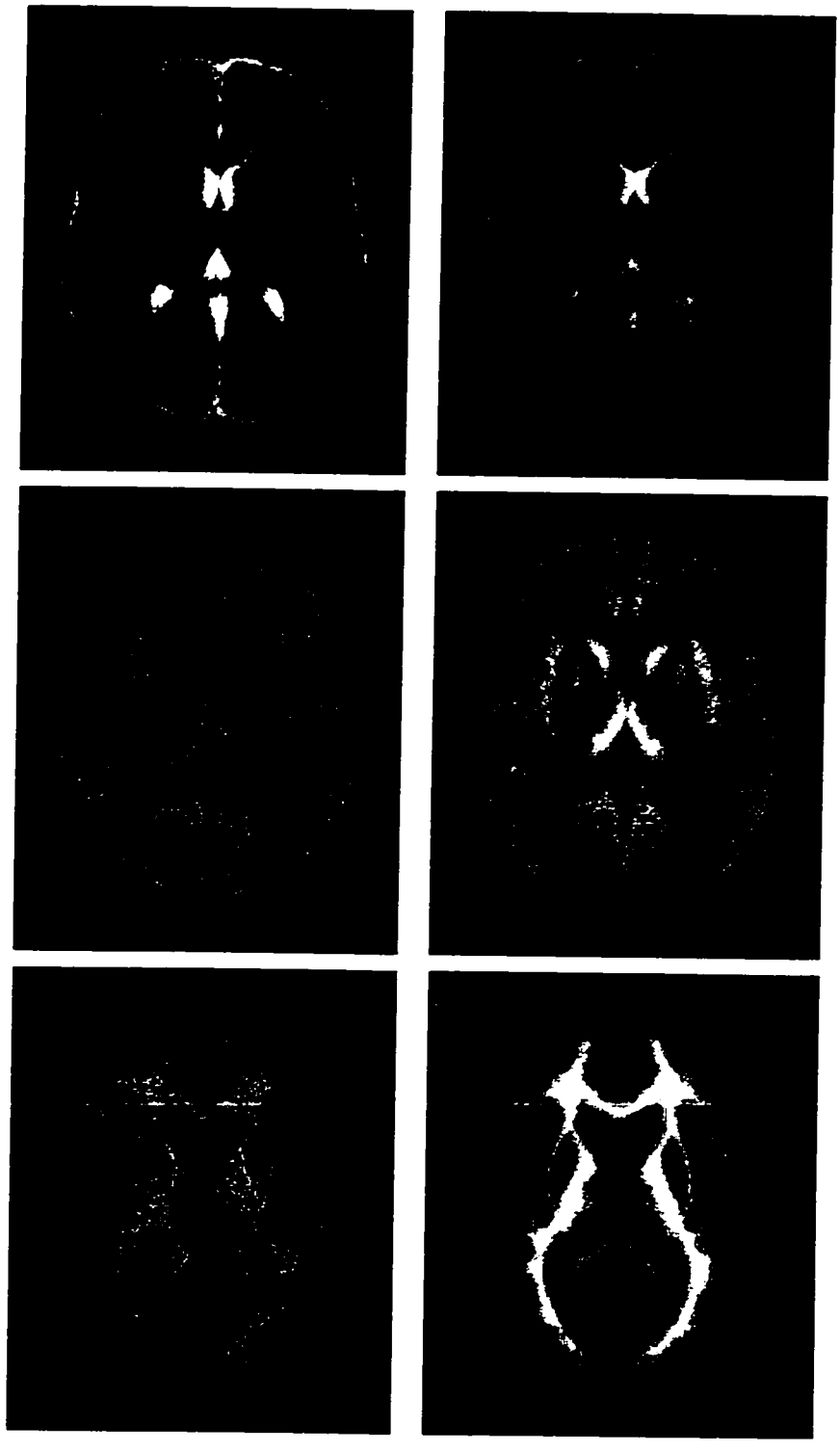
Figure 4.5: Gray matter tissue probability map created by (a) ANN, (b) MD and (c) C4.5 algorithms.

cusses the experiments carried out to establish the usefulness of these tissue probability maps in automating the training of supervised classification algorithms.



(a) (b) (c)

Figure 4.6: White matter tissue probability map created by (a) ANN, (b) MD and (c) C4.5 algorithms.



(a) (b)

Figure 4.7: Sample image slices for comparative view from (a) first generation (b) second generation cerebro-spinal fluid, gray matter and white matter tissue probability maps.

Chapter 5

Classifier Validation Methodologies in MRI

5.1 Introduction

Tissue classification is used to obtain quantitative measures of morphological parameters like brain volume or lesion load, diagnosis of disease and assessment of response to therapy. This chapter describes the experiments and analysis used to assess the reliability and accuracy of the various classifiers described in Chapter 3.

5.2 Similarity measures

In order to rate the performance of a particular classifier, operating under different sets of parameters, it is necessary to compare the resulting classification against a known truth, a *gold standard*, which is a labeled volume (i.e. each voxel having a label to indicate tissue type membership). The challenge is to establish a suitable coefficient of agreement between the standard and a particular classification. This coefficient will serve as a relative measure of performance for that specific classification.

There has been extensive research conducted in diagnostic and radiological circles in measuring agreement between two experts on categorical data. This is especially the case in psychiatry, where several indices have been devised to rate the agreement (or disagreement) among psychiatrists in assessing specific conditions of mental patients [Cohen, 1960; Cohen, 1968; Fleiss, 1975; Bartko and Carpenter, 1976; Hubert, 1977; Williams, 1987; Bartko, 1991]. Some of these studies have performed comparative analysis of different techniques focusing on the validity of these measures of similarity in various fields from psychiatry [Bartko and Carpenter, 1976; Bartko, 1991], to remote sensing applications in geological surveys [Rosenfield and Fitzpatrick-Lins, 1986].

The type of data classified brain volumes fall into, can be considered either *dichotomous*, belonging or not belonging to a particular class, say some form of lesion (+,−), or *polychotomous*, belonging to one of several classes, (CSF, GM, WM, ...).

Comparison of a classified volume against a gold standard can be viewed as measuring agreement between two raters or judges, where one of them (the gold standard) is considered correct. In order to measure agreement between the gold standard and a particular classified volume on dichotomous data, their assessment of each voxel is presented in a tabulated format like that of Table 5.1, where agreement between the volumes is represented by cells a and d, and disagreement by cells b and c on the presence or absence (+, −) of a trait. Some researchers consider a as the number of True Positives (TP) and d as True Negatives (TN), whereas b denotes the number of False Negatives (FN) and c as False Positives (FP)[Williams, 1987].

In the case where polychotomous data is considered, Table 5.1 could be easily extended to contain as many classes as necessary. In Table 5.2, N samples of data (representing an entire volume) are placed into any of T classes designated by $C_1, \ldots, C_T$, where each cell n_{ij} indicates a count that the gold standard places a sample in the i^{th} class and the particular classification places the same sample in the j^{th} class. Agreement between the classified volume and the gold standard is represented by the left-to-right diagonal, cells $n_{ii}, (1 \leq i \leq T)$. The two tables described above are considered to be classification

Gold Standard	Classification		
	+	−	Row totals
+	a	b	a+b
−	c	d	c+d
Column totals	a+c	b+d	N

Table 5.1: Sample dichotomous confusion matrix

Gold Standard	Classification				
	C_1	C_2	...	C_T	Row totals
C_1	n_{11}	n_{12}	...	n_{1T}	n_{1+}
C_2	n_{21}	n_{22}	...	n_{2T}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_T	n_{T1}	n_{T2}	...	n_{TT}	n_{T+}
Column totals	n_{+1}	n_{+2}	...	n_{+T}	$n_{++}(=N)$

Table 5.2: Sample polychotomous confusion matrix

Gold Standard	Classification		
	C_1	$C_2 \quad \ldots \quad C_T$	Row totals
C_1	a	b	$a+b$
C_2 $\vdots$ C_T	c	d	$c+d$
Column totals	$a+c$	$b+d$	N

Table 5.3: Sample polychotomous confusion matrix, collapsed on class C_1.

error matrices, or *confusion matrices*, from which similarity measures can be calculated to indicate degrees of agreement or disagreement between two volumes. Any polychotomous confusion matrix (Table 5.2) can easily be represented as a set of T dichotomous confusion matrices (Table 5.1) by collapsing the matrix on each class of interest. For example, to collapse a polychotomous matrix on a particular class C_i the elements of the respective dichotomous matrix can be calculated as follows:

$$\begin{aligned} a &= n_{ii} \\ b &= n_{i+} - a \\ c &= n_{+i} - a \\ d &= N - (a+b+c) \end{aligned} \tag{5.1}$$

Thereafter, coefficients of agreement for each individual class can be calculated from the respective collapsed dichotomous matrix for that particular class. Table 5.3 shows a polychotomous matrix collapsed on class C_1

5.2.1 Traditional measures of similarity

Bartko [Bartko and Carpenter, 1976] defines *reliability* as the consistency with which a given trait is measured, and *validity* as the extent by which an index measures a trait. He provides an extensive critique of the frequent misuse of numerous measures of similarity such as percent agreement, chi-square, product moment correlation, and rank order correlation. Numerous other measure like Yule's Y [Yule, 1912] have also been considered. Of the above, **percent agreement**, P_0, deserves particular attention, since it has been used extensively in the literature to designate classification accuracy [Rosenfield and Fitzpatrick-Lins, 1986; Bartko and Carpenter, 1976]. It is defined as follows:

$$P_0 = \sum_{i=1}^{T} n_{ii}/N \quad \forall_i \tag{5.2}$$

Williams [Williams, 1987] also defines *Accuracy* as a function of percent agreement (distributed over two terms called *Specificity* and *Sensitivity*), defined as follows:

$$Sensitivity = TP/S \tag{5.3}$$

$$Specificity = TN/H \tag{5.4}$$

$$Accuracy = \frac{TP+TN}{S+H} \tag{5.5}$$

where $S = TP + FN$ and $H = FP + TN$. Note that *Accuracy* can also be represented by the following expression:

$$Accuracy = (Sensitivity)\left(\frac{S}{S+H}\right) + (Specificity)\left(\frac{H}{S+H}\right) \tag{5.6}$$

Looking closely at Equation 5.5, one can see that it is the same as Equation 5.2, since the term $TP + TN$ is the same as $a + d$, which is the sum of left to right diagonals represented as $\sum_{i=1}^{T} n_{ii}$ and the term $S+H$ is the same as N. Intuitively, percent agreement gives the illusion of being a proper measure of similarity. However, it gives inflated confidence levels since it does not consider that there could be some agreement due to chance alone. A more sophisticated measure of agreement that is chance-corrected, and can function as interclass correlation coefficient [Fleiss, 1975; Bartko and Carpenter, 1976], is considered in Section 5.2.2.

5.2.2 Kappa

Cohen [Cohen, 1960] has developed a coefficient of agreement called *kappa*, which is a chance-corrected percent agreement with a statistical base [Bartko, 1991]. *Kappa* is defined as:

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{5.7}$$

where P_0 is defined in equation. 5.2 and

$$P_c = \sum_{i=1}^{T} n_{i+} n_{+i} / N \tag{5.8}$$

P_0 is the proportion of samples agreed to by the two raters, and P_c is the proportion of agreement due to chance, where n_{i+} are the row totals and n_{+i} the column totals for each class (table 5.2). $\kappa = 1$ is obtained whenever there is perfect agreement between the two raters. $\kappa = 0$ indicates that agreement is due to chance alone. κ between 0 and 1 reflects agreement greater than chance, whereas negative values indicate agreement less than chance. It should be noted that all disagreements in *kappa* are treated equally. If it is

necessary that certain disagreements be treated more serious than others, Cohen [Cohen, 1968] has also developed *Weighted Kappa*, a similarity measure that takes into consideration the seriousness of disagreement. In this thesis, all disagreements were considered of equal weight.

Fleiss [Fleiss, 1975] gives a critical review of a dozen or so measures of similarity, emphasizing the need for chance correction to be incorporated into such measures. He also points out that no index of agreement is informative by itself, and any index should be expressed in terms of agreement greater or less than chance alone. He concludes that *kappa* is one similarity measure that is justifiable as chance-corrected; it also serves as an interclass correlation coefficient.

Most coefficients of agreement take into consideration the entire confusion matrix and give a single value denoting similarity. However, sometimes it is necessary to have measures of similarity that concern each class by itself, or combine specific classes together. For example, one might be interested in obtaining classifier accuracy for cerebro-spinal fluid (one particular class), or brain parenchyma, where gray matter and white matter are combined into a single class. It might be necessary for a particular class be excluded from the calculation of coefficient of agreement (like background). Bishop [Bishop et al., 1975] derived a method where the numerator and denominator of overall *kappa* can be calculated by summing respective numerators and denominators from classes of interest in a confusion matrix defined as follows:

$$\hat{\kappa_i} = \frac{Nn_{ii} - n_{i+}n_{+i}}{Nn_{i+} - n_{i+}n_{+i}} \tag{5.9}$$

Above, $\hat{\kappa_i}$ is the the maximum likelihood estimate of the conditional agreement between observers for a given category. In this manner, *kappa* per class or per several classes, can be calculated. Rosenfield [Rosenfield and Fitzpatrick-Lins, 1986] used *kappa* in the field of remote sensing applications in geological surveys, noting that *kappa* could be used as a measure of accuracy as a whole or on a per category basis. He concluded

Class Label	Classification				
Gold Standard	0	1	2	3	Row totals
0	49172	0	0	0	49172
1	0	37141	16146	434	53721
2	0	179	31986	8869	41034
3	0	3590	13986	34882	52458
Column totals	49172	40910	62118	44185	196385

Table 5.4: A Sample confusion matrix

	Measures of Agreement					
Class	*Kappa*(κ)	*Collapsed*(κ)	*Sensitivity*	*Error*	*Specificity*	*Accuracy*
0	1.000	1.000	1.000	0.000	1.000	1.000
1	0.610	0.718	0.691	0.309	0.974	0.896
2	0.678	0.492	0.780	0.221	0.806	0.801
3	0.568	0.632	0.665	0.335	0.935	0.863
Total	0.708	0.708	0.780	0.220	*n/a*	*n/a*

Table 5.5: Various measures of similarity based on Table 5.4

that *kappa* should be adopted by the remote sensing community as a measure of accuracy for thematic classification.

5.2.3 An example

Table 5.4 represents a sample confusion matrix obtained from an actual comparison of a classified volume with a gold standard, and Table 5.5 shows a variety of similarity indices calculated, based on results presented in Table 5.4. It is interesting to note that Accuracy (last column) reports inflated values compared to *kappa* (first column) and Collapsed κ (second column, calculated by collapsing the polychotomous matrix into a dichotomous

Class Label	Classification				
Gold Standard	0	1	2	3	Row totals
0	12220	12144	12136	12443	48943
1	12265	12347	12258	12272	49142
2	12212	12222	12504	12113	49051
3	12372	12359	12247	12271	49249
Column totals	49069	49072	49145	49099	196385

Table 5.6: A Sample confusion matrix from random image volumes.

one, Equation 5.1) for individual classes.

In order to understand the importance of chance correction in a similarity measure, two randomly generated volumes were created using a random number generator. The resulting confusion matrix is given in Table 5.6, and the calculated similarity measures in Table 5.7. Accuracy gives high similarity values even though the data are completely random volumes. *Kappa*, taking into consideration chance correction, reports almost zero. Note that in Tables 5.5 and 5.7, column *Total* values for *specificity* and *sensitivity* are not applicable (indicated by n/a), as they are per class indices.

Given these results, and the recommendations from several critical reviews [Rosenfield and Fitzpatrick-Lins, 1986; Fleiss, 1975; Bartko and Carpenter, 1976], *kappa* was adopted as the measure of similarity between particular classifications and the gold standard in experiments conducted for this thesis.

5.3 Validation methods

As mentioned previously, validation of brain tissue classification is a critical step in assessing the performance of different classifiers under varying conditions of MR imaging. And since MR imaging, like many other medical imaging modalities, is an *in vivo* study, vali-

	Measures of Agreement					
Class	*Kappa*(κ)	*Collapsed*(κ)	*Sensitivity*	*Error*	*Specificity*	*Accuracy*
0	−0.000	−0.000	0.250	0.750	0.750	0.625
1	0.002	0.002	0.251	0.749	0.751	0.626
2	0.006	0.006	0.255	0.745	0.751	0.627
3	−0.001	−0.001	0.249	0.751	0.750	0.624
Total	0.002	0.002	0.251	0.749	*n/a*	*n/a*

Table 5.7: Various measures of similarity based on Table 5.6

dation becomes even more challenging. Numerous techniques have been used by several researchers, each having its own advantages and disadvantages. This section will briefly describe and comment on some of the various techniques of validation found in MR image segmentation literature [Zijdenbos and Dawant, 1994; Clarke et al., 1995].

5.3.1 Validation using physical phantoms

Several researchers have used physical phantoms, which are cylindrical structures with compartments of known volumes, sometimes roughly shaped like human brains, containing different paramagnetic substances to imitate various tissue relaxation parameters [Cline et al., 1991; Gerig et al., 1992: Jackson et al., 1993; Mitchell et al., 1994].

Physical phantoms, though convenient to use, have several disadvantages. The spatial tissue distribution in the brain, with its high geometric complexity, makes the construction of such compartmentalized physical phantoms very difficult, if not impossible. Therefore multiple class distribution and partial volume cannot be easily represented. Furthermore, when simulating pathology, compartments constructed to reflect tumors or lesions, are simplistic in shape and over-estimated in volume, owing to practical construction limitations. Moreover, physical phantoms when placed in an MR scanner affect the main magnetic field differently than human subjects do, and consequently induce different RF

non-uniformities. Owing to these drawbacks, physical phantoms with their simplistic nature of construction, are of limited use in validating the performance of classifiers when confronted with the complexities and intricacies of human brains.

5.3.2 Validation using gross anatomy and histo-pathology

Another validation method could potentially involve correlation of classified MR images with histo-pathological or post-mortem gross anatomical examinations. Some researchers [Taxt et al., 1992] have resorted to histo-pathology of surgically removed tumors to validate classifier performance on MR images acquired prior to the excision. Even though these methods might seem appropriate for volume measurement, they cannot confirm shape and location of regions of interest. They are limited to pathological tissues that are marked for excision during surgery. Besides being labor intensive, logistic difficulties of post-mortem analysis such as feasibility and proper excision, make this method of validation highly impractical, especially when considering healthy tissues.

5.3.3 Validation using manual labeling

One of the most intuitive gold standards is expert opinion, where an expert outlines regions of interest or classifies an entire MR image manually. The resulting image is considered to be the gold standard against which automated methods are compared. Several researchers [Vannier et al., 1991; Gerig et al., 1992; Zijdenbos et al., 1994] have relied on experts to validate classification. However, this method has potential problems. Studies of inter-operator variability, have demonstrated that sufficient variation (as much as 40% in some cases [Zijdenbos et al., 1994]), makes the determination of a gold standard difficult. One can consider the computer's classification as yet another expert opinion. Thereafter, one can resort to correlation studies to see if there is sufficient agreement among the experts, and to what extent does the computer's interpretation agree with the human experts. This method of validation is acceptable, provided that a sufficient

number of expert opinions are polled to establish a statistically significant result. This labor-intensive validation method is difficult to implement for large numbers of data sets and with an adequate number of experts.

5.3.4 Validation using test sets

Traditionally in pattern recognition problems, in order to validate the performance of classifiers, there are usually *test sets* (disjoint from training sets) in addition to the training sets, where the correct class of data to be classified is known prior to classification. Such tests can be used to validate the performance of the classifier.

In brain imaging, the test and the training sets have to be provided manually by an expert, who selects image voxels that belong to particular classes. Even though this method is fairly straight forward, it suffers from the same subjectivity problems as validation using manual labeling, since only voxels or regions of typical intensities are usually chosen, excluding partial volume cases. This leads to the test set being underrepresented.

5.3.5 Validation using digital phantoms and MR simulators

Computer simulations in numerous research fields have been extremely helpful in modeling the behavior of systems under study and the parameters under which they operate. MR imaging research has been no exception. MR simulators have played a key role in pulse sequence research, modeling of noise, and study of RF inhomogeneity.

The output of MR simulators makes an excellent candidate for classification validation, where robustness of a particular classifier can be tested by first corrupting the simulated image with noise, RF inhomogeneity, and degrade resolution. Also, pathology can be introduced in a systematic fashion to simulate tumors, lesions, trauma, etc. In this manner, one source of uncertainty can be introduced at a time, and the resulting classification uncertainty can be related to the source in a quantifiable manner. The simulation process starts from a *master digital phantom* (MDP), which is a labeled brain

image volume indicating "truth," where each voxel is pre-classified to belong to different tissue types to reflect proper neuroanatomy. Regardless of any errors in the classification process, the results are taken as correct by definition from that point on. Various forms of image degradation can then be added to simulate the effects of the scanner during data acquisition, but this truth remains constant. The disadvantage of this system is the fact that some subtle non-linearities found in MR imaging systems, such as gradient field inhomogeneities cannot simply be abstracted mathematically [Peterson et al., 1993].

It is important to note that the use of simulated and real MR image volumes is necessary but not sufficient condition to validate classification. This is because in the case of MR simulation, the resulting simulated images are somewhat simplistic. However, classification of simulated images can be compared against a gold standard. In the case of real MR volumes, they truly represent a real world condition, but they lack a gold standard against which, their resulting classifications could be compared.

5.4 Concluding remarks

After reviewing numerous similarity measures, while assessing their strengths and weaknesses, *kappa* was the preferred similarity index. Furthermore, several validation methodologies have been described, each having advantages and disadvantages. Most methods suffer from bias and variability, some are invasive in nature. In the absence of a true gold standard, MR simulators with their flexibility, sophistication, and ease of use, are favored over other, more complicated, labor-intensive methods of validation. The following chapter discusses the motivation and the development of a *Simulated Brain Database* to be used in the validation process.

Chapter 6

Simulated Brain Database

6.1 Introduction

This chapter explains the motivation behind the creation of a **Simulated Brain Database** (SBD) that was used to carry out the validation experiments designed to meet the objectives listed in Chapter 1.

6.2 Motivation

As discussed in Section 5.3.5, MR simulators are used to create simulated brain images that include various aspects of MR imaging, such as noise, RF inhomogeneity and slice thickness (Chapter 2). By varying specific MR imaging parameters, a set of brain images can be simulated, thus creating an image database. This database can be used to test medical image processing and pattern recognition algorithms ranging from classification to RF correction, all in a controlled environment. For example, by varying parameters that control RF inhomogeneity, a range of simulated MR images containing different levels of RF inhomogeneity can be created. Since the source of all the simulations is a single digital phantom (which is a labeled brain volume representing the gold standard), one can

establish the sensitivity of any particular classification algorithm with respect to levels of RF inhomogeneity in an MR image, classifier accuracy can be determined as a function of RF inhomogeneity. In this fashion, by testing several classifiers, one can tell which classifier is more resistant or susceptible to varying levels of RF inhomogeneity.

6.3 Creation of the database

Prior to the construction of the database, it is necessary to establish the baseline conditions of the varying MR imaging parameters that can be adjusted as follows:

- Estimate normal levels of noise, inherently present in MR images
- Estimate typical RF inhomogeneities for different imaging sequences
- Establish different levels of slice thicknesses that are part of clinical and research imaging protocols

6.3.1 The MR simulator

The MR simulator [Kwan et al., 1996] used in this thesis performs a discrete-event simulation of MR signal production. It is composed of two modules; *signal production*, and *image production*. The signal production is accomplished by the simulation of different pulse-sequence effects on tissue magnetization state (Chapter 2). The results of the signal production are used by the image production module to generate the simulated MRI image volume. A fuzzy digital phantom (Section 6.3.2), indicating the spatial probabilistic distribution of tissue types, is used to calculate the signal intensities at each voxel position to produce the simulation. Depending on the specified sampling resolution, partial volume effects are subsequently evaluated. Signal reception is modeled through the use of an RF-coil model, which takes into consideration the effects of noise and RF inhomogeneities, according to specified imaging parameters.

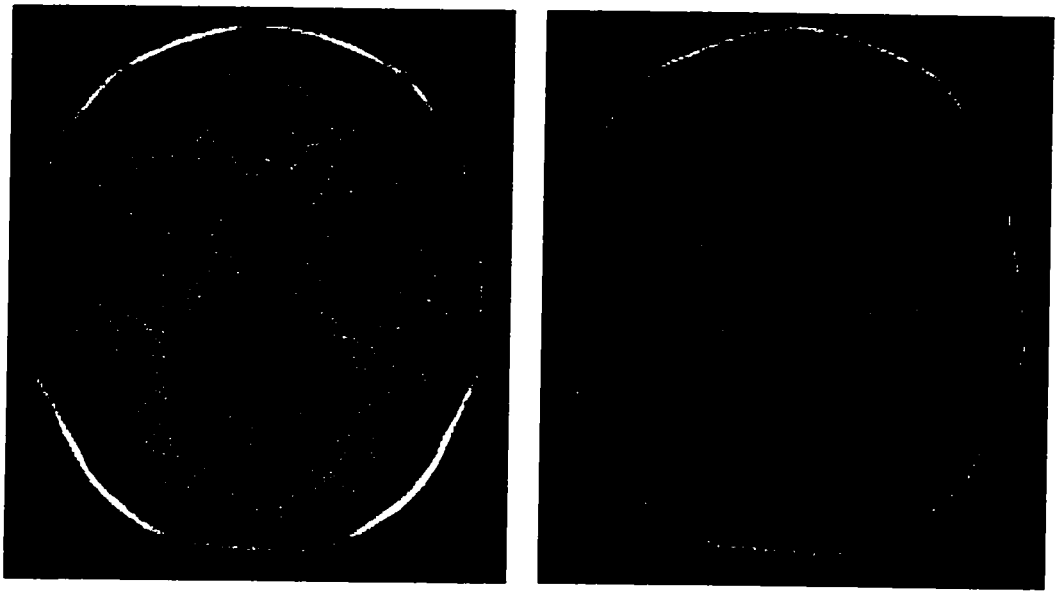
6.3.2 Construction of the fuzzy digital phantom

In order to simulate MR images, the MR simulator [Kwan et al., 1996] needs a *fuzzy digital phantom* that serves as the standard from which different MR images of varying conditions can be simulated. This fuzzy phantom is a probabilistic labeled brain image volume, where each voxel value is the probability of its belonging to different tissue types. In order to create such a phantom, a healthy volunteer was scanned 27 times using T_1-weighted imaging sequence. All the individual MR volumes were transformed into stereotaxic space (Section 4.2) to make them suitable for comparison and averaging. The transformed volumes comprised 181 slices of 217 x 181 voxels, each having a size of $1mm^3$. The 27 T_1-weighted volumes were co-registered in 3D [Collins et al., 1994] and intensity averaged into a single volume to increase the signal to noise ratio. Superior neuro-anatomical detail was visible in the resulting average volume.

Figure 6.1a is a transverse slice, demonstrating the quality of the MR images obtained by averaging 27 volumes from the same subject. Compared to Figure 6.1b, which is one of the 27 volumes, it exhibits superior neuro-anatomical detail that can be classified easily to capture the complex structure of the human brain.

From this average volume, and with the help of the neuro-anatomist at the MNI, a carefully selected training set was chosen as a starting point representing four classes; background, cerebro-spinal fluid, gray matter, white matter. Then the average MR volume was classified using fuzzy Minimum Distance algorithm into the above mentioned four classes. It is important to note that any classifier giving fuzzy class membership could have been used. The reason fuzzy Minimum Distance algorithm was chosen because qualitatively, the resulting fuzzy volumes (representing each class) had less sharp boundaries between different tissue types. Furthermore, visually it looked more realistic. As discussed in Section 5.3.5, it should be noted that errors resulting from mis-classification of the MDP do not matter, since they are considered by the simulator to be correct by definition.

The resulting classified volume was carefully analyzed by the neuro-anatomist, who



(a) (b)

Figure 6.1: (a) An image slice from the average of 27 T_1-weighted image volumes, demonstrating superior signal to noise ratio. (b) an image slice from a single T_1-weighted image volume.

corrected the numerous areas that the algorithm mis-classified. Subsequently, the following six tissue types; fat, muscle, skin, skull, glial matter (a structure around the ventricles) and connective tissue were manually added to the above mentioned corrected volume.

This volume served as the digital phantom that had correct neuro-anatomical representation (the gold standard) from which all MRI simulations were generated. The advantage of using fuzzy, rather than discrete digital phantom, is the fact that each voxel has anatomical partial volume information, represented in terms of probabilistic values for each tissue type. This is because at a microscopic level, the transition from one tissue type into another in the brain, is not sharp, but happens progressively. For example, a particular voxel might be 70% GM and 30% WM and 0% CSF. Figure 6.2 shows transverse slices from this fuzzy digital phantom, for each of CSF, GM and WM classes. Figure 6.3 shows how the fuzzy phantoms (like the ones in Figure 6.2) representing the ten tissue classes, can be combined into a discrete digital phantom suitable for classification comparison.

This is accomplished by assigning each voxel to the most probable tissue class from the fuzzy volumes.

(a) (b) (c)

Figure 6.2: Transverse image slices of fuzzy (a) CSF, (b) GM and (c) WM phantoms.

Figure 6.4 shows a real and a simulated MR image and Figure 6.5 shows MR simulations of T_1-, T_2- and PD-weighted image volumes.

6.3.3 Estimation of normal noise levels

As mentioned in Chapter 2, noise from a variety of sources is present in MR images. In order to create a series of simulations of varying noise levels, it is necessary to estimate the normal levels of noise present in MR images for a particular scanner. Within the simulator, definition of signal–to–noise (SNR) ratio is given by the following relation [Nishimura, 1993]:

$$SNR = \frac{signal\ amplitude}{standard\ deviation\ of\ noise} \tag{6.1}$$

Figure 6.3: Discrete version of the digital phantom, produced by combining several fuzzy phantoms.

Signal amplitude is represented by voxel intensities of the brightest tissue class, which for T_1-weighted images is the white matter tissue, and for T_2- and PD-weighted images it is cerebro-spinal fluid.

MR image volumes from five healthy volunteers obtained using the Philips Gyroscan ACS-2 scanner at the MNI, were examined to estimate the level of noise present in typical scans. This was done by calculating the mean intensity of the brightest tissue class for respective T_1-, T_2- and PD-weighted imaging sequences and the standard deviation of the voxel intensities in the background, by taking sample voxels in respective regions of interest (Equation 6.1). The resulting noise levels ranged between 2 to 4%. Therefore 3% was taken to be the level of noise normally present in MR images at the MNI.

Since MR scanners are constantly being improved, there exists a huge population of MR scanners with wide ranging conditions depending on age, sophistication and field strength. In the simulated brain database, noise levels were 0%, 1%, 3% (normal noise),

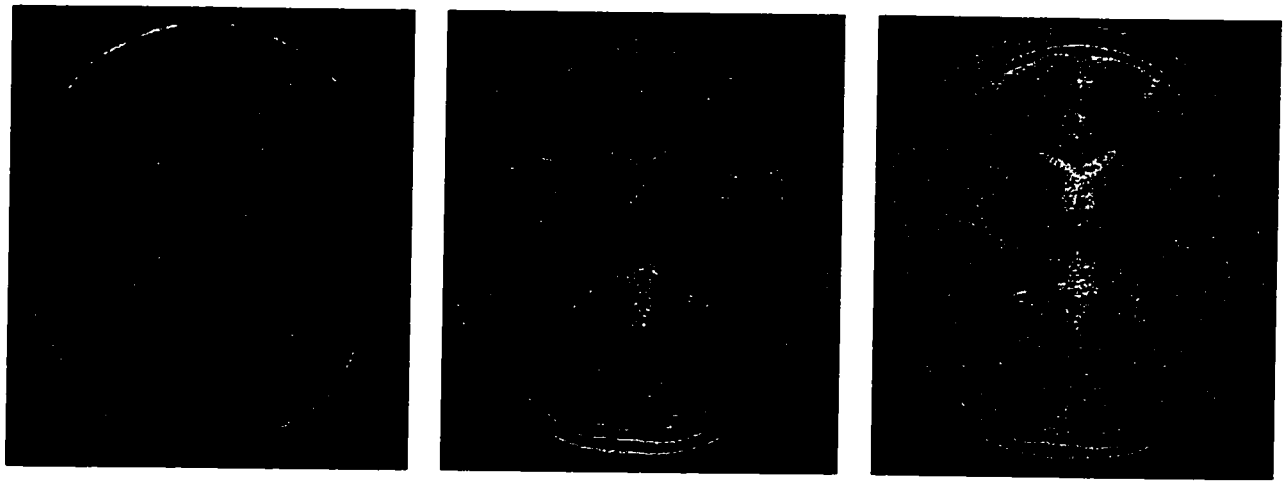
(a) (b)

Figure 6.4: (a) A real MR image and (b) a simulated MR image. Note that the simulated image has no Choroid Plexus in the lower ventricles (a structure responsible for producing CSF, pointed to by arrows in the first image). This is because in the digital phantom, this structure was misclassified as CSF.

5%, 7%, and 9%. With the present MR scanners, 9% noise is considered to be very noisy. Figure 6.6 shows T_1-weighted MR simulations with 3 levels of noise, demonstrating the resulting image degradations.

6.3.4 Estimation of typical RF inhomogeneities

RF inhomogeneity introduces a slowly-varying intensity non-uniformity over the whole image. In order to estimate RF inhomogeneities present in different MR images, it is necessary to apply an RF correction method to each set of images, in order to estimate typical RF field variations. Since RF inhomogeneities are pulse-sequence-and-subject dependent, a triplet set of T_1-, T_2- and PD-weighted image volumes of the healthy volunteer (who supplied the 27 T_1-weighted image volumes) were obtained. A spline curve-fitting

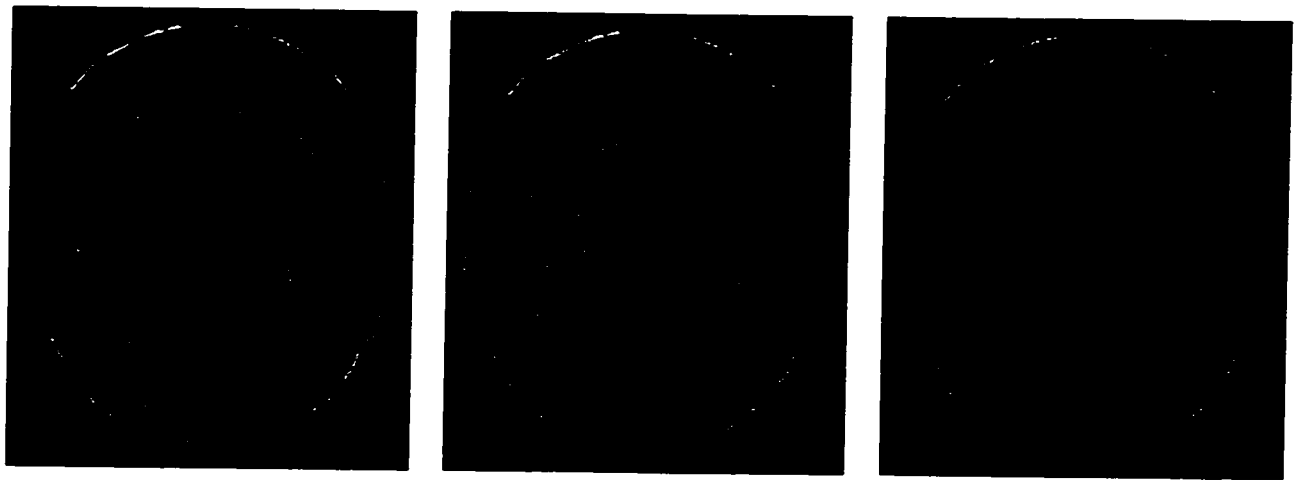


(a) (b) (c)

Figure 6.5: Sample simulated (a) T_1-weighted, (b) T_2-weighted and (b) PD-weighted images using the phantom of Figure 6.2, with normal levels of noise and RF inhomogeneity.

method was used to correct for RF inhomogeneities, where 500 reference points per tissue class were chosen, spanning the entire image volume. Afterwards, a thin-plate spline surface was fitted to the intensity values of these reference points [Dawant et al., 1993]. This fitted thin-plate spline was subsequently used to correct the volumes for RF inhomogeneities. Figure 6.7 shows sample fields obtained for each of T_1-, T_2- and PD-weighted image volumes, using the above mentioned RF correction method.

The RF fields are represented as magnitude images, ranging in values centered around 1.0, where the minimum and the maximum of these magnitude images signify the level of inhomogeneity. For example, an RF field might have magnitude values ranging between 0.8 and 1.2, signifying a 40% change in intensity values across an image volume. In establishing normal levels of RF inhomogeneity, MR images from 12 different MR scanners across North America were corrected for RF inhomogeneities. These images were obtained as part of the AutoImmune project; a drug trial study for Multiple Sclerosis, taking place at the MNI. The probability densities of the resulting RF fields were plotted as a function of field strength (Figure 6.8). Figure 6.9 shows horizontal and vertical intensity profiles



(a) (b) (c)

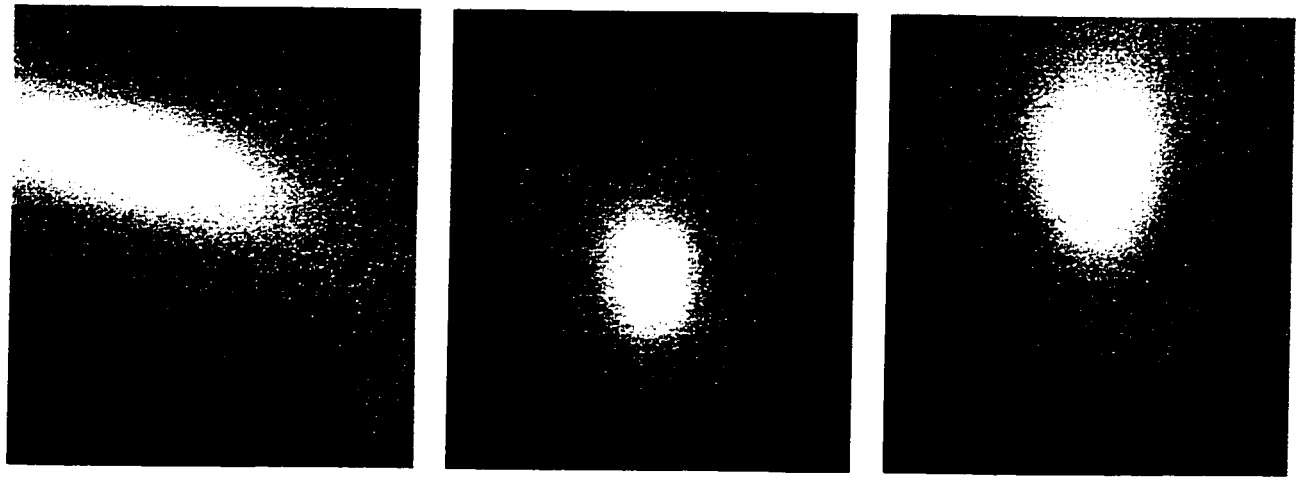
Figure 6.6: Sample simulated T_1-weighted images with (a) 0%, (b) 3% and (c) 9% noise.

across the fields obtained in Figure 6.7. It can be seen that the majority of RF field variations were between 0.9 and 1.1, signifying 20% variation in intensity. Therefore 20% was determined to represent the magnitude of normal RF inhomogeneity.

In order to vary the severity of the RF inhomogeneity, one can adjust the field minimum and the maximum, while preserving the shape of the fields and their probability densities, to obtain desired variations. In creating the simulated brain database, six different levels of RF were considered: 0% , 10%, 20% (normal RF), 30%, 40% and 50%. Figure 6.10 shows two T_1-weighted images with no RF and 50% RF inhomogeneity. Notice the level of intensity variations found across the upper and lower right corner of the second image.

6.3.5 Selection of slice thicknesses

Slice thickness selection is normally established by clinical and research protocols that dictate time constraints and the resolution at which a particular study is to be carried out. Typically, the slice thickness of MR scans for clinical purposes ranges between 3–5 millimeters. On the other hand, research protocols, requiring higher resolution, usually



(a) (b) (c)

Figure 6.7: Sample RF inhomogeneity fields generated by correcting (a) T_1-weighted, (b) T_2-weighted and (c) PD-weighted image volumes using a method described in Section 6.3.4.

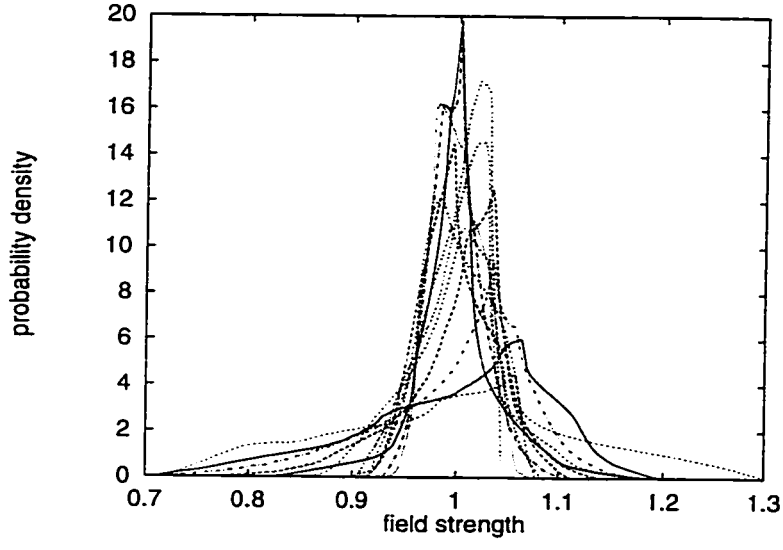
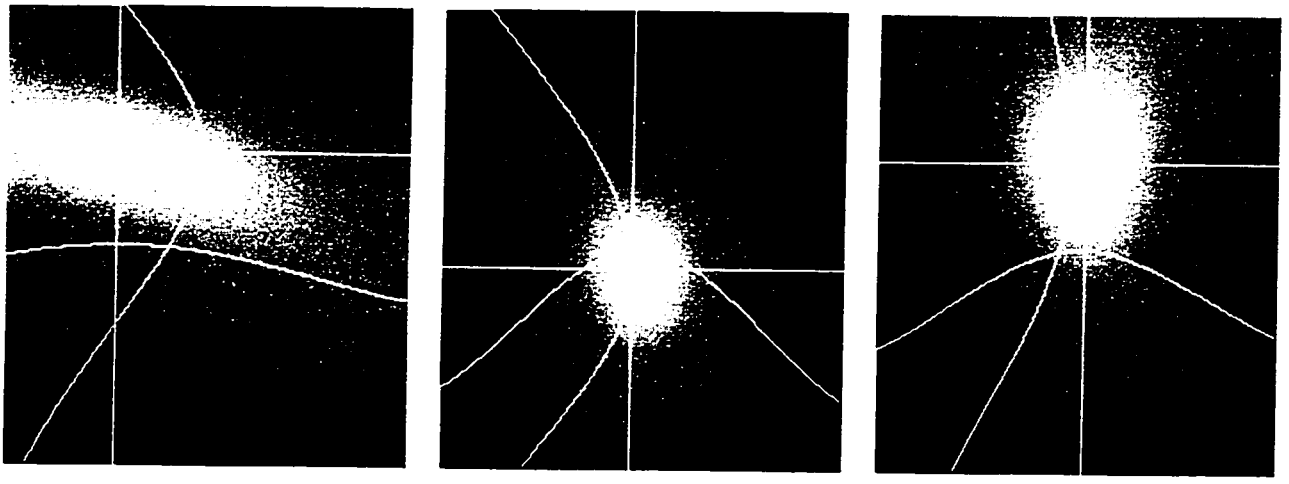


Figure 6.8: RF field strengths as a function of probability density from 12 different MR scanners.

acquire images at 1mm slice thickness. As slice thickness increases, partial volume becomes a significant factor, where adjacent tissue types are grouped together into a single voxel. In order to study the effects of partial volume of the MR scanner, the SBD con-



(a) (b) (c)

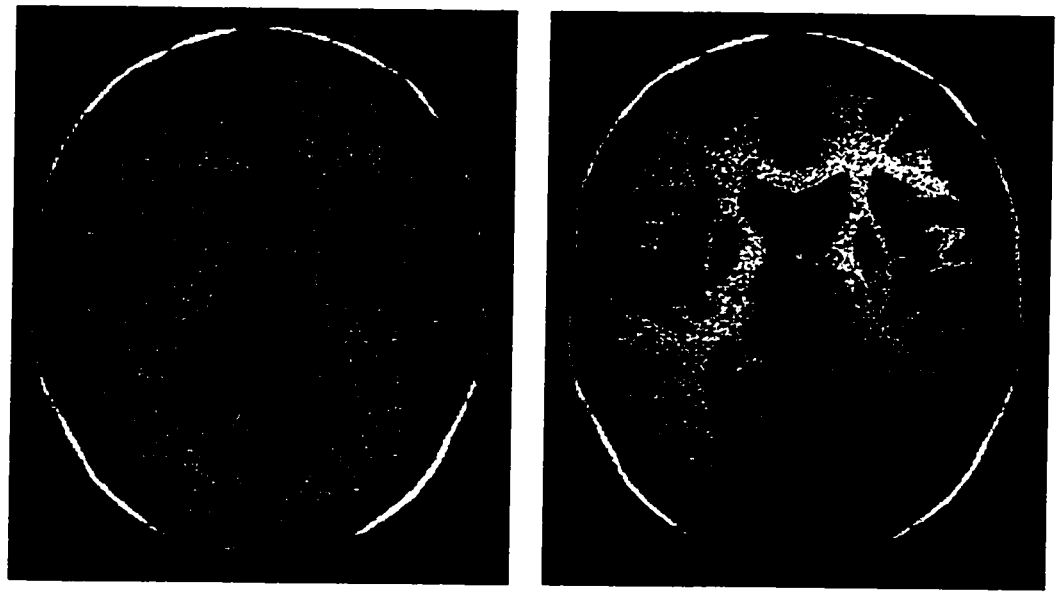
Figure 6.9: RF inhomogeneity field intensity profiles of (a) T_1-weighted, (b) T_2-weighted and (c) PD-weighted image volumes of Figure 6.7.

tains images of 1mm, 3mm and 5mm slice thicknesses. Figure 6.11 shows T_1-weighted MR simulations with 1mm, 3mm and 5mm slice thicknesses, demonstrating the effects of partial volume.

6.3.6 Creation of the database

Having established the baseline values of noise and RF inhomogeneities, and the range over which these parameters will vary, it is helpful to put into perspective the different permutations of image parameters. There are 3 imaging sequences (T_1-, T_2- and PD-weighted), 6 levels of noise, 6 levels of RF inhomogeneity, and 3 slice thicknesses (1mm, 3mm, 5mm). Together, the combinations yielded potentially 324 different image volumes, requiring considerable storage, processing and analysis (each volume occupies more than 14MB of disk space for a total of 4.5GB of storage).

In order to avoid the combinatorial explosion of imaging parameters in a meaningful way, a systematic approach had to be followed to limit the variations in the noise, RF



(a) (b)

Figure 6.10: Sample simulated images with (a) no RF and (b) 50% RF inhomogeneity. Note the intensity variations in the white matter of the second image, as compared to the first.

inhomogeneity and slice thickness. The essence of the MR simulation experiments was to study the impact of different parameters of MR imaging on brain tissue classification. To do that, one parameter was systematically varied, while keeping all others constant. For example, noise levels and slice thickness should be kept to typical normal values, when studying the effects of RF inhomogeneity on different classifiers. The database was generated such that the six levels of noise were created at normal levels of RF inhomogeneity and slice thickness. Furthermore, six levels of RF inhomogeneity were created at normal levels of noise and slice thickness. Finally three slice thicknesses were generated at normal levels of noise and RF inhomogeneity. Since three imaging sequences (T1, T2, PD) were considered in all combination, there were 39 image volumes. These were far fewer image volumes as compared to 324 combinations. These, however, could be managed in a realistic fashion.

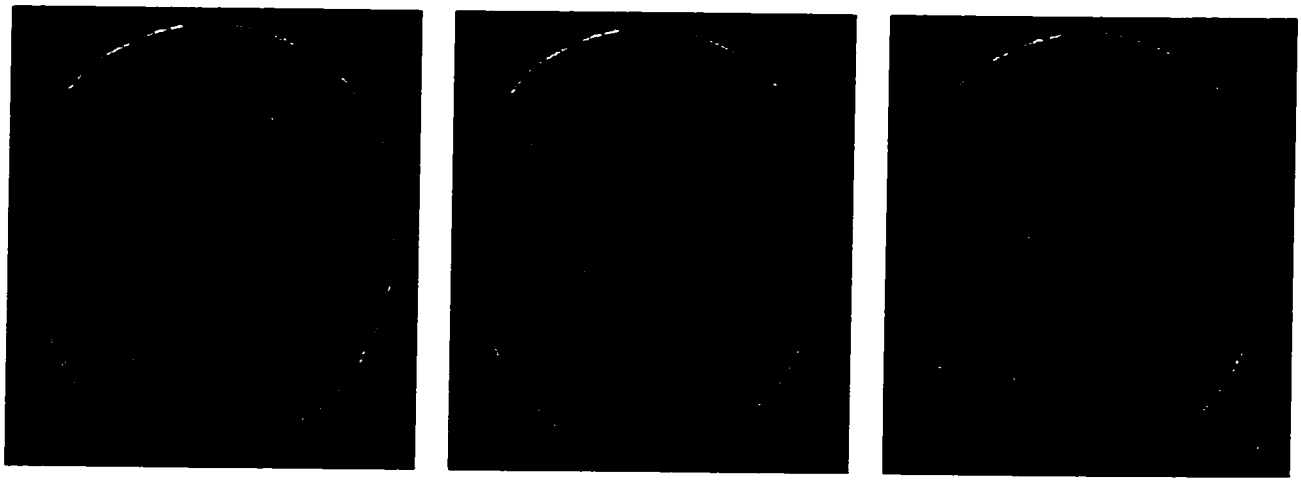


Figure 6.11: Sample simulated T_1-weighted images with (a) 1mm, (b) 3mm and (c) 5mm slice thickness. Note the level of blurring between GM and WM class borders as the slice thickness is increased.

6.4 Concluding remarks

This chapter has discussed the motivation and the creation of the simulated brain database in order to provide a controlled environment for measuring the effects of MR imaging parameters on image processing and pattern recognition algorithms. For the purpose of this thesis, such a database was invaluable in providing a systematic means of studying effects of noise, RF inhomogeneity and slice thickness on various tissue classification algorithms (Chapter 3), as well as for testing the automation of training sample selection (Chapter 4).

Chapter 7

Experimental Results

7.1 Introduction

This chapter describes the experiments carried out to establish the usefulness of the tissue probability maps in automating the training set selection process, and to compares the performance of five supervised and two unsupervised classification algorithms under varying MR imaging conditions.

7.2 Rationale and design of the experiments

The experimental program was designed to determine the feasibility of automating supervised classifications by means of tissue probability maps, and compare their performance (alongside that of experts) under varying conditions of MR imaging. These goals lead to the following specific questions:

- How similar are automatically generated training sets to those manually chosen by experts?

- How sensitive is each individual classifier to manually and automatically generated training sets under normal conditions of MR imaging?
- How do the classification results obtained by using automatically generated training sets compare to those obtained from manually chosen ones?
- How do increased levels of noise, RF inhomogeneity and slice thickness affect the performance of each classifier?
- Which classifiers are most sensitive to each of the above conditions? Which one functions best overall ?
- Does the highest ranking algorithm based on simulated MR data sets perform as well when classifying real MR data sets?

7.3 Methods

The following sections describe the experimental methods employed in carrying out the experiments designed to answer the questions posed in the previous section.

7.3.1 Brain volume data sets

All classifications were carried out using three features (triplets); T_1-, T_2- and PD-weighted image volumes. Chapter 6 described the creation of the simulated brain database that provided the brain volumes of varying conditions of MR imaging. Five separate sets of brain images were used to conduct the experiments described in this chapter. These were:

1– A triplet set of simulated normal brain volumes.

2– Six simulated brain volume triplet sets with varying levels of noise, (Section 6.3.3).

3– Six simulated brain volume triplet sets with varying levels of RF inhomogeneity, (Section 2.4.2).

4– Three simulated brain volume triplet sets with varying slice thicknesses (Section 6.3.5).

5– A single triplet set of a real brain.

7.3.2 Training sets

In this chapter, the term *trainer*, denotes either a human expert or an automated method of choosing a set of training samples. Several such sets were obtained through human and automatic trainers as follows:

Manually selected training sets

In order to obtain manual training samples, human operators with neuro-anatomical knowledge used interactive tools to view a particular brain volume, and choose a set of voxels from various tissue classes. Three operators (*Expert 1, Expert 2, Expert 3*), volunteered their services. Six manually selected training sets containing 25 and 50 sample voxels from each tissue type were obtained from the simulated normal T_1-weighted brain image volume. The experts preferred T_1-weighted images because of good contrast between gray and white matter tissues.

Automatically generated training sets

Chapter 4 described the creation of the brain tissue probability maps. Five probability thresholds (100%, 90%, 80%, 70%, 60%), were chosen to automatically generate ten training sets, also containing 25 and 50 training samples, by the sub-sampling method explained in Section 4.3.1.

7.3.3 Index of performance

Classification performance was determined by the index of agreement *Kappa*, measured between individual classifications and the MDP from which the simulations were generated. Kappa values (Section 5.2.2) were calculated over all tissue classes, spanning the entire brain volume, but excluding areas outside the brain. The kappa values obtained in the following experiments were interpreted in a relative setting.

7.3.4 Classification parameters

Artificial neural nets

The learning rate and momentum of a neural network affects the rate at which it convergences in training. The momentum prevents the network from stopping at a local minimum in its gradient decent [Caudill and Butler, 1992]. The number of hidden neurodes in the network should be large enough to form decision boundaries necessary to classify the different tissue types. However, choosing a very large number of hidden neurodes might be counter-productive, as the weight distribution cannot be estimated reliably with a given set of training samples [Lippmann, 1987]. Learning and topological parameters of the ANNs in MR image classification has been studied by Ozkan [Özkan et al., 1993]. He showed no significant variation in the performance of the ANN classifier, as the number of hidden neurodes increased beyond four. Furthermore, he concluded that the nature of the problem of classifying MR data permitted the use of unitary learning rate, and the momentum was found to be important only when the learning rate was small. In that respect, the ANN topology used in this thesis was as follows: three neurodes at the input layer (corresponding to the 3 features; T_1-, T_2- and PD-weighted image volumes), one hidden layer containing 10 neurodes, and an output layer of 4 neurodes (corresponding to the 4 classes of interest; BCK, CSF, GM, WM). The maximum number of training iterations was set to 500, the learning rate was 0.8 and the momentum was 0.3. The delta error rate for convergence was set to 10^{-6}.

C4.5 Decision tree

For the C4.5 classifier, training involved the generation and pruning of 20 decision trees, and choosing the tree with the best error rate to classify the brain volume.

k-Nearest Neighbors

In the k-NN classifier, k is usually set to a small odd fraction of the number of training samples to avoid ties [Duda and Hart, 1973; Clarke et al., 1993]. The larger the value of k, the better estimation of of *a posteriori* probability of a given class, at a very high computational demand. The number of nearest neighbors for the kNN classifier was set to 5.

7.3.5 Unsupervised classifiers

Initial experiments showed that the performance of the unsupervised classification algorithms (HCM and FCM) was unsatisfactory when the only information presented to the algorithms (in addition to the data sets) was the number of tissue types to classify. In order to improve the classification performance, a method presented by Bensaid [Bensaid et al., 1996] was implemented. This was a semi-supervised method, where trainer supplied samples were used to guide the initial cluster centroids in the right direction. When a training set was used in this manner, dramatic improvements of classification performance were observed for both HCM and FCM algorithms. Therefore, subsequent testing of these unsupervised algorithms were done using such training sets.

7.3.6 Validation based on real MR data sets

Section 5.3 described some of the difficulties associated in doing quantitative appraisal of the classification performance in real data sets. These difficulties were especially evident when results from real MR volumes needed to be validated (in the absence of the gold

standard). Ideally, sufficient expert opinion is needed to establishing a mean (and variance) of performance ranking of all classifiers by several experts (Section 5.3.3). Given the labor-intensive nature of this task, the experts were reluctant to appraise the performance of classifiers on a large number real brain volumes, since each real MR volume had seven classified volumes corresponding to each algorithm. However, they volunteered to appraise classifier performance on a single real data set. Therefore, a real brain volume (from the ICBM data set) was classified by all seven algorithms using automatically generated training sets. This was to determine whether the best ranking classifier tested on the simulated data sets, performed equally well on real brain volumes. The classification results of the real MR volume were verified qualitatively by the three experts, each of whom spent considerable time analyzing the performance of the results.

7.3.7 Computational machinery

All the software tools necessary for the experiments were written in the C/C++ programming language, exceeding 40,000 lines of code. The code for the C4.5 classifier was obtained from Morgan Kaufman Publishing [Quinlan, 1993] and incorporated in the main classification tool, the "black box" (Figure 7.1). The ANN classifier was part of an interactive medical image analysis environment called MIDAS, written at Vanderbilt University [Zijdenbos et al., 1994; Zijdenbos and Dawant, 1994]. The ANN classifier was extracted from MIDAS and incorporated into the classification "black box". The rest of the algorithms (except BAYES) were developed by the author at the McConnell Brain Imaging Center.

All the experiments were conducted on a cluster of twelve Silicon Graphics Computer Systems (11 workstations and a multi-processor compute server) with varying CPU power (36 - 175 MHz). Table 7.1 shows average CPU time in minutes for different classifiers and training sets, calculated over all the workstations. Due to the long classification time required by the unsupervised classifiers, only training sets containing 50 samples were used.

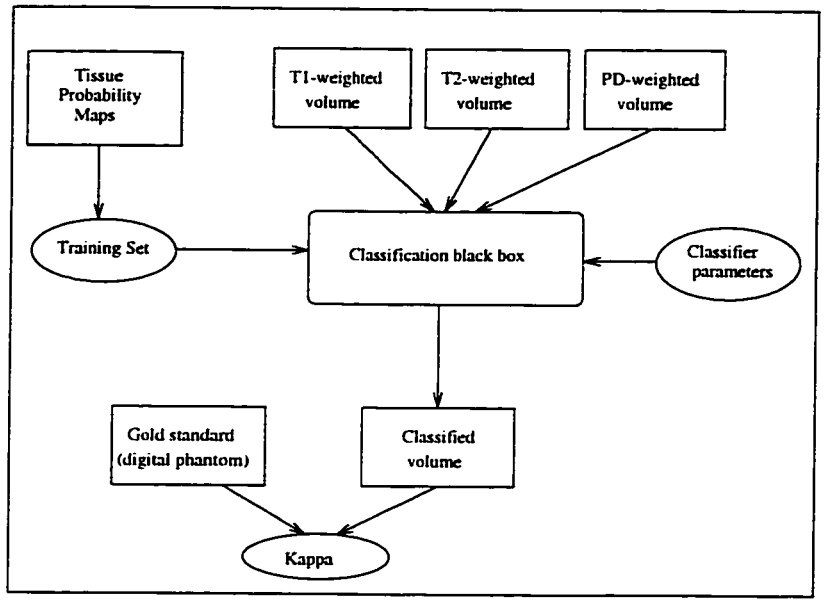


Figure 7.1: A flow diagram showing the classification process.

	ANN	*BAYES*	*C4.5*	*kNN*	*MD*	*HCM*	*FCM*
Samples(25)	4.16	6.38	3.56	140.10	4.16	*n/a*	*n/a*
Samples(50)	4.32	8.01	3.73	250.18	4.32	239.97	424.16

Table 7.1: Average CPU time in minutes for different classifiers and training sets.

7.4 Experiments and results

The following experiments were conducted to evaluate the usefulness of the brain tissue probability maps in automatically generating training sets, and to test performance of several training sets (manual and automated), under normal and varying conditions of MR imaging.

7.4.1 Usefulness of TPM in generating training sets

The training sets obtained from all the trainers (human and automated), were used to calculate the mean and the standard deviation of voxel intensities for each of the tissue

Trainer	*Samples*(25)		*Samples*(50)	
	Mean	*S.D.*	*Mean*	*S.D.*
Expert 1	233.74	33.30	232.56	35.52
Expert 2	220.00	22.55	223.98	30.10
Expert 3	204.98	60.27	210.78	52.23
AvgExpert	219.58	42.92	222.44	41.12
Auto 100	231.81	33.15	227.52	26.97
Auto 90	220.00	31.21	226.44	47.44
Auto 80	239.95	78.11	227.95	59.46
Auto 70	280.53	103.46	267.87	99.52
Auto 60	277.74	120.66	262.61	95.18

Table 7.2: Mean and standard deviation of voxel intensities of CSF tissue class for a T_1-weighted brain volume, calculated from training samples of several automatic and human trainers.

types CSF, GM and WM from the T_1-weighted normal simulated brain volume. Any of the T_1-, T_2- and PD-weighted image volumes could have been used, the reason for choosing T_1-weighted images was because they showed the most contrast between the three tissue types. The automated methods were named after their threshold values (*Auto 100*,..., *Auto 60*). Tables 7.2 to 7.4 list the mean and standard deviation of intensity values for CSF, GM and WM tissue classes calculated from the above mentioned training sets. AvgExpert refers to the average obtained from all human trainers.

The results presented in Tables 7.2 to 7.4, show that the mean intensities and the standard deviations of all tissue types were fairly constant for 25 and 50 sample training sets across all three experts, except for Expert 3 who had a slightly larger standard deviation for the CSF tissue class. On the other hand, in automatically generated training sets, the standard deviation of the tissue intensities increased dramatically as the prob-

Trainer	*Samples*(25)		*Samples*(50)	
	Mean	*S.D.*	*Mean*	*S.D.*
Expert 1	533.15	32.67	534.76	34.00
Expert 2	541.31	34.05	542.81	33.82
Expert 3	542.17	34.89	542.59	38.03
AvgExpert	538.87	33.67	540.05	35.30
Auto 100	513.19	33.68	519.41	35.11
Auto 90	505.89	75.53	508.36	59.93
Auto 80	497.95	117.01	506.97	94.33
Auto 70	513.19	75.99	511.90	75.11
Auto 60	520.49	114.05	534.44	112.33

Table 7.3: Mean and standard deviation of voxel intensities of GM tissue class for a T_1-weighted brain volume, calculated from training samples of several automatic and human trainers.

Trainer	*Samples*(25)		*Samples*(50)	
	Mean	*S.D.*	*Mean*	*S.D.*
Expert 1	720.52	49.43	721.43	43.38
Expert 2	727.61	34.77	726.86	35.30
Expert 3	717.95	22.87	718.38	29.82
AvgExpert	720.74	39.31	720.34	37.03
Auto 100	740.27	28.07	742.42	26.64
Auto 90	742.20	22.54	735.87	25.54
Auto 80	723.96	47.01	724.28	43.16
Auto 70	723.96	57.66	721.70	50.44
Auto 60	687.04	86.15	663.54	120.86

Table 7.4: Mean and standard deviation of voxel intensities of WM tissue class for a T_1-weighted brain volume, calculated from training samples of several automatic and human trainers.

Classifier	*Samples*	*Expert1*	*Expert2*	*Expert3*	*AvgExpert*	*Auto100*	*Auto90*
ANN	25	0.907	0.895	0.860	0.887	0.905	0.860
	50	0.901	0.907	0.866	0.891	0.903	0.905
BAYES	25	0.890	0.908	0.895	0.898	0.897	0.803
	50	0.900	0.901	0.889	0.897	0.885	0.854
C4.5	25	0.773	0.778	0.801	0.784	0.833	0.702
	50	0.868	0.866	0.801	0.845	0.872	0.773
kNN	25	0.845	0.828	0.850	0.841	0.848	0.857
	50	0.846	0.835	0.857	0.846	0.863	0.865
MD	25	0.802	0.799	0.822	0.807	0.806	0.832
	50	0.802	0.803	0.824	0.810	0.812	0.830

Table 7.5: Kappa values obtained for the supervised classifiers for five trainers at 25 and 50 samples under normal MR imaging conditions (noise level = 3%, RF inhomogeneity = 20%, slice thickness = 1mm).

ability threshold was reduced, even though the mean tissue intensities were somewhat uniform. Furthermore, the CSF tissue class intensities were over estimated in comparison to the experts as the threshold dropped below 80%. This was expected because of normal neuro-anatomical variability: as the tissue probability threshold was reduced, voxels from adjacent tissue classes were mixed in, thus changing the mean and increasing the variance of the class intensities. Since the standard deviations increased dramatically for probability thresholds lower than 90%, especially for CSF and GM tissue classes, *Auto 80*, *Auto 70* and *Auto 60* trainers were discarded from subsequent experimentation.

7.4.2 Sensitivity of the classifiers to different training sets

In order to determine how sensitive each of the five supervised classification algorithms were to manually and automatically selected training sets, the following experiment was

carried out.

Of the sixteen training sets (3 human and 5 automatic trainers having chosen two sets each) described in Section 7.3.2, ten were chosen from five trainers (3 humans, 2 automated). These sets were used to train and classify the simulated normal brain volume with the five supervised methods. Table 7.5 shows the tabulated kappa values of each of the five classifiers, under normal MR imaging conditions for simulated data sets.

For most classifiers, the *AvgExpert* and the *Auto 100* trainers did not demonstrate any noticeable difference in performance, as the number of training samples increased from 25 to 50 samples, except for the C4.5 classifier, where training sets with larger number of samples were preferred. For the *Auto 90* trainer, increasing the sample size from 25 to 50 voxels showed improvement in performance for the ANN, Bayesian and C4.5 classifiers. This was possibly due to the fact that as the probability threshold was reduced to 90%, the resulting training samples were more varied, and a larger number of samples were needed to establish proper decision boundaries. Some classifiers, like the MD, showed no noticeable improvement to an increase in the number of training samples, while others like C4.5 did. In general, increasing the training set size only gave marginally different overall results. Hence, in order to reduce the dimensionality of the search space (type of classifier, noise, RF inhomogeneity, slice thickness, different trainers, ...), only 50-sample sets were included in subsequent experiments. Also, Table 7.5 indicates that the automatic trainers yielded results no worse than those obtained with human trainers. Hence, only automated trainers were used for all subsequent experiments.

7.4.3 Classifier performance under varying conditions of MR imaging

To demonstrate classifier performance as a function of noise, RF inhomogeneity and slice thickness, the following experiments were conducted. The five training sets were used to train the seven algorithms and classify the brain image volumes in the simulated brain

database. The database had six varying levels of noise and RF inhomogeneity and three different slice thicknesses (Section 6.3). In subsequent graphs, kappa values of the classifiers, trained on sets supplied by human trainers, were averaged (labeled *AvgExpert*) in order to compare the automated trainers to average human trainer.

Varying levels of noise

Figure 7.2 demonstrates the performance (kappa values) of each of the supervised algorithms, and Figure 7.3 for the unsupervised algorithms as a function of increased levels of noise. As expected, increasing noise levels reduces classifier performance. However, among all the classifiers, the ANN classifier (and to some extent the BAYES classifier) was the least affected by noise. In contrast, the MD classifier was the most affected. In most classifiers (except for BAYES), the automated trainer *Auto 100* compared favorably to the average human trainer on all levels of noise. The *Auto 90* trainer did worse than the rest of the trainers for C4.5 classifier, as the performance fluctuated. It was not clear why the MD and kNN classifiers preferred the *Auto 90* trainer over *Auto 100.*

Varying levels of RF inhomogeneity

Figure 7.4 demonstrates the performance of each of the supervised algorithms, and Figure 7.5 for the unsupervised algorithms as a function of increased levels of RF inhomogeneity. The ANN and BAYES classifiers were the least affected by RF inhomogeneity, whereas the MD classifier was the most affected. The *Auto 100* trainer in general, did slightly better than *AvgExpert* on all levels of RF inhomogeneity for kNN, MD and C4.5 classifiers, and most levels for the ANN classifier. As before, it was also not clear why the MD and kNN classifiers preferred the *Auto 90* trainer over *Auto 100.*

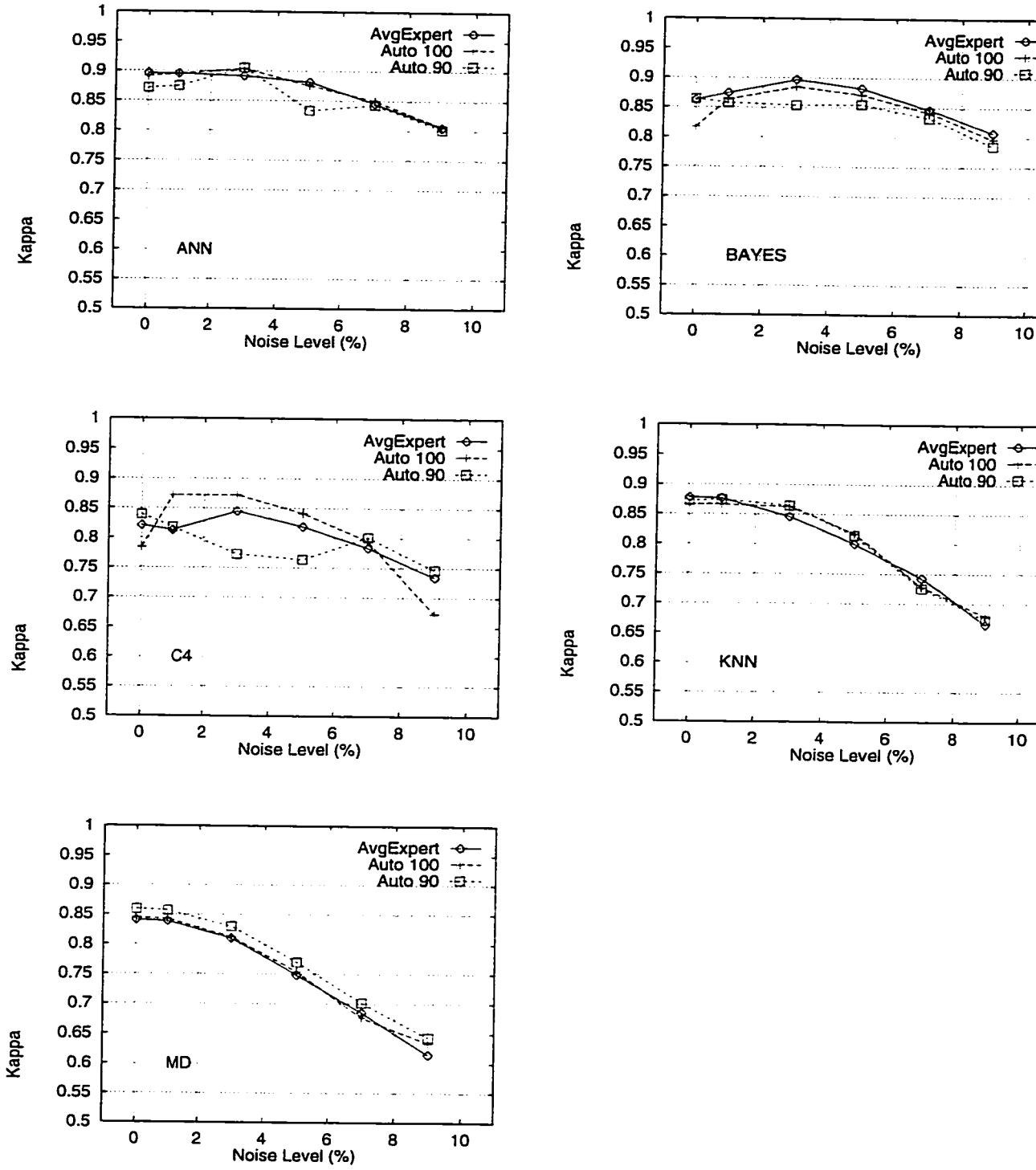


Figure 7.2: Performance of the supervised classifiers, using manual and automated training sets of 50 samples, on T_1-, T_2- and PD-weighted image volumes, under varying condition of noise (RF inhomogeneity = 20%, slice thickness = 1mm).

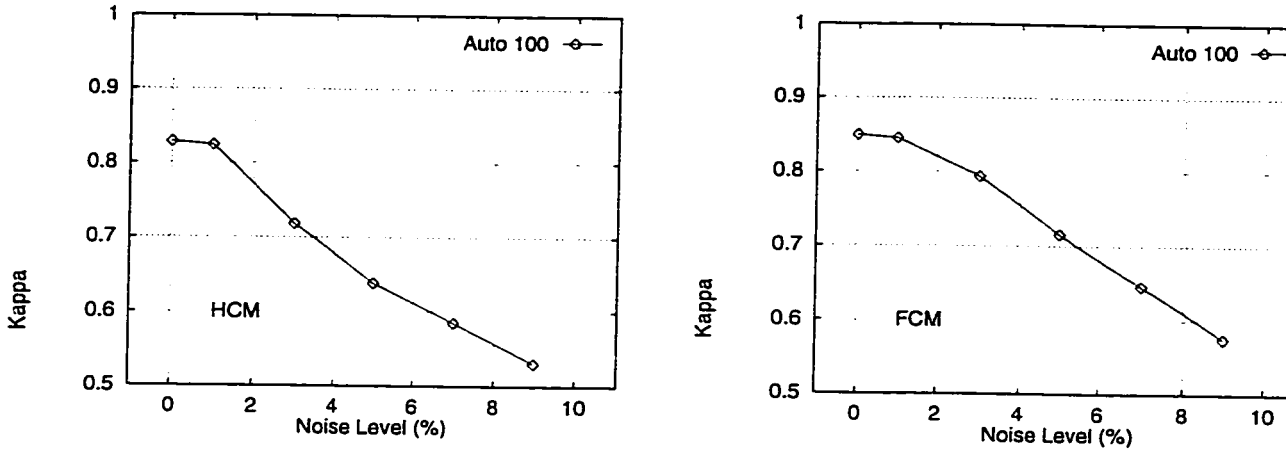


Figure 7.3: Performance of the unsupervised classifiers, using automated training sets of 50 samples, on T_1-, T_2- and PD-weighted image volumes, under varying condition of noise (RF inhomogeneity = 20%, slice thickness = 1mm).

Varying slice thicknesses

When classifying brain volumes having 3mm or 5mm slice thickness, experts choose training samples from respective volumes, since the only information available for the expert are the images themselves. So the experts were asked to choose training samples from brain volumes simulated at 3mm and 5mm slice thicknesses. In general, they expressed that as slice thickness became larger, it became more difficult to choose training samples with confidence. This was expected, since partial volume effect mixed intensity signals of neighboring tissues into single voxels, thus producing blurred voxel intensities. As for the automatic trainers, the same training sets generated earlier were used. Figure 7.6 demonstrates the performance of each of the supervised algorithms, and Figure 7.7 for the unsupervised algorithms as a function of increased slice thicknesses. Overall, the *Auto 100* and *Auto 90* trainers did no worse or better than *AvgExpert* in all classifiers except for the C4.5 classifier, where the *Auto 90* trainer , as before, showed performance fluctuations. For the kNN classifier, the *Auto 90* trainer performed marginally better than *AvgExpert*. The *Auto 100* trainer performed slightly better for the BAYES classifier.

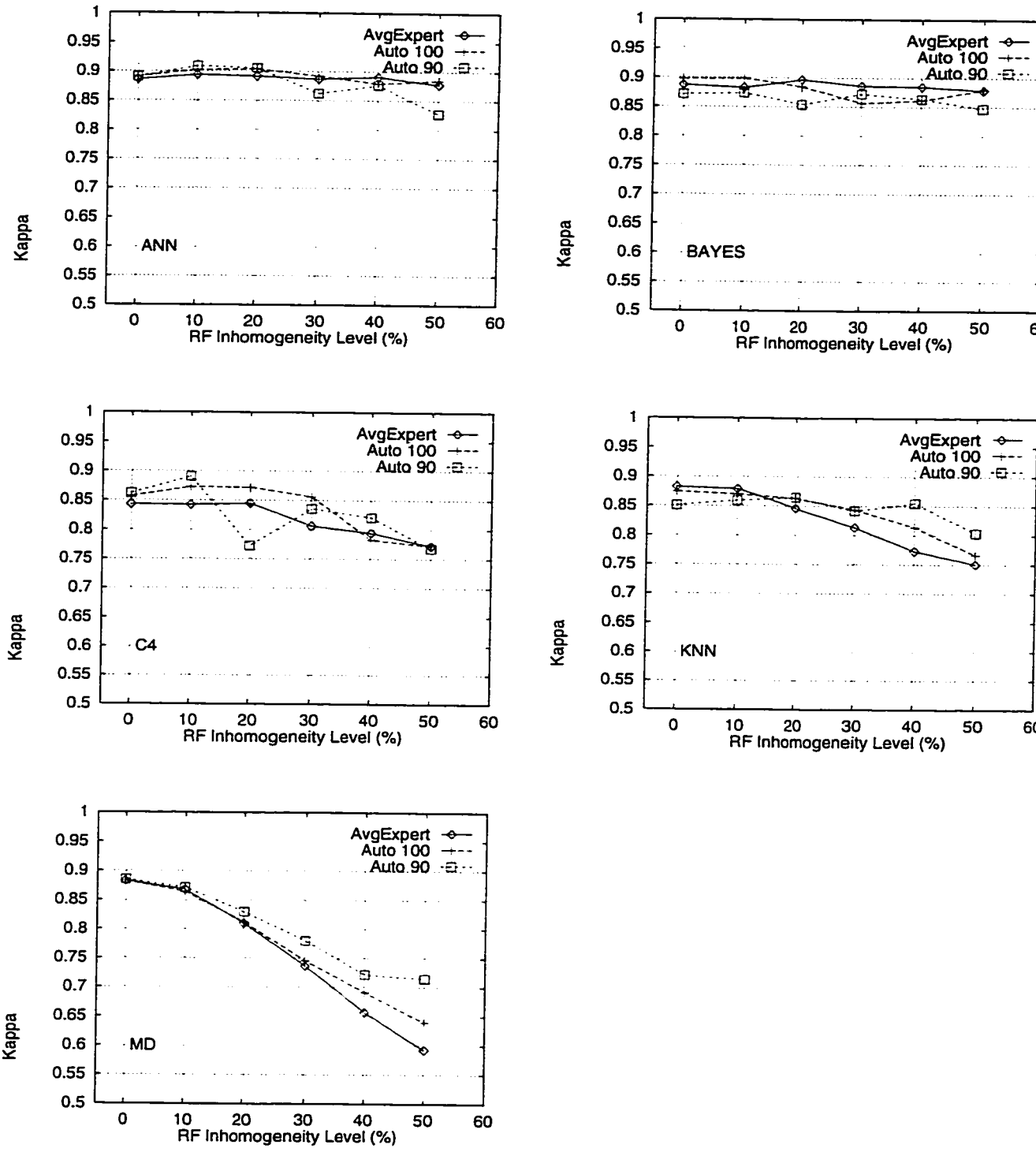
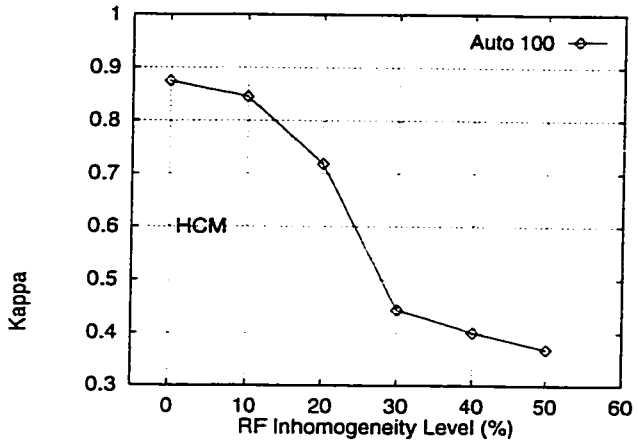


Figure 7.4: Performance of the supervised classifiers, using manual and automated training sets of 50 samples, on T_1-, T_2- and PD-weighted image volumes, under varying condition of RF inhomogeneity (noise level = 3%, slice thickness = 1mm).



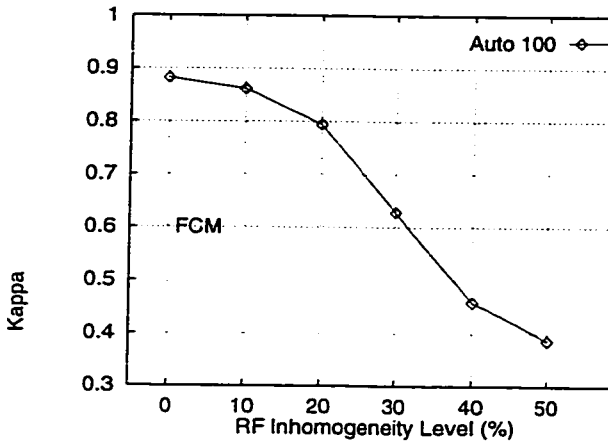
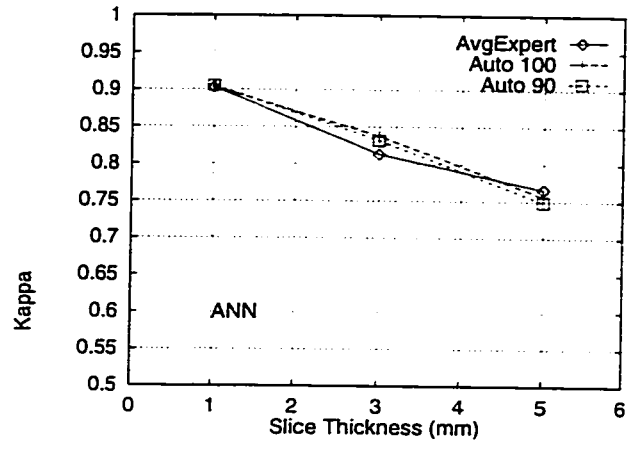
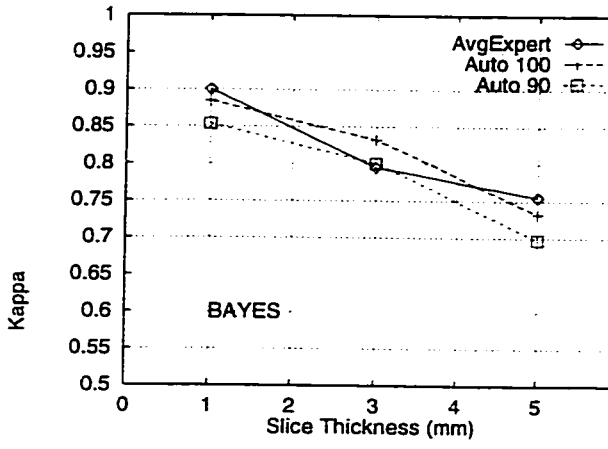


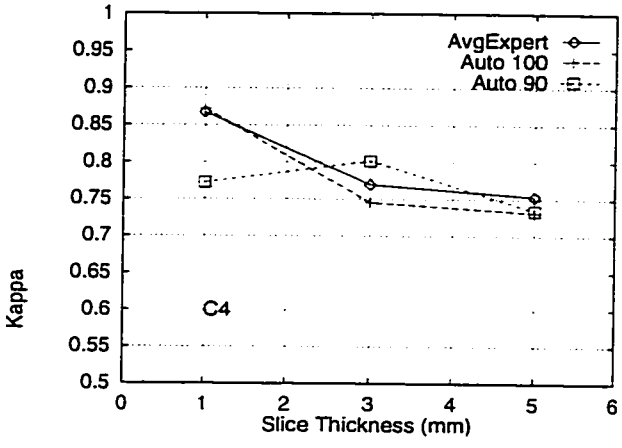
Figure 7.5: Performance of the unsupervised classifiers, using automated training sets of 50 samples, on T_1-, T_2- and PD-weighted image volumes, under varying condition of RF inhomogeneity (noise level = 3%, slice thickness = 1mm). Note the change in the origin of the ordinate.

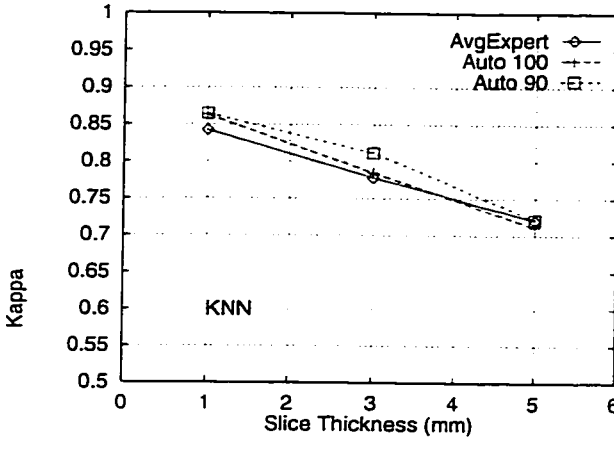
7.4.4 Result of real MR volumes

In order to evaluate the performance of the classifiers on a real MR data set, each expert was presented with the seven classified versions of the real data in a blind test, and asked to rate the volumes from best to worst. As mentioned in Section 7.3.6, it was difficult for the experts to do a quantitative assessment of the results obtained. Each had a different interpretation as to what parts of the brain were misclassified by various classifiers. They stated that some algorithms (BAYES and C4.5) overestimated gray matter tissue in the anterior part of the brain and underestimate it in the posterior part, whereas others (MD and C4.5) overestimated white matter tissue in the periphery. Even though there were disagreements among the experts, as to the relative ranking of all the classifiers, when asked which classified volume among the seven, best represented the underlying neuro-anatomy of the real MR volume, they all chose the ANN-classified volume.









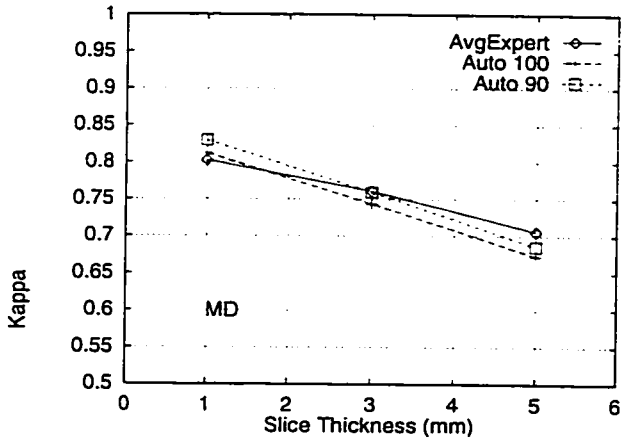
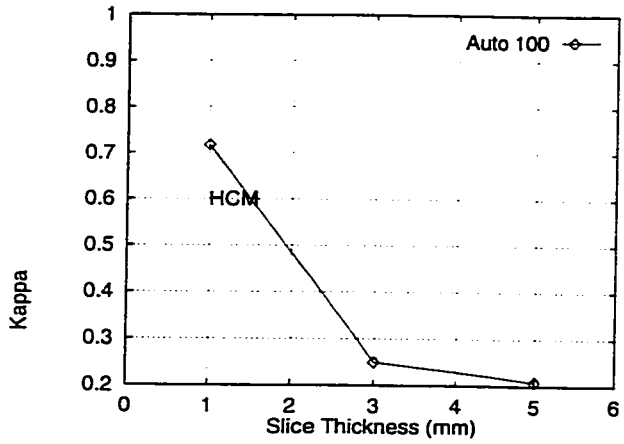


Figure 7.6: Performance of the supervised classifiers, using manual and automated training sets of 50 samples, on T_1-, T_2- and PD-weighted image volumes, under 1mm, 3mm, and 5mm slice thicknesses (noise level = 3%, RF inhomogeneity = 20%).



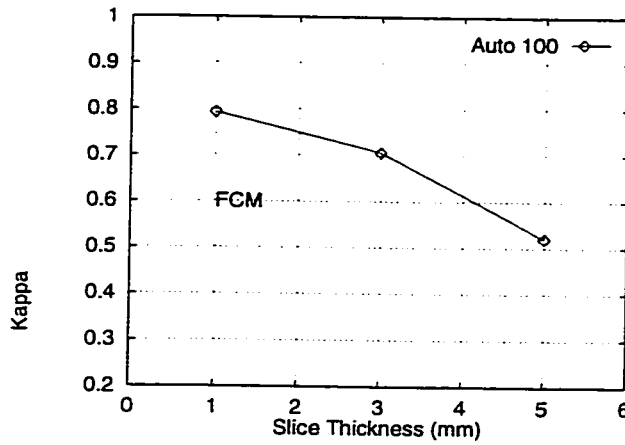


Figure 7.7: Performance of the unsupervised classifiers, using automated training sets of 50 samples, on T_1-, T_2- and PD-weighted image volumes, under 1mm, 3mm, and 5mm slice thicknesses (noise level = 3%, RF inhomogeneity = 20%). Note the change in the origin of the ordinate.

7.5 Discussion

Experiments were carried out to determine the performance of five supervised and two unsupervised classification algorithms, using manually and automatically generated training sets. Figures 7.9, 7.10 and 7.11 show the performance of all classifiers as a function of increased levels of noise, RF inhomogeneity and slice thickness respectively, using the *Auto 100* trainer. Figure 7.8 shows a single image slice from the simulated T1-weighted image volume, the digital phantom, and the results produced from each of the individual algorithms. Furthermore, the *Auto 100* trainer was also used to classify a real MR brain data set, using all seven classifiers. Analyzing the results obtained, the following conclusions were drawn:

1- The brain tissue probability maps were used successfully in automatically generating training sets, thus eliminating the need for manual interaction with the use of supervised classification algorithms. Overall, the automatic trainers (especially *Auto 100*) produced slightly more accurate results than their human counterparts.

2- The ANN classifier was the best overall performer, being the most resistant to noise and slice thickness and almost unaffected by RF inhomogeneity. Due to complex and parallel nature of the ANN classifier, one cannot simply gain insight in its reasoning by analyzing weight distribution on inter-connected neurodes. One "hopes" that the network develop expertise necessary to capture patterns that it had trained on [Schalkoff, 1992]. Two layer ANNs form unbounded convex polygonal decision boundaries in feature space [Lippmann, 1987]. This might be one possible explanation to the superior performance of the ANN classifier in discriminating non-linearities found in MR data. The BAYES classifier was the next best algorithm, where its relative immunity to RF inhomogeneity (which elongates clusters in feature space) may be due to the insensitivity of the Mahalanobis distance to exaggerated hyper-ellipsoid clusters.

3- Of the supervised algorithms, the MD classifier performed the worst, and was most influenced by the MR imaging artifacts. This was expected. As noise and RF inhomogeneity levels increased, clusters in feature space got more diffused and elongated (Section 3.3), and the Euclidean Distance measure of the MD classifier favored hyper-spherical clusters.

4- The C4.5 classifier showed fluctuating results to automatically generated training sets, especially at 90% probability threshold. It was not entirely clear why this was the case.

5- In the absence of noise (and at normal levels of RF inhomogeneity and slice thickness), the BAYES and C4.5 classifiers did worse than the rest. This was possibly due to their inability to establish some variance that would allow them to operate optimally.

6- Unsupervised methods showed dramatic improvements when initial clustering was guided through the use of training sets. No difference was observed between *Auto 100* and *Auto 90* trainers. The FCM algorithm performed better than Hard C Means algorithm, but both performed worse that the supervised classifiers.

7- The C4.5 classifier was the fastest, while the FCM classifier the slowest.

8- Training sets containing 50 samples performed a little better than 25 sample set for most classifiers.

9- The human experts stated that the result produced on real MR brain volume by the ANN classifier was the best among all classifiers.

7.6 Concluding remarks

A method for automating the training samples selection process was proposed, and its usefulness tested. It was determined that brain tissue probability maps can successfully be used to automate supervised classification algorithms. Furthermore, the performance of seven classifiers (5 supervised and 2 unsupervised) were tested under varying condition of MR imaging. The Artificial Neural Nets classifier was the best classifier under all conditions of MR imaging for both simulated and real data. The next chapter summarizes the work produced by this thesis, pointing to areas that could potentially be improved in the future.

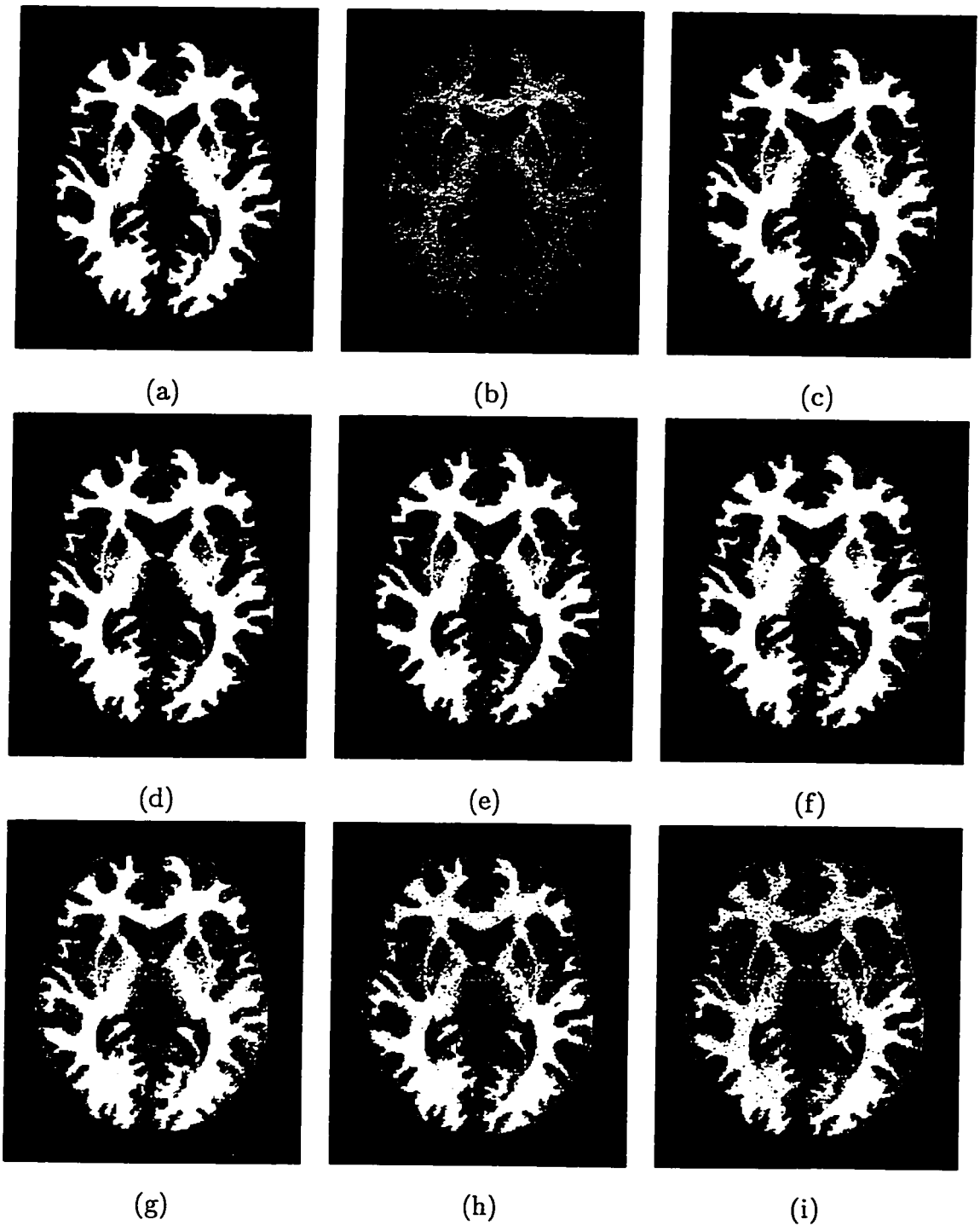


Figure 7.8: Sample images from (a) digital phantom (b) T_1-weighted simulated MR image; classified using (c) kNN (d) ANN (e) BAYES (f) C4.5 (g) MD (h) FCM (i) HCM classification algorithms on T_1-, T_2- and PD-weighted image volumes, under normal MR imaging conditions (noise level = 3%, RF inhomogeneity = 20%, slice thickness = 1mm).

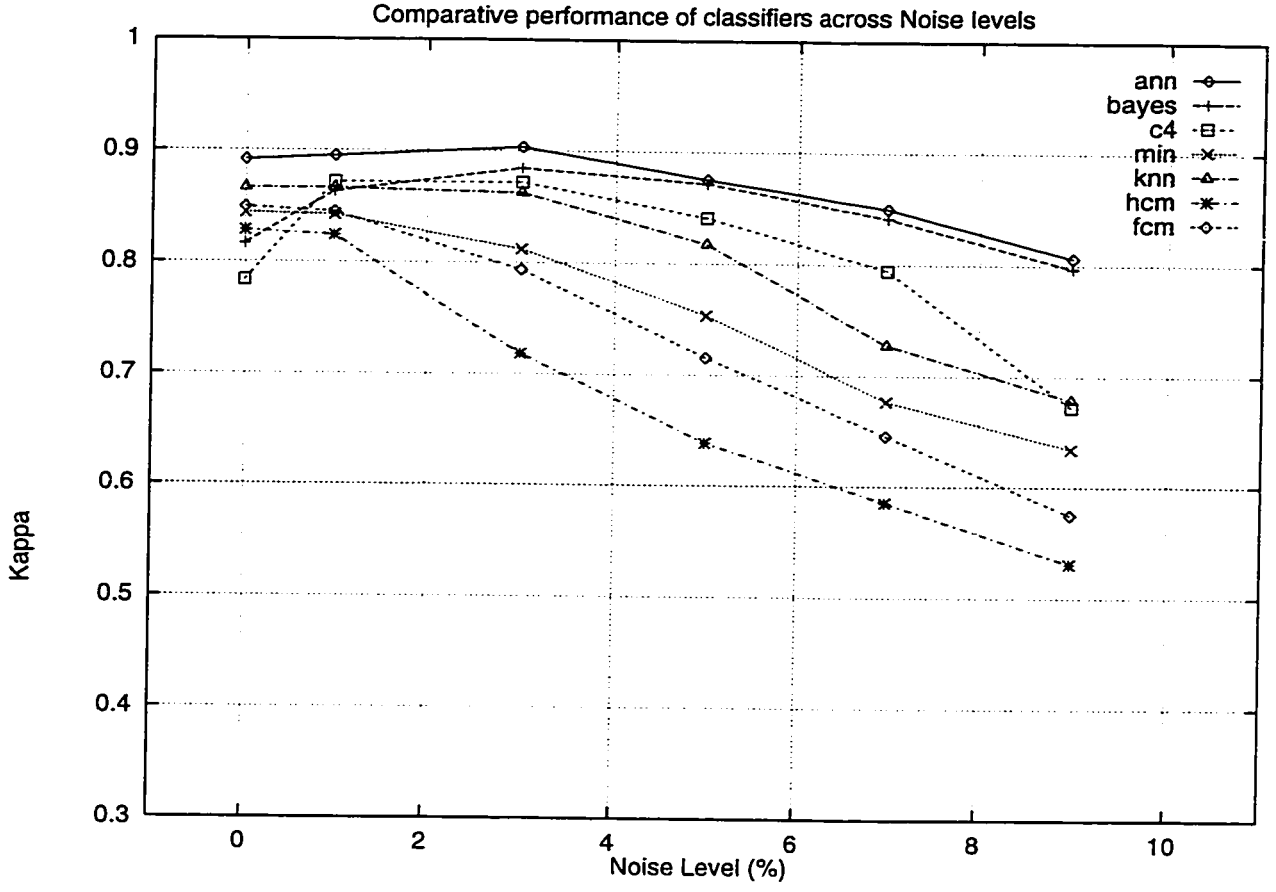


Figure 7.9: Performance of the all classifiers using an automatic trainer at 100% probability threshold, 50 training samples, on T_1-, T_2- and PD-weighted image volumes, under varying conditions of noise (RF inhomogeneity = 20%, slice thickness = 1mm).

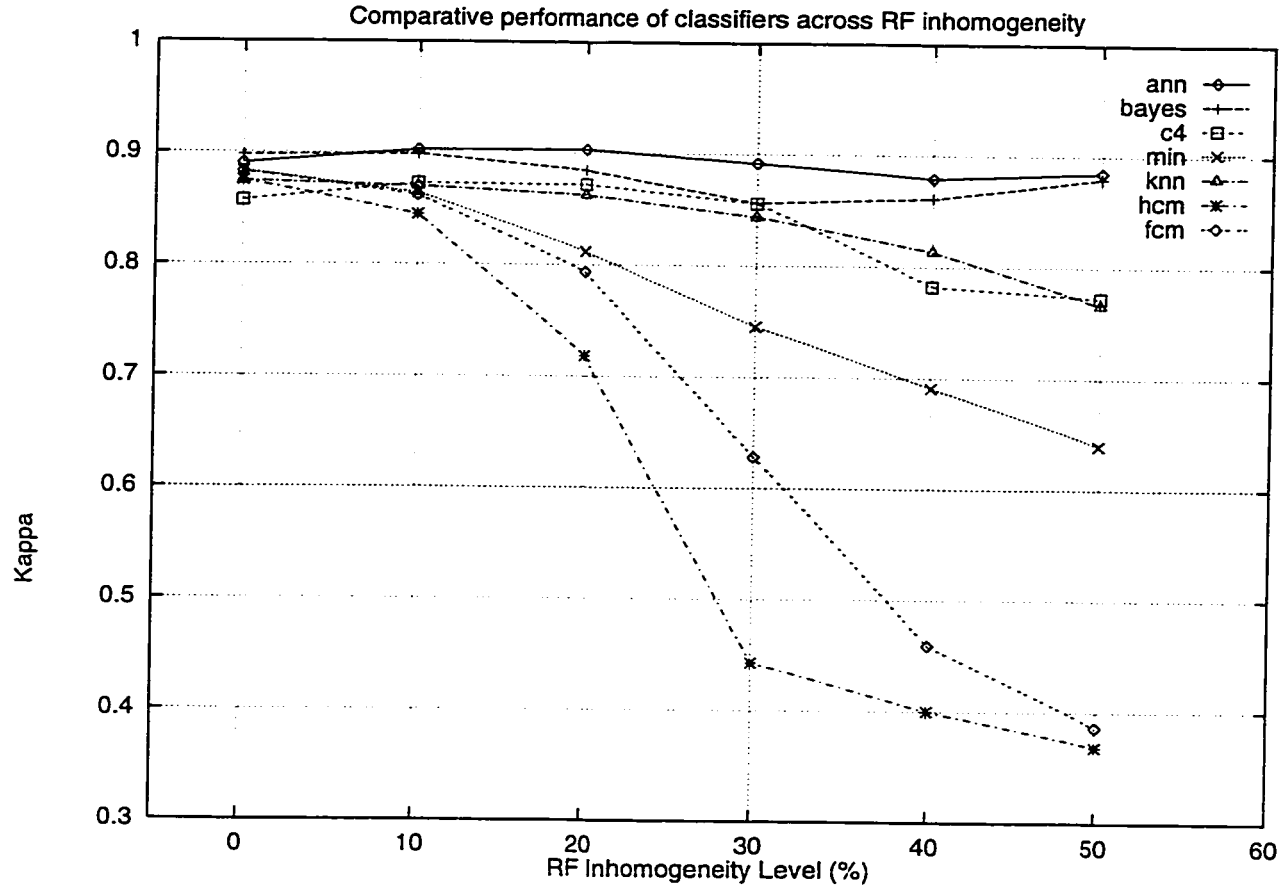


Figure 7.10: Performance of the all classifiers using an automatic trainer at 100% probability threshold, 50 training samples, on T_1-, T_2- and PD-weighted image volumes, under varying conditions of RF inhomogeneity (noise level = 3%, slice thickness = 1mm).

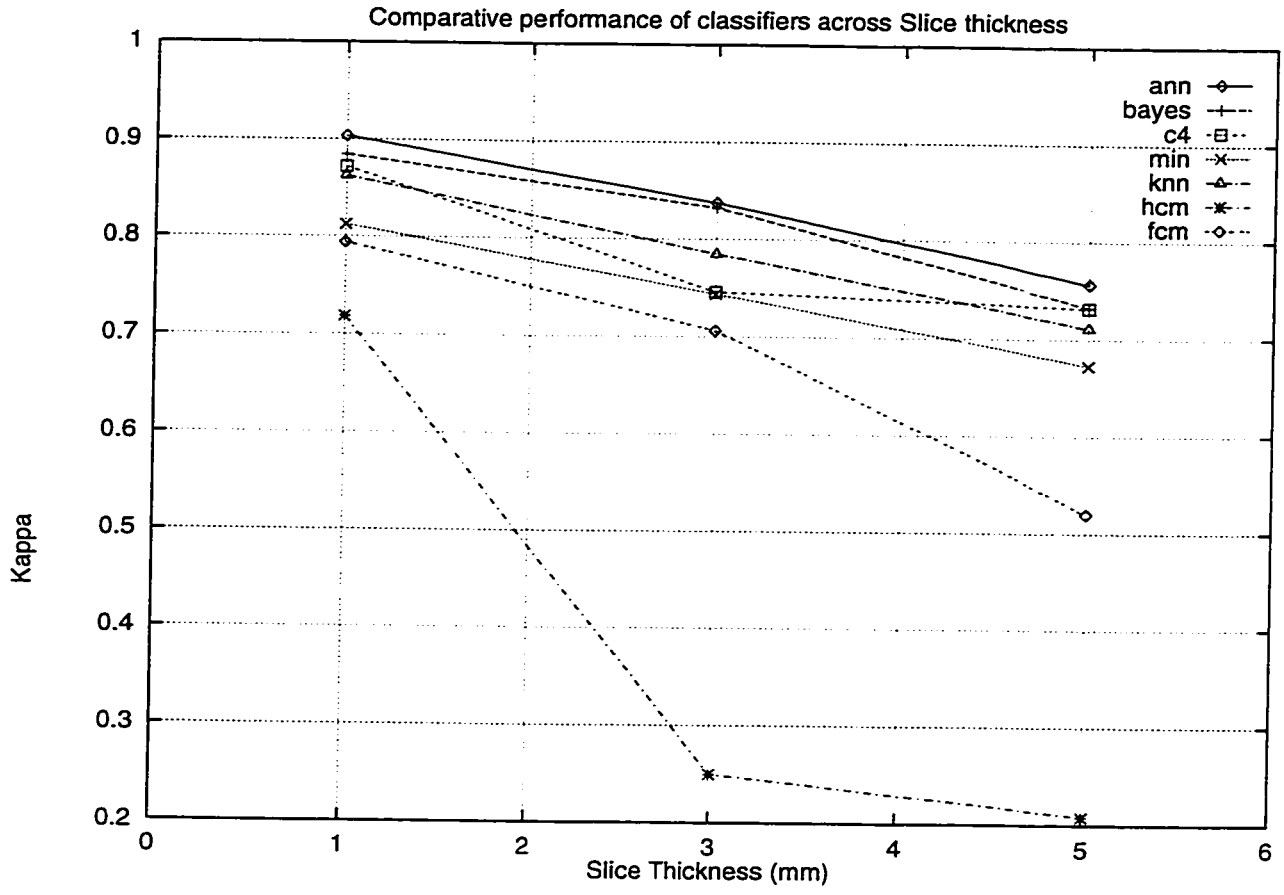


Figure 7.11: Performance of the all classifiers using an automatic trainer at 100% probability threshold, 50 training samples, on T_1-, T_2- and PD-weighted image volumes, under three slice thicknesses, 1mm, 3mm and 5mm (noise level = 3%, RF inhomogeneity = 20%).

Chapter 8

Conclusion and Future Work

8.1 Conclusions

The main objectives of this thesis, as listed in Section 1.4 were the following:

1– Develop a controlled test environment, where different classification algorithms can be implemented easily and their performance tested in a brain imaging context.

2– Provide a mechanism for automating supervised classification algorithms through the use of brain tissue probability maps. These maps provide *a priori* knowledge of neuro-anatomy. where training samples are extracted from probabilistic brain locations, eliminating the need for user intervention.

3– Test the performance of five supervised (Artificial Neural Networks, Bayesian. k - Nearest Neighbors, C4.5 decision tree, Minimum Distance) and two unsupervised (Hard C-Means, Fuzzy C-Means) classification algorithms under varying conditions of MR imaging.

The classification "black box" was designed to be versatile enough to allow algorithms to be included with ease and simplicity. The use of the digital phantom and the MR simu-

lator [Kwan et al., 1996] was instrumental in creating the *Simulated Brain Database* where the impact of MR imaging artifacts were studied in a carefully controlled environment.

The brain tissue probability maps were used successfully to generate training sets, thus automating the supervised classification process. Results obtained through automated trainers were comparable to those obtained through human trainers.

The performance of five supervised (Artificial Neural Networks, Bayesian, k-Nearest Neighbors, C4.5 decision tree, Minimum Distance) and two unsupervised (Hard C-Means, Fuzzy C-Means) classification algorithms was compared under varying conditions of MR imaging artifacts. The Artificial Neural Networks classifier gave the best overall results. It proved to be the most resistant to noise and partial volume effect, and almost not affected by RF inhomogeneity. Its performance was also the best in a real MR image volume. In contrast the worst performer was the Hard C Means unsupervised classifier.

8.2 Future work

One of the main objectives of this thesis has been the automation of the training set selection process. Although the experimental results obtained through automatic trainers were favorable, the results were based on only a few training sets. Further work should be done to establish statistical significance through the use of numerous training sets containing the same or varied number of training samples. Section 4.3.1 described a mechanism of obtaining different realizations of similar training sets, containing the same number of samples by having random sub-sampling intervals.

Because of inherent neuro-anatomical variability, and depending on the probability threshold chosen to select training samples from the tissue probability maps, the training set may include a few samples that do not fall within the intensity profile of the tissue class the classifier is training on. These samples might in turn confuse the classifier, as they are considered to be *false positives* training samples. Fuzzy algorithms can be used to associate certainties with classification results. This in turn, can be helpful in two

respects:

- Fuzzy classifications can be used to eliminate false positive training samples, making sure that the samples obtained with the aid of tissue probability maps produce proper tissue intensities.
- Fuzzy class certainties can be used to reveal distribution of different tissue types in a single voxel to characterize partial volume effect.

The elimination of false positive training samples could be accomplished by initially generating a training set with a large number of samples, then performing a fuzzy classification on this training set. The resulting fuzzy class certainties could be used to reject samples that did not produce a particular level of certainty at particular voxel positions. Afterward, the "cleaned" training set could be used to train and classify the original data set.

In the human brain, tissue types do not have the same density across the entire brain. For example, gray matter tissue is dense in some parts of the brain, while sparse in others, the relative intensity of the tissue depending on density. Researchers are sometimes interested in classifying only small sections of the brain. It may be possible to limit the search space of tissue probability maps in generating training samples to small regions of interest. This could improve classification results, since tissue intensities specific to those areas of interest could train the algorithm better. Moreover, RF inhomogeneity, being a smoothly varying field, has less of an impact on localized tissue.

In addition to multi-spectral MRI, Positron Emission Tomography (PET) images could be incorporated to increase the dimensionality of the feature space of brain images being classified. PET simulators [Ma et al., 1993] have been successfully used to study the effects of partial volume on quantification of brain metabolism [Rousset et al., 1993]. Simulated PET images could be used in conjunction with simulated MR images to determine whether PET improves the discriminating power of classification algorithms.

Recently, genetic algorithms [Goldberg, 1989] have been developed that follow evolutionary trends in producing the fittest discriminating function that classify data best. The use of these algorithms may improve on results obtained in this thesis.

8.3 Concluding remarks

Artificial intelligence [Rich and Knight, 1991; Shinghal, 1992] and machine learning [Michalski, 1983] are rapidly developing areas of computer science and technology, where principles of signal and image processing [Schowengerdt, 1983; Schalkoff, 1989], pattern recognition [Duda and Hart, 1973; James, 1985; Schalkoff, 1992] and computer graphics [Foley et al., 1990], are being used as components of sophisticated medical systems ranging from radiological analysis of diagnostic images to image-guided neuro-surgery. As research advances are made in artificial intelligence, more sophisticated algorithms will become available, continuously raising the issue of optimal functionality. This thesis has presented a method to automate the brain tissue classification process, which in many respects is a precursor to numerous other research and clinical studies in neurology. It has also developed an environment where future classification algorithms in artificial intelligence and machine learning can be readily tested in a brain imaging context.

Abbreviations and Glossary

Acquisition: The process of measuring a signal and storing it into an image file.

Anterior Commisure: A structure deep in the brain defined as the origin in the stereotaxic atlas.

ANN: Artificial Neural Netorks.

BP: Back Propagation.

Cerebro-Spinal Fluid (CSF): A fluid that fills the ventricles.

Classification: The process of assigned meaningful labels to different brain tissue types.

Computerized Tomography (CT): An imaging modality utilizing X-rays (ionizing radiation) to acquire multi-slice anatomical images of the body.

Digital Phantom: A labeled brain volume where each voxel is assigned to any of the normal tissue types CSF, GM, or WM, used for validation.

Echo Time (TE): The time at which MR signals are measured, after the delivery of an RF pulse.

FCM: Fuzzy C-Means.

Gray Matter (GM): A type of tissue found in the brain this is predominantly made of neuronal dendrites.

HCM: Hard C-Means.

ICBM: International Consortium for Brain Mapping.

kNN: k Nearest-Neighbor.

Longitudinal Relaxation: Tissue specific time constant (T1) characterized by the rate at which the longitudinal magnetization vector returns to equilibrium after an RF pulse.

Magnetic Resonance Imaging (MRI): Measuring the electro-magnetic energy emitted by nuclei in response to Radio Frequency pulses, reconstructed into an image.

MDP: Master Digital Phantom.

MD: Minimum Distance.

MNI: Montréal Neurological Institute.

Multi-spectral: The condition denoting the nature of MR data, where underlying anatomy of the imaged organ is represented by multi-contrast images of varying characteristics.

Neuro-anatomical variability: The condition represented by individuals having different sized and shaped brains.

Noise: Variation of the signal measuring capability of the imaging hardware due to Brownian motion of electrons in the circuitry.

Partial Volume: Whenever signals from more than one tissue type a mixed into a single voxel.

Physical Phantom: A compartmentalized construction, filled with paramagnetic substances to resemble tissue characteristics of a real brain, used for validation.

Positron Emission Tomography (PET): An imaging modality that measures the extent of biochemical reactions taking place in the body, such as measuring glucose or oxygen consumption, or blood flow.

***PD*-weighted images:** The image acquired mostly due to Proton Density.

RF Inhomogeneity: A variation in intensity of similar tissue types, due to magnetic pulse variations.

Radio Frequency (RF) Pulse: Electro-magnetic radiation imparted to a collection of nuclei in magnetic equilibrium.

Repetition Time (TR): The rate at which a sequence of RF pulses is repeated.

Scatter Plots: Multi-dimensional histogram.

Segmentation: The process of delineation of MR images into distinct anatomical regions.

SNR: Signal-to-Noise Ratio.

Similarity Measure: A index indicating the extent of agreement between a classified volume and the gold standard.

Simulated brain database (SBD): A collection of simulated MR brain volumes with varying conditions of MR artifacts, used for validation.

Spins: The momentum of atomic nuclei with an odd number of protons and neutrons.

Stereotaxic Space: A standard anatomically-based frame of reference, where different sized and shaped brains can be transformed into and meaningfully compared on a voxel-by-voxel basis.

Supervised Classification: A type of classifier, where the algorithm is trained on specimen of known classes, to later classify specimen of unknown classes.

T_1-weighted images: The image acquired mostly due to longitudinal relaxation time, T1.

T_2-weighted images: The image acquired mostly due to transverse relaxation time, T2.

TPM: Tissue Probability Maps.

Training set: A set of voxels representing characteristic tissue intensities on which supervised classifiers train.

Transverse Relaxation: Tissue specific time constant (T2), characterized by the rate at which the transverse magnetization vector returns to equilibrium after an RF pulse.

Unsupervised Classification: A type of classifier, where the algorithm requires no training.

Validation: The process of verifying if a particular classification algorithm performed acceptably.

Ventricles: Cavities inside deep brain structures.

Voxel: A 3D pixel.

White Matter (WM): A type of tissue found in the brain that is predominantly made of neuronal axons.

References

Allen, P. S. (1992). Some fundamental principles of nuclear magnetic resonance. In Bronskill, M. J. and Sprawls, P., editors, *The Physics of MRI 1992 AAPM Summer school Procceedings*, pages 15–31. American Institite of Physics.

Auer, P., Holte, R., and Maass, W. (1995). Theory and application of agnostic pac-learning with small decison trees. In *Proceedings of Twelveth International Conference on Machine Learning*, pages 21–29, San Francisco, CA.

Bartko, J. J. (1991). Measurement and reliability: statistical thinking considerations. *Schizophrenia Bulletin*, 17(3):483–489.

Bartko, J. J. and Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervervous Mental Disorders.*, 163(5):307–317.

Bensaid, A. M., Hall, L. O., Bezdek, J. C., and Clarke, L. P. (1996). Partially supervised clustering for image segmentation. *Pattern Recognition*, 12(5):859–871.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum, New York.

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

Bishop, Y. M., Fienber, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge MA.

Carpenter, M. B. (1985). *Core text of neuroanatomy.* Williams and Walkins.

Caudill, M. and Butler, C. (1992). *Understanding Neural Networks: Computer Explorations.* MIT Press, Cambridge, MA.

Cendes, F., Leproux, F., Melonson, D., Ethier, R., Evans, A., Peters, T., and Andermann, F. (1993). MRI of amygdala and hippocampus in termporal lobe epilepsy. *Journal of Computer Assisted Tomography*, 17:206–210.

Clarke, L., Velthuizen, R., Camacho, M., Heine, J., Vaidyanathan, M., Hall, L., Thatcher, R., and Silbiger, M. (1995). MRI segmentation: Methods and applications. *Magnetic Resonance Imaging*, 13(3):343–368.

Clarke, L. P., Velthuizen, R. P., Hall, L. O., Bezdek, J. C., Bensaid, A. M., and Silbiger, M. L. (1992). Comparison of supervised pattern recognition techniques and unsupervised methods for MRI segmentation. *Proceedings of the SPIE. Medical Imaging VI: Image Processing*, 1652:668–677.

Clarke, L. P., Velthuizen, R. P., Phuphanich, S., et al. (1993). MRI: stability of three supervised segmentation techniques. *Magnetic Resonance Imaging*, 11(1):95–106.

Cline, H. E., Dumoulin, C. L., Hart Jr., H. R., Lorensen, W. E., and Ludke, S. (1987). 3D reconstruction of the brain from magnetic resonance images using a connectivity algorithm. *Magnetic Resonance Imaging*, 5(5):345–352.

Cline, H. E., Lorensen, W. E., Kikinis, R., and Jolesz, F. (1990). Three-dimensional segmentation of MR images of the head using probability and connectivity. *Journal of Computer Assisted Tomography*, 14(6):1037–1045.

Cline, H. E., Lorensen, W. E., Souza, S. P., et al. (1991). 3D surface rendered MR images of the brain and its vasculature. *Journal of Computer Assisted Tomography*, 15(2):344–351.

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

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

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

Dawant, B. M., Zijdenbos, A. P., and Margolin, R. A. (1993). Correction of intensity variations in MR images for computer-aided tissue classification. *IEEE Transactions on Medical Imaging*, 12(4):770–781.

Dougherty, R., J. K., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of Twelveth International Conference on Machine Learning*, pages 194–202, San Francisco, CA.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York, NY.

Ehricke, H.-H. (1990). Problems and approaches for tissue segmentation in 3D-MR imaging. In *Proceedings of the SPIE: Medical Imaging IV: Image Processing*, volume 1233, pages 128–137.

Evans, A., Kamber, M., Collins, D., and MacDonald, D. (1994). An MRI-based probabilistic atlas of neuroanatomy. In Shorvon, S. et al., editors, *Magnetic Resonance Scanning and Epilepsy*, chapter 48, pages 263–274. Plenum Press.

Evans, A. C., Collins, D. L., and Milner, B. (1992a). An MRI-based stereotactic atlas from 300 young normal subjects. *Proceedings of the 22nd Annual Symposium, Society for Neuroscience.*, 18:408.

Evans, A. C., Marrett, S., Neelin, P., Collins, D. L., Worsley, K., Dai, W., Milot, S., Meyer, E., and Bub, D. (1992b). Anatomical mapping of functional activation in stereotactic coordinate space. *NeuroImage*, 1(1):43–53.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31:651–659.

Foley, J. D., van Dam, A., Feiner, S. K., and Hughes, J. F. (1990). *Computer Graphicss: Principles and Practice.* Addison-Wesley Publishing Company, Reading MA.

Fox, P., Mikiten, S., Davis, G., and Lancaster, J. (1994). BrainMap: A database of functional brain mapping. In Thatcher, R., Hallett, M., Zeffiro, T., John, E., and Heurta, M., editors, *Functional Neuroimaging, technical foundations*, pages 95–109. Academic Press, San Diego, CA.

Fox, P. T., Perlmutter, J. S., and Raichle, M. E. (1985). A stereotactic method of anatomical localization for positron emission tomography. *Journal of Computer Assisted Tomography*, 9(1):141–153.

Gerig, G., Martin, J., Kikinis, R., Kübler, O., Shenton, M., and Jolesz, F. A. (1992). Unsupervised tissue type segmentation of 3D dual-echo MR head data. *Image and Vision Computing*, 10(6):349–360.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Publishing Company, Reading MA.

Gutfinger, D., Hertzberg, E., Tolxdorff, T., et al. (1991). Tissue identification in MR images by adaptive cluster analysis. In *Proceedings of SPIE. Medical Imaging V: Image Processing*, volume 1445, pages 288–296.

Hall, L. O., Bensaid, A. M., Clarke, L. P., Velthuizen, R. P., Silbiger, M. S., and Bezdek, J. C. (1992). A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. *IEEE Transactions on Neural Networks*, 3(5):672–682.

Hartigan, J. A. (1975). *Clustering algorithms.* John Wiley and Sons, New York, NY.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, 84(2):289–297.

Jack, Jr., C. R., Gehring, D. G., Sharbrough, F. W., Felmlee, J. P., Forbes, G., Hench, V. S., and Zinsmeister, A. R. (1988). Temporal lobe volume measurement from MR images: Accuracy and left-right asymmetry in normal persons. *Journal of Computer Assisted Tomography*, 12:21–29.

Jack, Jr., C. R., Sharbrough, F. W., Twomey, C. K., Cascino, G. D., Hirschorn, K. A., Marsh, W. R., Zinsmeister, A. R., and Scheithauer, B. (1990). Temporal lobe seizures: Lateralization with MR volume measurements of the hippocampal formation. *Radiology*, 175:423–429.

Jackson, E. F., Narayana, P. A., Wolinsky, J. S., and Doyle, T. J. (1993). Accuracy and reproducibility in volumetric analysis of multiple sclerosis lesions. *Journal of Computer Assisted Tomography*, 17(2):200–205.

James, M. (1985). *Classification Algorithms.* William Collins Sons, London, England.

Just, M. and Thelen, M. (1988). Tissue characterization with T1, T2, and proton density values: results in 160 patients with brain tumors. *Radiology*, 169(3):779–785.

Kabani, N. J., MacDonald, D., Evans, A., and Gopnik, M. (1996). Neuroanatomical correlates of familial language imparement: A preliminary report. *Journal of Neurolinguistics.* Submitted.

Kamber, M., Collins, D. L., Shinghal, R., Francis, G. S., and Evans, A. C. (1992). Model-based 3D segmentation of multiple sclerosis lesions in dual-echo MRI data. In *Proceedings of the SPIE. Visualization in Biomedical Computing*, volume 1808, pages 590–600, Chapel Hill, North Carolina.

Kamber, M., Shinghal, R., Collins, D. L., Francis, G. S., and Evans, A. C. (1995). Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. *IEEE Transactions in Medical Imaging*, 14(3):442–453.

Kulkarni, A. D. (1994). *Artificial Neural Networks for Image Understanding.* Van Nostrand Reinhold, New York, NY.

Kwan, R. K.-S., Evans, A. C., and Pike, G. B. (1996). An extensible MRI simulator for post-processing evaluation. In Höhne, K. H. and Kikinis, R., editors, *Visualization in Biomedical Computing. 4th International Conference, VBC '96. Hamburg, Germany, September 1996. Proceedings.*, volume 1131 of *Lecture Notes in Computer Science*, pages 135–140, Berlin. Springer-Verlag.

Lee, J., Reutens, D., Dubeau, F., Evans, A., and Andermann, F. (1995). Morphometry in termporal lobe epilepsy. *Magnetic Resonance Imaging*, 13(8):1073–1080.

Lim, K. O. and Pfefferbaum, A. (1989). Segmentation of MR brain images into cerebrospinal fluid spaces, white and gray matter. *Journal of Computer Assisted Tomography*, 13(4):588–593.

Lippmann, R. P. (1987). Introduction to computing with neural nets. *IEEE ASSP*, pages 4–22.

Ma, Y., Kamber, M., and Evans, A. (1993). 3D simulation of PET brain images using segmented MRI data and positron tomography characteristics. *Computerized medical imaging and graphics*, 17:365–371.

Mazziotta, J. C., Toga, A. W., Evans, A. C., Fox, P. T., and Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development. *Neuroimage*, 2:89–101.

McVeigh, E. and Atalar, E. (1992). Balancing contrast, resolution, and signal-to-noise ration in magnetic resonance imaging. In Bronskill, M. J. and Sprawls, P., editors, *The Physics of MRI 1992 AAPM Summer school Procceedings*, pages 234–267. American Institite of Physics.

Michalski, R. (1983). *Machine learning: An artifical intelligence approach (Vol. 1)*. Morgan Kaufmann, San Mateo, CA.

Mitchell, J., Karlik, S. J., Lee, D. H., and Fenster, A. (1994). Classification and analysis of multiple sclerosis lesions in spin-echo MR exams. *SPIE proceedings*, 2359:362–372.

Neelin, P., Crossman, J., Hawkes, D., Ma, Y., and Evans, A. (1993). Validation of an MRI/PET landmark registration method using 3D simulated PET images and point simulation. *Computerized Medical Imaging and Graphics*, 17:351–356.

Nishimura, D. G. (1993). *Introduction to Magnetic Resonance Imaging.* Verify, Standford University, CA.

Özkan, M., Dawant, B. M., and Maciunas, R. J. (1993). Neural-network-based segmentation of multi-modal medical images: A comparative and prospective study. *IEEE Transactions on Medical Imaging*, 12(3):534–544.

Peterson, J., Christofferson, J., and Golman, K. (1993). MR simulation using k-space formalism. *Magnetic Resonance Imaging*, 11:557–568.

Philips, M. S. (1984). *Introduction to MR Imaging.* Philips Medical Systems, The Netherlands.

Plewes, D. and Bishop, J. (1992). Spin-echo MR imaging. In Bronskill, M. J. and Sprawls, P., editors, *The Physics of MRI 1992 AAPM Summer school Procceedings*, pages 167–187. American Institite of Physics.

Quinlan, J. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufman, New York.

Rich, E. and Knight, K. (1991). *Artificial Intelligence.* McGraw-Hill, New York, NY.

Rosenfield, G. H. and Fitzpatrick-Lins, K. (1986). A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 52(2):223–227.

Rousset, O. G., Ma, Y., Léger, G., Gjedde, A. H., and Evans, A. C. (1993). Quantification of brain functions. In Uemura, K. et al., editors, *Tracer Kinetics and Image Analysis in Brain PET*, pages 113–123. Elsevier Science Publishers B.V.

Schalkoff, R. (1989). *Digital image processing and computer vision.* John Wiley and Sons, New York, NY.

Schalkoff, R. (1992). *Pattern Recognition - Statistical, Structural and Neural Approaches.* John Wiley and Sons, New York, NY.

Schowengerdt, R. A. (1983). *Techniques for Image Processiong and Classification in Remote Sensing.* Acedemic Press, New York, NY.

Shinghal, R. (1992). *Formal Concepts in Artificial Intelligence.* Chapman and Hall, London, UK, Co-published in the US with Van Nostrand, New York.

Simmons, A., Tofts, P. S., Barker, G. J., and Arridge, S. R. (1994). Sources of intensity nonuniformity in spin echo images. *Magnetic Resonance in Medicine*, 32:121–128.

Sprawls, P. (1992). The magnetic resonance image: A physical perspective. In Bronskill, M. J. and Sprawls, P., editors, *The Physics of MRI 1992 AAPM Summer School Proceedings*, pages 1–14. American Institite of Physics.

Talairach, J., Szikla, G., and Tournoux, P. (1967). *Atlas d'anatomie stereotaxique du telencephale.* Masson, Paris.

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

Taxt, T., Lundervold, A., Fuglaas, B., Lien, H., and Abeler, V. (1992). Multispectral analysis of uterine corpus tumors in magnetic resonance imaging. *Magnetic Resonance in Medicine*, 23:55–76.

Vaidyanathan, M., Clarke, L., Velthuizen, R., Phuphanich, S., Bensaid, A., Hall, L., Bezdek, J., Greenberg, H., Trotti, A., and Silbiger, M. (1995). Comparison of supervized MRI segmentation methods for tumor volume determination during therapy. *Magnetic Resonance Imaging*, 13(5):719–728.

Vannier, M. W., Butterfield, R. L., Jordan, D., et al. (1985). Multispectral analysis of magnetic resonance images. *Radiology*, 154(1):221–224.

Vannier, M. W., Butterfield, R. L., Rickman, D. L., et al. (1987). Multispectral magnetic resonance image analysis. *CRC Critical Reviews in Biomedical Engineering*, 15(2):117–144.

Vannier, M. W., Pilgram, T. K., Speidel, C. M., Neumann, L. R., Rickman, D. L., and Schertz, L. D. (1991). Validation of magnetic resonance imaging (MRI) multispectral tissue classification. *Computerized Medical Imaging and Graphics*, 15(4):217–223.

Vannier, M. W., Speidel, C. M., and Rickman, D. L. (1988). Magnetic resonance imaging multispectral tissue classification. *News in Physiological Sciences (NIPS)*, 3:148–154.

Williams, L. E. (1987). *The Diagnostic Process*, volume 3. CRC Press, Boca Raton FL.

Yule, G. (1912). On the methods of measuring association between two attributes. *Journal of Royal Statistical Society*, 75:581–642.

Zijdenbos, A. P. and Dawant, B. M. (1994). Brain segmentation and white matter lesion detection in MR images. *Critical Reviews in Biomedical Engineering*, 22(5&6):401–465.

Zijdenbos, A. P., Dawant, B. M., and Margolin, R. A. (1993). Measurement reliability and reproducibility in manual and semi-automatic MRI segmentation. In *Proceedings of the IEEE-Engineering in Medicine and Biology Society (EMBS)*, pages 162–163, San Diego, CA.

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., and Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724.

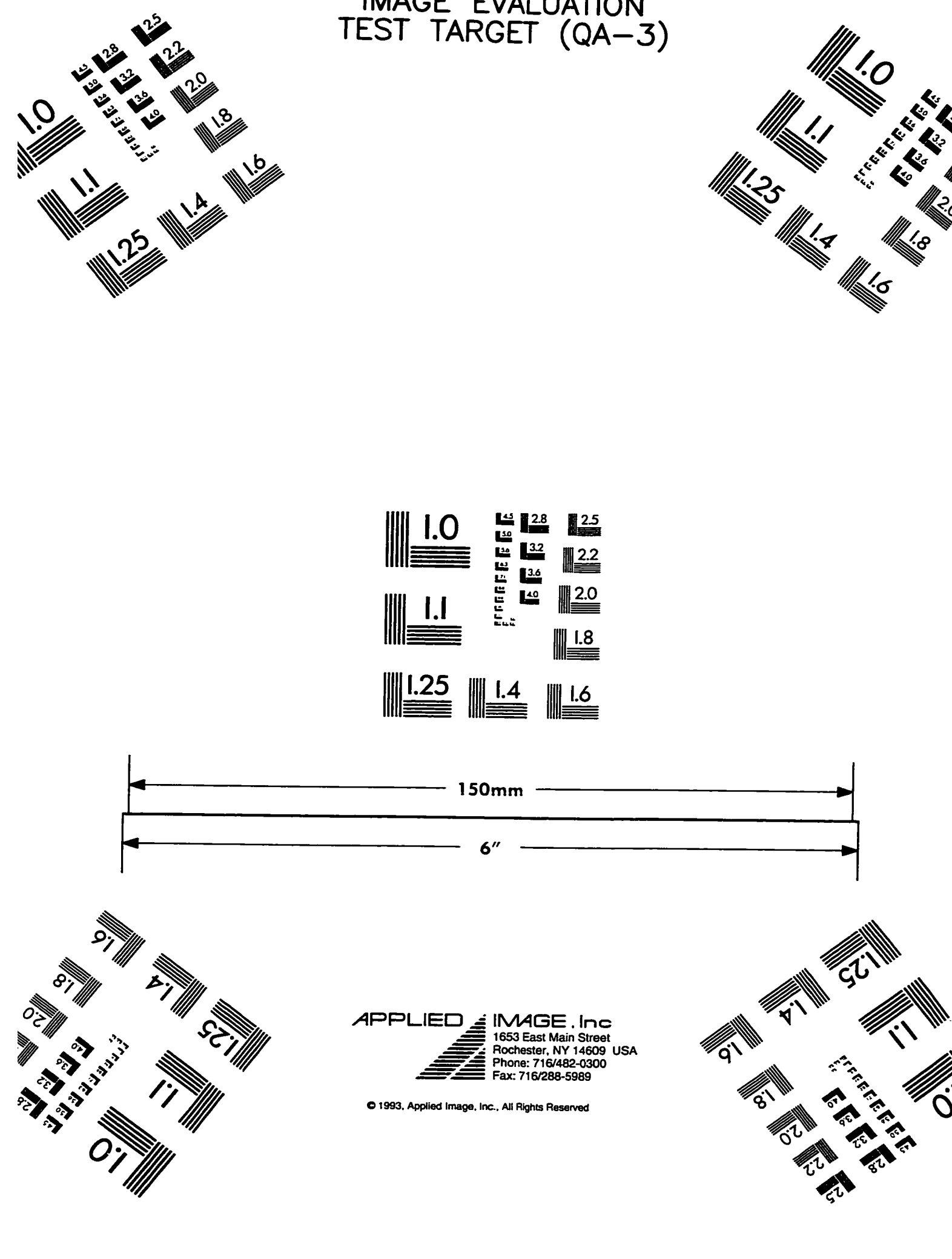


IMAGE EVALUATION
TEST TARGET (QA-3)
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
150mm
6"
APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989
© 1993, Applied Image, Inc., All Rights Reserved