A Robust Framework for Medical Image Segmentation through Adaptable Class-Specific Representation

A thesis submitted to Middlesex University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Casper Falkenberg Nielsen

School of Computing Science

Middlesex University

August 2002

THIS THESIS CONTAINS A CD WHICH WE ARE NOT PERMITTED TO COPY

PLEASE CONTACT THE UNIVERSITY IF YOU WISH TO SEE THIS MATERIAL

Abstract

Medical image segmentation is an increasingly important component in virtual pathology, diagnostic imaging and computer-assisted surgery. Better hardware for image acquisition and a variety of advanced visualisation methods have paved the way for the development of computer based tools for medical image analysis and interpretation. The routine use of medical imaging scans of multiple modalities has been growing over the last decades and data sets such as the Visible Human Project have introduced a new modality in the form of colour cryo section data. These developments have given rise to an increasing need for better automatic and semi-automatic segmentation methods. The work presented in this thesis concerns the development of a new framework for robust semi-automatic segmentation of medical imaging data of multiple modalities. Following the specification of a set of conceptual and technical requirements, the framework known as ACSR (Adaptable Class-Specific Representation) is developed in the first case for 2D colour cryo section segmentation. This is achieved through the development of a novel algorithm for adaptable class-specific sampling of point neighbourhoods, known as the PGA (Path Growing Algorithm), combined with Learning Vector Quantization. The framework is extended to accommodate 3D volume segmentation of cryo section data and subsequently segmentation of single and multi-channel greyscale MRI data. For the latter the issues of inhomogeneity and noise are specifically addressed. Evaluation is based on comparison with previously published results on standard simulated and real data sets, using visual presentation, ground truth comparison and human observer experiments. ACSR provides the user with a simple and intuitive visual initialisation process followed by a fully automatic segmentation. Results on both cryo section and MRI data compare favourably to existing methods, demonstrating robustness both to common artefacts and multiple user initialisations. Further developments into specific clinical applications are discussed in the future work section.

Acknowledgements

I would like to thank my director of studies Dr. Peter Passmore for inspiration and support and for giving me the freedom to develop my own ideas, while always keeping me focussed. Thanks to Professor Harold Thimbleby for bits of useful advice along the way, many of which left me pondering for some time. Thanks to Dr. Ian Mitchell for helping to tell the novel from the trivial in my early work, and to Dr. Paul Cairns and Brian Kunding for valuable input about algorithm formalisation. Thanks to Dr. Richard Spencer-Smith for originally introducing me to the work of Kohonen, Professor Steve Torrance for helping me to appreciate the wonders of emergence and to Dr. Thomas Bending for discussions about graph theory. Thanks to Professor Nigel Foreman and Stephen Nunn for advise on statistics. A special thanks to Professor Huw Jones for assessing my work and offering advice at my PhD transfer.

On the more practical side I would like to thank the Department of Surgical Oncology and Technology at St. Mary's Hospital for their collaboration at various times. A special thanks to Dr. Paul Ziprin for his help in recruiting participants for my human observer experiments, and to Professor Ara Darzi for making it possible. Thanks to Keith Humphries at the Radiological Services Unit, Hammersmith Hospital, for valuable discussions about visualisation and clinical applications at various stages throughout my work. Thanks to Dr. Vijay Jayaram for sharing ideas about evaluation of segmentation quality and for helping to test the computer based human observer experiments. Thanks also to Dr. Igor Aizenberg for his encouraging words following my first publication.

I thank the other research students, past and present, at the School of Computing Science, Middlesex University for their friendship and support. A special thanks goes to Dr. Konstantinos Giannakis for sharing ideas and good advice and for generally being a cool guy. A big thank you to our research administrator Sardia Alhassan for all your help from day one. Thanks to my friends outside the academic world who have stuck by me, especially to Stephen Fernandez, Pornthip, Simon, Romana, Ritu, William, to Katy for keeping me in touch with California and to those back in

Denmark. Thanks to Pia, Per, Berit, Annette, Sidsel, Jane and Palle for all the great times during my medical studies at University of Copenhagen.

Finally I would like to thank my parents for their encouragement always and for their love and support. I would not have been where I am now without you – literally. All my love to Pixie who has been sharing my life with me throughout most of my PhD. Thanks for putting up with it and for your unconditional support.

Images from the Visible Human Project appear courtesy of the United States National Library of Medicine. The author is a license holder.

The IBSR MR brain data sets 788_6_m and 1320_2_max and their manual segmentation were provided by the Center for Morphometric Analysis at Massachusetts General Hospital.

Acknowledgements of other images and resources utilised in this project are given in the relevant sections of the text.

This project was funded by an NFFR grant from Middlesex University.

London, United Kingdom
August, 2002

Casper F. Nielsen

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Introduction and background.

> Virtual Reality (VR), telepresence, interactive computer aided instruction, holographic visualization: This is the language of the new generation of medicine. If there is but one island of certitude within this technological maelstrom it is the realization that we have irrevocably crossed the threshold from the Industrial Age into the Information Age: Medicine is dead, long live Medicine. There no longer is Medicine, rather it is Information with a medical flavor [1].

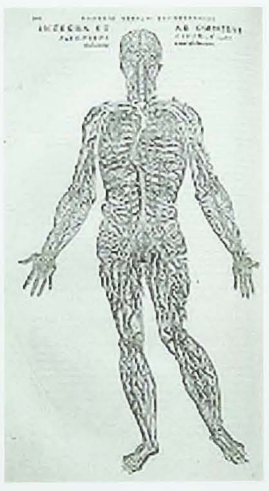
The above statement about traditional medical practice versus medicine in the modern age can be found in a state-of-the-art paper by Richard Satava in the 1996 proceedings of Medicine Meets Virtual Reality - an international conference devoted to the fusion of traditional medical science and modern computer based technologies. The fact that this conference has taken place for the 10th consecutive year in 2002 shows that technologies under the broad term *medical informatics* [2] have become an established field of research. The term of course is an umbrella for several specialised strands, including advanced database design, expert systems, video conferencing and ever increasingly visualisation and image analysis. It is the latter that the research described in this thesis is concerned with. Visualisation and image analysis in medicine can in turn be broken down into several different areas, including image processing and enhancement, virtual reality, augmented reality, computer assisted surgery, and systems for education, training and simulation. As each of these areas grow and mature, the boundaries between them become blurred, as they merge into integrated systems.

In medical imaging, radiologists and medical doctors have traditionally had to rely on their interpretation skills of 2D projections of 3D anatomical structures. Such was the imaging produced by the classic modality X-ray [3,4] discovered in 1895 and later on by ultrasound [5] (1960's). In the early 1970's Magnetic Resonance Imaging (MRI)

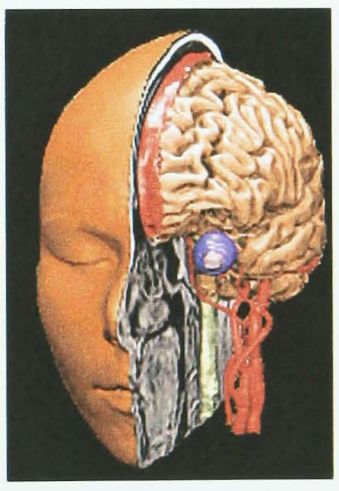
[6] and Computed Tomography (CT) [7] were introduced. These two new imaging techniques were capable of capturing precisely located sequences of slice images through the patient's body. This type of data opened up the possibility for creating 3D visualisations and the area of volume imaging was born. In 1979 Herman and Liu published the Cuberille algorithm (and its application to medical image reconstruction) [8], which became one of the first widely used algorithms for volume visualisation. Ultrasound could initially not easily provide volumes of localised slice images, because the sequence of acquired images depended on a person manually moving a transducer over the patient's body. This was later made possible using tracking of the probe and registration with the patient. This modality is known as 3DUS (3D Ultrasound) [9,10].

In volume rendering, models consist of voxels, a 3D equivalent of the pixel in 2D images. There are a variety of different rendering techniques (see [11] for a review), but traditionally volume rendering is achieved through ray tracing, summing up voxel values along rays, possibly with a particular weighing for each voxel. In 1987 Lorensen and Klein published their paper on the Marching Cubes algorithm [12] for extracting surface models from 3D volumes. Although surface models do not possess the ability of volume models for the viewer to "look inside", they provide the opportunity for more interactive viewing of 3D surfaces on less sophisticated hardware, due to the substantially smaller requirements for memory and processing speed. Marching Cubes works through triangulation, approximating structures with polygons where cells are intersected by the isosurface. The more triangles in a model, the greater the accuracy of the representation, but at the expense of processing overhead. Other types of surface fitting, such as the use of splines [13] and NURBS [14] have been successfully used in surface modelling to accurately visualise curves with less computational overhead, and a number of decimation algorithms (algorithms for polygon reduction), such as [15,16] have been introduced since Marching Cubes. They use local information to determine if two or more triangles or other primitives can be merged within predefined restrictions to reduce the number of polygons in a model, without seriously affecting the accuracy of the appearance. These techniques along with advances in processing power and rapidly decreasing prices on Random Access Memory and permanent storage devices have helped to bring 3D visualisation

within the reach of not only most doctors and medical institutions, but even home users. Volume-based rendering [17] is a compromise between volume rendering and surface rendering. In this approach only selected voxels are rendered, providing faster rendering than volume rendering and greater detail then surface rendering. Prior segmentation and classification is however required to label each voxel. Researchers at University of Hamburg produced the first full 3D reconstruction of the brain of a living human being, based on MRI scans [18] in 1987. This group has also done pioneering work in multimodal reconstruction with their VoxelMan range of 3D multimodal anatomical atlases [19] (see fig. 1.1). Different imaging modalities are better for visualising different types of tissue, for example MRI for soft tissue, CT for bone and DSA (Digital Subtraction Angiography) for blood vessels. VoxelMan features composite 3D models from different imaging modalities - imaging scans of the same specimen registered and fused to give the viewer the best of all the modalities. VoxelMan uses volume-based rendering and surface models and is implemented in Java, making it available on a variety of platforms and on machines with hardware requirements no higher than the average new PC of the late 1990's.



(a)



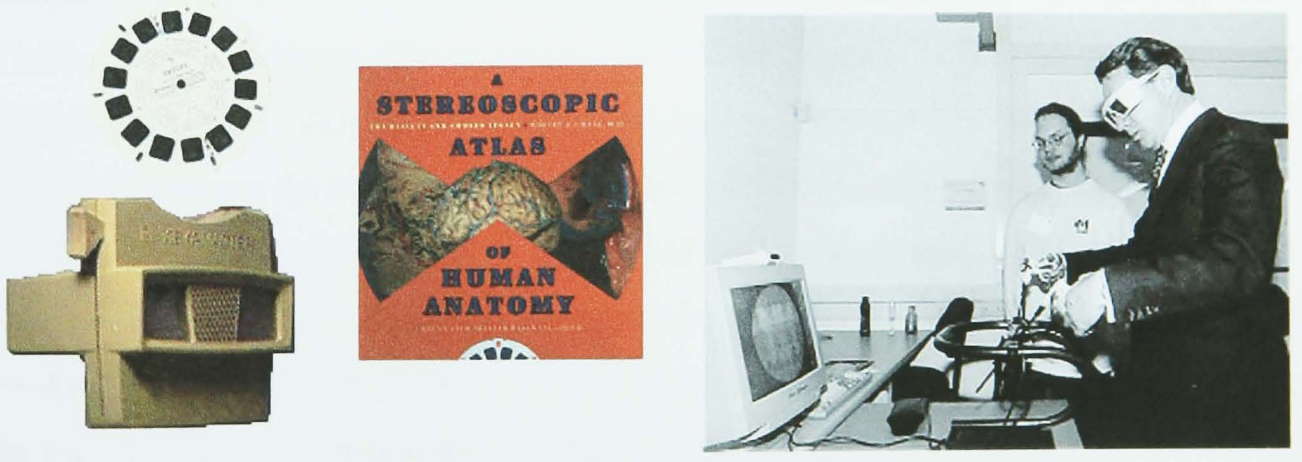
(b)

Fig. 1.1. Visualisation of anatomy throughout the ages. (a) Hand drawing by Vesalius, 1543 (from "De Humani Corporis Fabrica") [20]. (b) The VoxelMan atlas [19] using a combination of volume and surface rendering of segmented slice data from multiple modalities, 1995.

In 1998 Mitsubishi Electric launched the VolumePro 500 graphics board for PC's and SGI boxes. It was the World's first dedicated volume rendering board and allowed for real-time manipulation of large volume data sets even on low end workstations. The product line has since been acquired by TeraRecon and their latest model is capable of

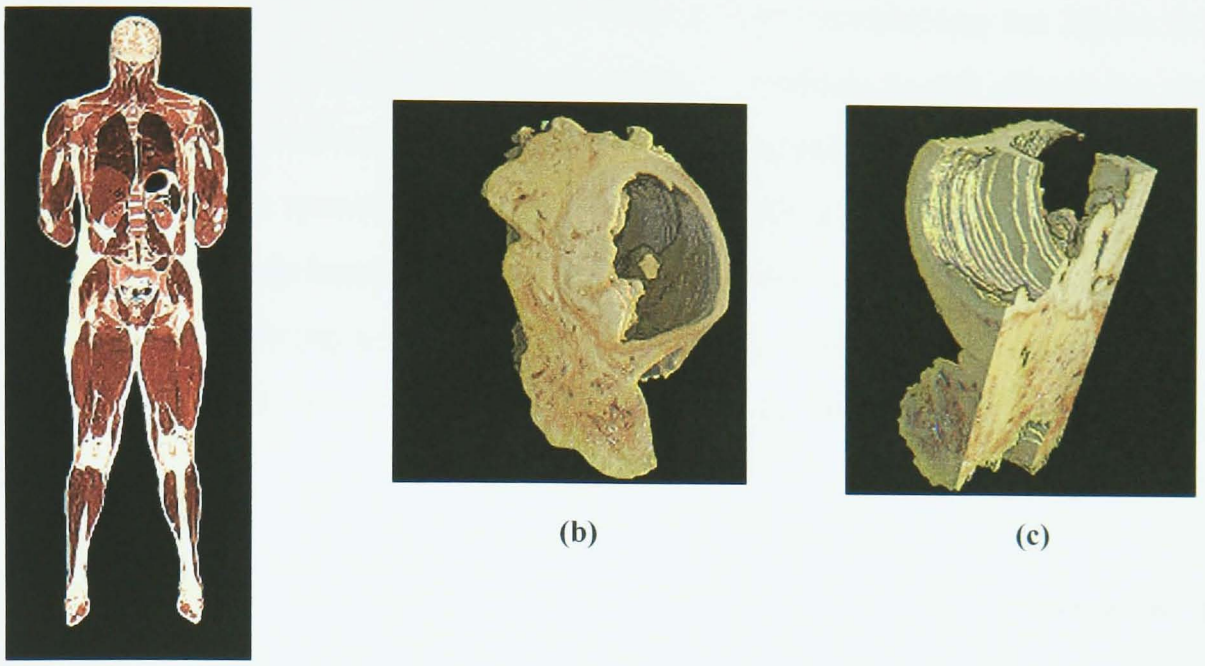
displaying volume models up to 512^3 (8-bit, 16-bit or 32-bit voxels) at a frame rate of up to 30 frames/second [21]. Hardware and software is thus rapidly catching up with the demand for visualisation of 3D volume data.

To achieve a better sensation of 3D for the viewer, stereoscopic viewing systems have become increasingly popular. Most systems use LCD shutter glasses synchronised with a monitor to display one image for each eye rapidly one after another (fig. 1.2(b)). Modelling the natural disparity of images between our two eyes is no new idea. The ViewMaster system [22] (fig. 1.2(a)) invented in 1939 was a cheap and easy way of creating stereo viewing and it is still in use today. It is a purely analogue device, which can be loaded with image reels containing two images per scene (one for each eye). From 1948 to 1962 Bassett and Gruber produced the "Stereoscopic Atlas of Human Anatomy" [23]. The project resulted in 1554 colour stereo images of dissections for use with the ViewMaster system. In modern digital visualisation systems, head mounted displays have been introduced for the fully immersive experience. Some of these are see-through and can be used to augment a real scene (see e.g. [24]). The CAVE [25] is another popular viewing system, where the viewer is situated inside a room with displays on the walls and on the floor.



(a) (b)

Fig. 1.2. The 3D visualisation of anatomical structures has been popular since before computers, let alone computer graphics, were available. (a) A typical 1960's model ViewMaster for stereo viewing of image pairs on cardboard reels. Also shown is the cover of Bassett and Gruber's "Stereoscopic Atlas of Human Anatomy". (b) A user interacting with a modern virtual reality endoscopic surgery simulator, getting a stereo view through LCD shutter glasses.



(a) (b) (c)

Fig. 1.3. 2D and 3D reconstructions from cryo sections. (a) Reconstruction of full coronal slice plane from Visible Human Male transverse slices (a full slice is shown in fig. 2.4). (b) Volume rendering of colon specimen from the St. Mary's Hospital project with viewplane parallel to the original slice plane. (c) A reconstructed slice plane through the colon specimen.

Apart from advances in hardware and software, what has also made products such as VoxelMan possible is the availability of complete, high quality, high resolution multimodal medical imaging data sets. The most ambitious example of this is the Visible Human Project (VHP) from the United States National Library of Medicine [26,27]. The VHP currently has two multimodal data sets of slice data from two complete cadavers, a male and a female (referred to as the Visible Human Male and Visible Human Female). The cadavers were frozen in gelatine, sliced in 1mm slices and photographed. Before and after being frozen the cadavers were CT and MRI scanned. Licenses for using the data sets are available to researchers worldwide. For the first time large scale volume models of real, photographed tissue were made possible (see fig. 1.3(a)). Other similar projects since have included the Stanford Visible Female [28] and the Whole Frog project from University of California [29,30] for use in schools to replace the dissection of real frogs in biology classes. In 1998 I (the author) was involved in a collaboration with the Minimal Access Surgery Unit (MASU) at St. Mary's Hospital in London on a project of visualising colon cancer [31]. A sectioned piece of colon with a large polyp (similar in many respects to a cancer tumour) was frozen, sliced, photographed and volume rendered (see fig. 1.3(b-

c)). The stage of colon cancer can be determined by establishing the layers in the colon that have been penetrated by the tumour. A Volume model allows for virtual dissection (virtual pathology) and one could imagine models being distributed over the Internet to enable specialists anywhere in the world to examine them. This type of technology has already been investigated for microscopy images and medical imaging scans within the growing area of telepathology [32]. Particularly The United States military has been interested in setting up systems to implement such technologies [33].

During the last five years, an alternative to the VHP has been in preparation. The Visible Korean Human [34] (a project of the School of Medicine, Ajou University, Korea), will provide a cryo section volume of a complete cadaver, setting new standards for spatial resolution. The data set will consist of 9000 cryo sections of 0.2mm slice thickness. Each pixel will represent an area of 0.2mm*0.2mm. This will allow for a much higher level of detail in visualisation than what was possible with the VHP. In addition to the cryo sections, the Visible Korean Human will offer 1800 CT and MRI scans of 1mm slice thickness. Similarly to the VHP the plan is to make the full data sets of the Visible Korean Human freely available to researchers.

A separate issue from the actual visualisation, is the ability to analyse medical imaging data. Being able to visualise the data in 3D is desirable, but in most cases it is important to be able to view a particular isolated structure within a volume. Examples include tumours and/or their blood supply and a particular organ or section of an organ for assessing shape and size. The area of segmentation and classification, rooted in machine vision, is one which has been actively researched for decades. The story goes that Marvin Minsky at MIT in 1966 set the solution to the problem of machine vision and scene interpretation as a summer project for an undergraduate student [35]. It turned out that the problem would take more than a summer to solve. In the year 2002 segmentation and classification still represents a major problem in many different application areas, not least in the medical area. With the increasing amounts of medical imaging data that has to be interpreted by radiologists (as imaging equipment becomes more advanced and routine screenings more common), the need for accurate, reliable automatic or semi-automatic and user friendly segmentation

techniques increases too. DSA [36] mentioned earlier is a new imaging modality, of which image processing is an integrated part, and image processing is a big research issue for the three main modalities MRI, CT and ultrasound and increasingly for volumes of photographed tissue (also known as anatomical slice data or cryo section data). Although countless automatic and semi-automatic segmentation techniques exist, a lot of segmentation work is still done by hand. This may not be much of a problem for 2D images, although batch processing is still desirable, but certainly for the segmentation of 3D volumes this represents a major challenge. A volume may contain hundreds or even thousands of slice images. Segmenting these by hand is an unacceptably time consuming task.

A number of established techniques are available for medical image segmentation. These range from simple intensity thresholding, over semi-automatic shape based analysis to fully automatic statistical classification systems and neural network or fuzzy logic based systems. A review is given in chapter 2. The simplest methods requiring the least amount of user interaction often do not provide the required level of accuracy. Other methods provide accuracy, but at the expense of time consuming user interaction during the segmentation process. The more complex automated methods, potentially providing higher accuracy, often require manual parameter settings and/or the selection of training data in some form. Results are very dependent on this selection and tweaking the parameters for optimal performance is a task suited only for image processing experts, which normally precludes medical professionals. Many approaches are very specialised and not easily adapted to other application areas. There are dedicated packages for medical image processing and visualisation, such as Analyze [37], which bring together a host of these techniques in one package. At least this makes them more accessible to medical professionals - having a large selection of tools with one familiar interface. However defining the best image segmentation pipeline for a specific problem and a specific modality in such an application is not trivial. There have been attempts to introduce standard software development kits for machine vision applications. The Image Processing Toolbox for MatLab [38] implements many popular image processing algorithms within the MatLab environment. There are comprehensive C++ libraries available both from the academic world (such as VXL from University of Oxford [39]) and the commercial

world (such as OpenCV from Intel [40]). So far however none of these have become established as a de facto standard.

This brief introduction has highlighted that there are many excellent visualisation techniques available for medical imaging data. The level of detail possible today in applications for education, training and simulation is impressive. Computer generated data visualisation has become part of everyday life and we humans are used to composing and assessing images visually. However most of the current applications for the visualisation of anatomy and quantification of tissue types have relied on time consuming manual or semi-automatic segmentation to classify the slice data. It could be argued that developments in medical image processing have not managed to keep up with developments in visualisation techniques and their applications.

There is a need for accurate semi-automatic segmentation, which, based on the user's goal, can optimally use the available image/volume information, and which can be applied successfully to a variety of imaging modalities. Medical professionals should not be required to have expert knowledge in computer graphics and image processing in order to use such a system. There is a need for robustness and an acceptable balance between automation and user control.

This thesis presents research into the development of a robust semi-automatic segmentation framework to support discrete 2D medical images and 3D image volumes of multiple modalities.

1.2 Research objectives.

The research documented in this thesis aims to:

- Develop a robust framework for accurate semi-automatic segmentation with intuitive but minimal human intervention for 2D and 3D medical images of multiple modalities.
- Evaluate the proposed framework both quantitatively and qualitatively

Development of the segmentation framework should be achieved through the following stages:

- The identification of key problems with traditional image segmentation algorithms in general, as well as specifically for medical applications
- The theoretical development of a segmentation framework and algorithms to implement it, addressing key problems, with the initial implementation for a single image modality
- The extension of the framework to multidimensional data sets of multiple modalities

It was important that evaluation throughout the project should be based mainly on standard test volumes and previously published images. Cryo section data sets from the Visible Human Project have provided the images and image volumes for the initial stages of development. Other non-medical images have been used as benchmark tests in comparison to previously published results using other algorithms. Further development for the MRI modality has used simulated data from the BrainWeb image database [41] (see chapter 6, section 6.3.1) and real clinical data from the Internet Brain Segmentation Repository [42] (see chapter 6, section 6.3.2).

The empirical evaluation of the segmentation framework is based on:

- The comparison of segmentation results achieved within the proposed framework and through the use of other established methods, using visual presentation, ground truth comparison and visual ranking by human observers
- The variability in segmentation accuracy for the same data sets, initialised by different users, as a measure of robustness

1.3 Thesis overview.

This chapter provided an introduction and motivation for the research project described in this thesis.

Chapter 2 is a literature review of image segmentation. The major algorithms and concepts in the field are presented, discussed and related to medical image segmentation.

Chapter 3 identifies and discusses the main problems faced by automatic and interactive segmentation systems generally and specifically for medical image segmentation. A set of conceptual requirements for a robust semi-automatic segmentation framework are proposed and subsequently a set of technical requirements for the implementation of this framework. The chapter finishes by outlining the research methodology.

Chapter 4 introduces a feature vector encoding for natural colour images and cryo sections and a vector quantization neural network approach to classification. The problem of segmentation near edges is addressed through the development of a new algorithm, the Path Growing Algorithm (PGA), for class-specific representation of segment classes. The required user interaction for initialisation is specified and preliminary results using the PGA integrated with neural network classification are given. The combination of the initialisation, feature encoding, path growing and classification is introduced as the segmentation framework for Adaptable Class-Specific Representation (ACSR). Different segmentation pipelines are proposed along with an automatic focussing of the PGA for higher efficiency. The ACSR framework is extended from the 2D domain to 3D volume segmentation and results based on cryo section volumes from the Visible Human Project are presented.

Chapter 5 discusses the issues of quantitative ground truth evaluation versus qualitative visual ranking by human observers, for the empirical evaluation of segmentation algorithms. Three experiments in ACSR segmentation are presented, investigating the effects of initialisation by different users, and segmentation accuracy measured using ground truth comparison on a set of standard natural colour test images and a series of brain cryo sections.

Chapter 6 describes the development of the ACSR framework for greyscale MRI segmentation. Three modified Path Growing Algorithms are presented and tested in a

comparative study based on simulated BrainWeb [41] T1-weighed MRI data. Two optimisations are introduced, one targeting the problem of noise and the other targeting inhomogeneity. The EQ [43] and N3 [44,45] inhomogeneity correction algorithms are tested as pre-processing tools and further results using the proposed optimisations and the best PGA from the initial study are given. Finally results on multispectral segmentation of T1 and T2-weighed volumes are presented and discussed.

Chapter 7 presents the results of a series of human observer experiments, evaluating the quality of ACSR segmentation. The effects of different initialisations and overall segmentation quality are analysed for natural colour images and cryo sections. MRI data is also evaluated qualitatively, investigating the issues of single-channel versus multispectral segmentation, the effects of using EQ and N3 on the final observed segmentation and the quality of ACSR segmentation, compared to a gold standard manual ground truth.

Chapter 8 summarises and concludes the work presented in the previous chapters and future work is discussed.

Appendix A is a collection of papers published as a result of the work described in this thesis. It contains full text versions of four international conference papers and two technical reports.

Appendix B contains the notation for the SOM and LVQ algorithms used, as well as a definition of Sammon's Mapping and a section on fuzzy logic.

Appendix C is a discussion of the complexity and computational overhead of the PGA and gives the definition of a nearest-neighbour search algorithm used in the implementation of the PGA.

Appendix D concerns the human observer experiments described in chapter 7 and contains written material given to participants, as well as screen shots and a diagram of the computer based experiments.

Appendix E gives instructions about the use of the companion CD.

The companion CD contains sets of images/volumes and their segmentations described in chapter 5, 6 and 7. The contents of the CD is not essential for the understanding of the work described in this thesis, but allows readers to perform their own visual assessment of images and volumes which are not reproduced in the thesis. Please consult appendix E and the file readme.txt on the root directory of the CD for further information.

Chapter 2

Background and related work

2.1 A literature review of image segmentation.

Image segmentation is the process of breaking an image into meaningful components. Segmentation goes hand in hand with classification, which is the process of putting labels on the segmented components. A segmentation depends on the desired classification scheme, since the same image might be segmented differently depending on the desired segment classes. Segment classes are identified through prior classification. It is also the case that classification depends highly on the achieved segmentation, which may not always be identical to the desired result (the so-called *ground truth segmentation*). Thus a classic dilemma is that "segmentation begs classification and classification begs segmentation" [46]. Borrowing a metaphor from the area of neural networks, segmentation approaches may be broken into two different types: supervised and unsupervised. Purely unsupervised approaches are normally referred to as automatic, while supervised approaches range from purely manual to semi-automatic. Manual segmentation is achieved by segmenting an image by hand. Because automatic and semi-automatic approaches do not always deliver the required level of accuracy, manual segmentation may be preferred, even if it is cumbersome [47]. Manual segmentation is also used to clean up images after crude automatic segmentation. The quote from [46] can be extended to: The more manual classification, the better the segmentation - the more manual segmentation, the better the classification. Establishing context is always a desirable achievement in machine vision [48]. Letting the user classify an area before segmentation makes the task much easier. Therefore many segmentation approaches are based on substantial interaction with the user and are thus time consuming.

This literature review is compiled from the perspective that segmentation of medical images should be considered a special case of the more general image segmentation problem. Most segmentation and image processing algorithms are transferable between different application areas. Simply looking at the literature, which is strictly inside the area of medical image analysis, is not sufficient. Thus most of the algorithms and models described in sections 2.1.1-2.1.7 are general image segmentation methods, which may all be applied to the special case of medical images. Section 2.2 summarises the review in the context of medical image segmentation and briefly discusses the advantages and drawbacks of model based vision in medical image analysis.

2.1.1. Edge detection and filtering.

Edge detection is based on the detection of significant intensity changes in an image, corresponding to high frequencies in the signal. Even before computers became generally available, edge detection was a known signal processing technique in the area of image transmission. As far back as 1959 Julesz published a paper on using edge detection as a simple form of compression for television signals [49], by only transmitting the higher frequencies (corresponding to points of high gradient magnitude). Edge detection as a biological process was suggested first by Hubel and Wiesel in 1968 [50] (visual perception in monkeys) and later by Marr [51] (human visual perception).

Edge detection algorithms work either in the spatial or frequency domain of an image. Examples of spatial filters for edge detection include the Rober's Cross [52], Sobel [53], Prewitt [54] and Canny [55,56] filters. These filters all work on the first derivative of the image (the intensity gradients). Gradients are found by convolving an image with a kernel and usually applying a threshold afterwards. This threshold value and the size of the kernel can be varied according to the application. The Sobel filter uses a smoothing operation first, making it less vulnerable to noise. The Laplacian filter [57] works in a similar way to the previously mentioned filters, but on the second derivative of an image. A positive peak in the first derivative caused by a

sudden increase in intensity in the image, results in a positive peak followed by a negative peak in the second derivative. This is known as a zero-crossing. The zero-crossing specifies the location of the edge. In theory this makes the Laplacian filter better at *locating* boundaries, but as a second derivative method it is also very sensitive to noise. Marr and Hildreth addressed this problem by using a smoothing operation first (similar to the Sobel filter). This filter is the Laplacian of the Gaussian (also known as the LoG operator) [58].

Frequency domain filters include the ideal high pass filter [59] and the high pass Butterworth filter [59]. Frequency filters suppress particular frequencies after transforming the image from the spatial to the frequency domain. This is achieved using the Fourier transform [57]. The Discrete Fourier Transform (DFT) for an *M*N* image is defined as:

$$F(m,n) = \frac{1}{\sqrt{MN}} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} f(k,l) e^{-2\pi j \left[\frac{km}{M} + \frac{ln}{N}\right]} \qquad \textbf{(2.1)}$$

f(k,l) is the input image in the spatial domain at position *k,l*. *e* is the natural exponent and *j* is the square root of –1. The DFT can also be written as a transformation matrix for a given *M*N* image for easy computation of the DFT of the image.

Lowpass filters suppress high frequencies resulting in smoothing. Highpass filters suppress low frequencies. This may be used for edge detection, but in most cases too many false edges are found. Spatial filters are generally preferred for edge detection, while highpass frequency filters are used to simply enhance high frequency information in images [59].

Filtering in the frequency domain is equivalent to convolution in the spatial domain. As previously mentioned this requires the definition of a finite kernel, which is shifted over the image matrix and multiplied with the corresponding points. The convolution equation for the image *f(x,y)* with the convolution kernel *h(m,n)* is given as:

$$g(x,y) = h(m,n) \otimes f(x,y) = \sum_m \sum_n h(m,n) f(x-m, y-n) \qquad \textbf{(2.2)}$$

The output image is $g(x,y)$. This means that the value at each point in the output image is found by multiplying the elements representing the $M*N$ neighbourhood around the point at $f(x,y)$ with the elements representing the $M*N$ kernel. An example is given in fig. 2.1, which shows an image filtered with the mean filter for smoothing. The mean filter kernel is an example of a uniform kernel, because all elements are the same. Examples of non-uniform kernels are given in chapter 4, section 4.3.1.

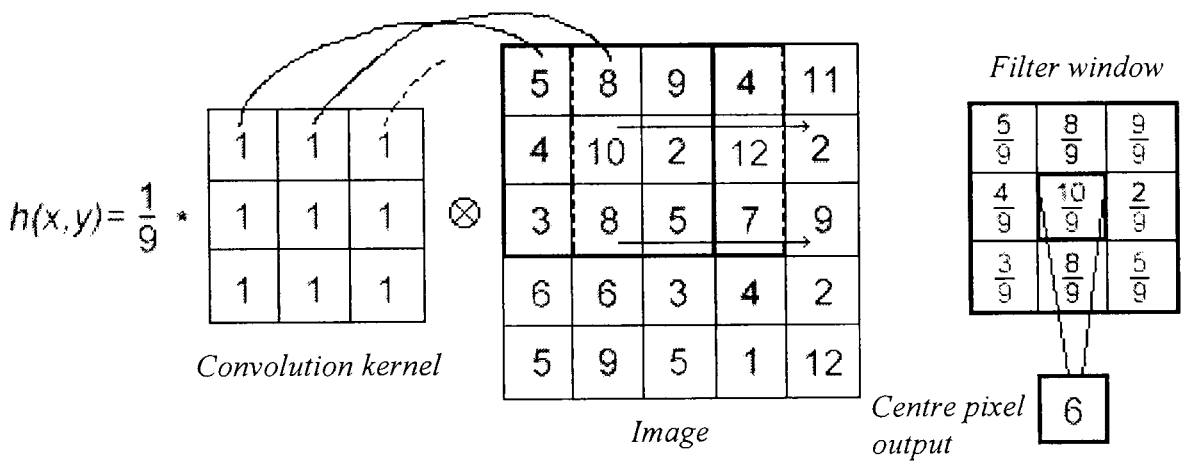


Fig. 2.1. The 3*3 convolution kernel for the mean filter applied to the top left 3*3 neighbourhood of a 5*5 image. Filter window is moved across the image and the kernel is multiplied with the neighbourhood surrounding the pixel in the image at which the output value is produced.

Rank order filters [60] are a special type of filters which require the ranking of point intensities inside the filter window. The most common example is the median filter [59], which ranks the intensities from lowest to highest value and assigns the median value to the centre point. This filter is better at preserving edges than the mean filter.

Edge detection is useful for segmenting the contour of objects on a homogeneous background. However, it has problems when it comes to more complex segmentation tasks. First of all edge detection is sensitive to noise. As previously mentioned this problem can be reduced by using pre-filtering. Obviously a lowpass filter in the frequency domain will suppress not only the noise, but also actual edge points. Some spatial filters for smoothing are better at preserving the actual edges while efficiently removing noise, but it is inevitable that some detail is washed out in the process. However, the biggest problem with edge detection is that there is no easy way of distinguishing between local high frequency information within a segment class and

high frequency information, which is actually a segment boundary. In the following an *edge* shall refer to any significant continuous change in frequency within an image and within and between segments, while a *boundary* shall refer to the border between segments only. Thus a boundary is an edge, but an edge is not always a boundary. What constitutes boundaries depends on the desired segmentation and is often entirely subjective. It may be desirable to perform a segmentation, where segment classes contain local high frequency information. In this case no edges should be identified as boundaries inside a segment. It is for this type of problem that edge detection fails to correctly produce only segment boundaries. Soft boundaries may appear too thick. Edge thinning [61] is a common image processing step following edge detection. Another is edge linking [61], which attempts to join fragmented edges into one continuous chain. A similar, common solution is the following segmentation pipeline (also known as a closing operation) [61]:

Segmentation → Dilation → Erosion → Thresholding

Dilation and erosion [61] are morphological operations, which enlarge and shrink image areas respectively. Following segmentation, dilation will join fragmented edges and the edges will then be trimmed to a smaller thickness using erosion. Thresholding in image processing is the process of masking out intensities above or below a threshold [57] (see section 2.1.2). It is useful for removing unwanted edge detections, based on the assumption that actual edges have higher magnitudes in the gradient image. One of the specific goals of the Canny filter [55] was to achieve accurate location of boundaries, taking the Laplacian of the Gaussian one step further. The filter combines the above mentioned processes into one. Smoothing is achieved by convolution with a 2D Gaussian function (or two 1D Gaussians, one for each axis). The gradient values are found and magnitudes and directions are calculated. Non-maximum values perpendicular to the direction of edges are suppressed. This is done by interpolating the values on each side of the current pixel in perpendicular direction to the edge within a 9-pixel neighbourhood. If the interpolated gradient magnitudes are greater than that of the current pixel then the pixel is suppressed. This is because another specific goal of the filter was that each edge should only give one strong response. It is effective for this purpose, but problems are encountered at sharp

curvatures of edges. Finally a thresholding operation is carried out. A lower and an upper limit is used. All pixels with values inside these two limits are accepted if they are connected to pixels with maximum responses. The Canny filter solves some of the problems encountered with other edge detection filters, but as already mentioned it is not well suited for detection of objects with sharp corners.

Multiscale edge detection [62] uses edge detection on multiple scales of an image to determine if edges are consistently detected over a number of scales. Scaling is achieved through a gradual blurring of the image. This makes it easier to accommodate textures, which exhibit local intensity changes inside strong object boundaries, as well as textures which exhibit close to no local intensity changes inside weak object boundaries, but there is no definitive way of combining the results from different scales.

Because edge detection is both fully automatic and generally fast, it is very commonly used as the first step in unsupervised segmentation methods. It is also part of many hybrid approaches. It may for example be combined with region growing (see section 2.1.3). Edge detection is rarely ever a perfect segmentation tool for natural images and requires manual post-processing. The best parameter settings for a given image also requires some experimentation.

2.1.2 Thresholding, histogram analysis and intensity occurrence matrices.

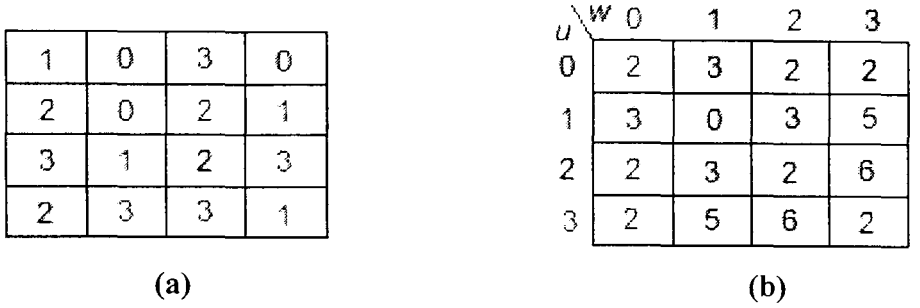
Thresholding [57] is the most simple form of segmentation available. If the intensity of an image point is within certain thresholds then the point is considered to be part of a segment with a particular classification. A number of different segment classes may be defined in terms of their upper and lower intensity thresholds. Adaptive thresholding [61] can use different thresholds for different image areas based on local histogram profiles. The image histogram [57] gives the frequencies of occurrences of all intensity levels (or bins representing a fixed range of intensities) in the image or image region. An image consisting of homogeneous segments on a high contrast background can be easily segmented based on its histogram, where peaks will

correspond to the intensities of regions. However the method is highly sensitive to artefacts.

In X-ray, MRI and CT scans grey levels directly correspond to tissue types (bone, soft tissue, air, fluid, etc.). This is why thresholding has been a popular segmentation method and most packages for the visualisation of medical images support thresholding. There are however serious problems such as different intensity ranges for different patients. The imaging equipment and in X-ray the choice of media (analogue film or digital) also affects the grey level ranges for the same tissue of the same patient for different imaging volumes. There can even be high levels of variability within the same volume due to inhomogeneity artefacts, which cause the boundaries between grey level ranges for different tissue classes to overlap (see chapter 6, section 6.6). Low resolution, sampling artefacts and slice thickness all contribute to partial volume artefacts, where one voxel may represent several tissue classes and have a grey level which reflects this, making it ambiguous. Noise is ever present caused by numerous sources, including the imaging equipment, the patient anatomy and the environment. These problems prohibit simple thresholding from producing acceptable results on most clinical data.

Although thresholding may describe a grey level distribution in terms of its intensity range, it does not take spatial patterns into consideration. In other words the relations between different grey levels within neighbourhoods of the same texture are not used. These are the relations that describe texture. The co-occurrence matrix [63] attempts to capture these relations. The co-occurrence matrix is a two-dimensional histogram describing the joint probability of two pixels with specific intensities occurring in a particular relation to each other (thus the term co-occurrence). The relation is defined in terms of a direction and a distance. Haralick [63] suggested that the co-occurrence matrices for the angles 0, 45, 90 and 135 degrees and distances of 1 and 2 pixels should be calculated as a minimum to produce good results. Any direction and distance is possible though, it only comes down to processing overhead. The co-occurrence matrix for a pixel neighbourhood is V^2 in size, where V is the number of intensity levels. Fig. 2.2 shows an example of a 4*4 neighbourhood containing pixels of 4 different intensity levels (0...3) and its co-occurrence matrix. Co-occurrence

matrices may be used on their own for pattern recognition or more often in combination with other statistical methods.



1	0	3	0
2	0	2	1
3	1	2	3
2	3	3	1

(a)

u \ w	0	1	2	3
0	2	3	2	2
1	3	0	3	5
2	2	3	2	6
3	2	5	6	2

(b)

Fig. 2.2. The co-occurrence matrix. (a) A 4*4 pixel neighbourhood showing intensity levels 0...3. (b) Co-occurrence matrix for the neighbourhood using the distance 1 and the angles 0, 90, 180 and 270 degrees. The cells show p_{uw} which is the number of occurrences of two pixels with intensity levels u and w at a distance of 1 between them. The rows and columns of the matrix represent u and w.

The run length matrix [64] is related to the co-occurrence matrix, and describes the occurrence of certain intensities within a local image area. The run length matrix describes the frequency of chains of points with the same intensity level, up to a specified run length, occurring in a particular direction. The lengths of the two dimensions of the matrix are defined by the maximum intensity level and the maximum run length. As for the co-occurrence matrix, the run length matrix can be defined for different directions and it may be used as a classifier in its own right or in combination with other statistical approaches.

2.1.3. Region growing, split-and-merge and watershed segmentation.

Edge detection finds boundaries, which define segments. Region growing on the other hand finds segments, which define boundaries. In traditional region growing a seed point is manually selected in an image within a segment and the classification is given manually. Based on its classification the segment is grown from the seed point according to a certain criteria. This criteria defines whether or not a neighbouring pixel is added to the region. It could be based on colour information or intensity. As a new pixel is added, the neighbours of this pixel are evaluated and the region keeps

growing until no neighbours of any pixel included in the region can be added. Adaptive region growing [65] attempts to improve on traditional region growing by analysing the region every time it is grown. If the current region does not match the previous classification as well as another, the classification is changed before the region is grown again. Region growing has some advantages over edge detection. Impulse noise does not affect the segment boundaries as much as it does in edge detection. On the other hand pixels with extreme values cannot be added to a region. Pixels in the boundary area may not fall within the specifications for a particular classification, which may result in uneven and incorrect boundaries. If a classification is given with the seed point in traditional region growing, it means that a pixel can either fit under that classification or not. There is no competition between segment classes. This means that too many or too few pixels may be added at boundaries. Adaptive region growing attempts to go beyond single pixel level and analyses the whole region continuously, so that newly added pixels may change the classification. Boundary pixels have little effect on a larger segment though and it is again possible that too many or too few pixels may be added at boundaries. Region growing is particularly good for semi-automatic segmentation of images with homogeneous textures, but requires a considerable amount of user interaction to specify the criteria for growth and selection of the seed points.

In order to reduce the problem of segmentation near boundaries, region growing has been combined with edge detection in hybrid segmentation algorithms (e.g. [66,67]), which all still rely on and assume homogeneity of the regions. A twist on the combination of edge detection and regions expanding from points comes in the form of watershed segmentation [68]. Watershed segmentation is based on the gradient magnitudes of an image and can be thought of as a flooding starting in all the valleys of a landscape (where the gradient magnitude determines the height of the terrain). Regions expand from local minima and stop when they collide with other expanding regions. It results in continuous boundaries located at the edges within the image. This is the advantage over edge detection, which as previously mentioned may require other processing to close discontinuities of the boundaries. The disadvantage is that a vast over-segmentation is produced and further processing is required to merge segments. The problem can be reduced by producing the gradient at a higher scale

(higher flood level). This will eliminate some of the smaller false segments inside the true segment boundaries, but may also eliminate whole true segments or fine details. See fig. 2.3.

Split-and-merge techniques [69] first over-segment an image based on some criterion of inhomogeneity and then merge some of the segments again according to the opposite criterion of homogeneity. The standard split-and-merge will consider the image as a rectangle and then split it into four equally sized rectangles. Each of these rectangles are again split up into four and this continues until every rectangle is considered to be a homogenous region. When the splitting has been completed, neighbouring regions will be compared and merged if they fit the homogeneity criterion. Split-and-merge is an excellent technique for quickly determining the presence of a simple object in a simple scene, but it is not suitable for precise segmentation with accurate boundaries.

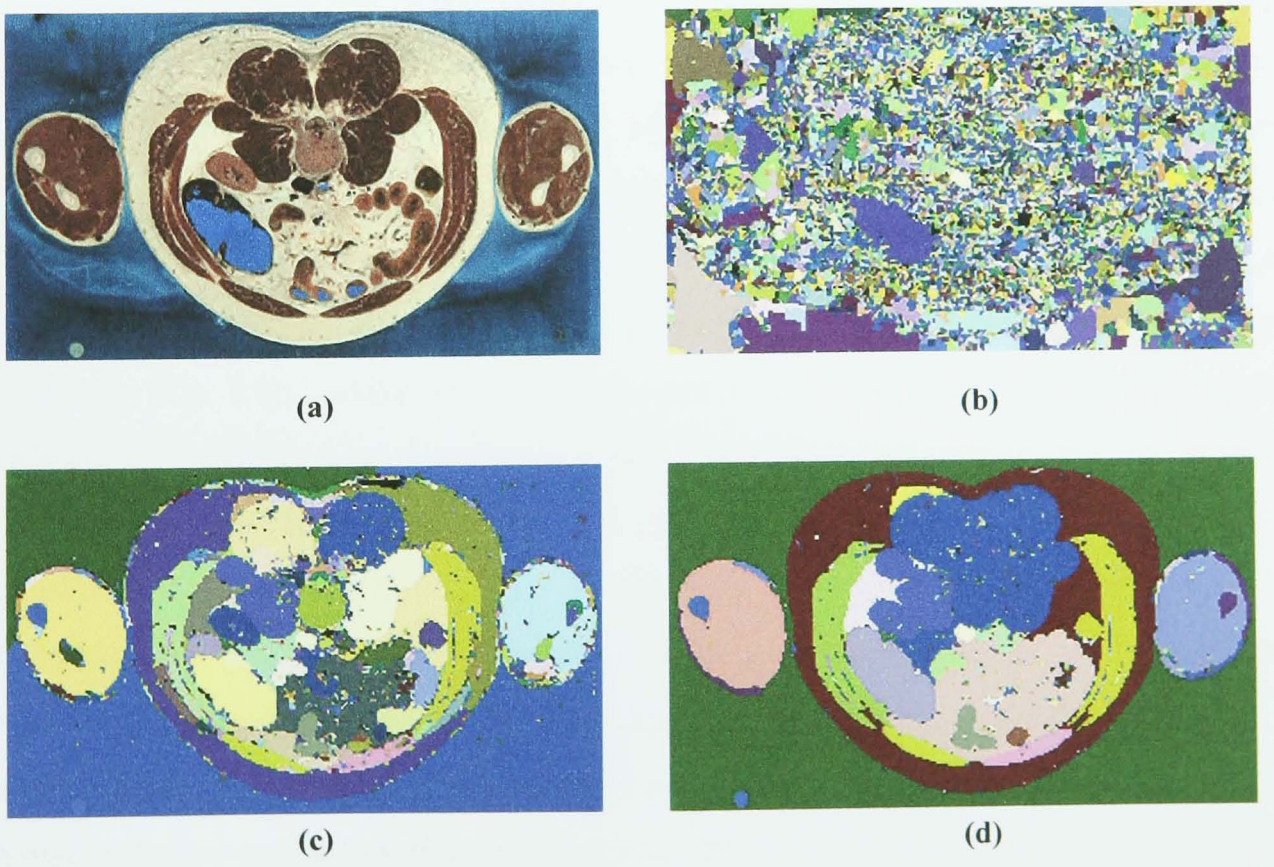


Fig. 2.3. An example of watershed segmentation. (a) Cryo section from the Visible Human Project. (b-d) Watershed segmentation with increasing flood level (reproduced from [70]).

2.1.4. Integral transforms, multifractals and texture analysis.

Beyond point intensities and colour, a segment may be described by texture as mentioned in section 2.1.2. Texture analysis can be seen as an extension of edge detection. It is a detection of patterns in intensity changes. As previously mentioned, the Fourier transform can transform an image from the spatial to the frequency domain and the dominant frequencies can be found. This can reveal the orientation of a regular texture and its frequency components (see fig. 2.4). The Fourier transform however will not distinguish textures, which contain the same frequencies, but where the frequencies occur at different times (fig. 2.5). Intensity changes at specific positions in a texture in the spatial domain correspond to changes in the signal at specific points in time in the frequency domain. In signal processing terms the Fourier transform is limited to identifying stationary signals (frequencies occur at all times) and cannot differentiate between non-stationary signals (frequencies occur at different times) containing the same frequency components. Because the Fourier transform uses a sampling window, which stretches from minus infinity to plus infinity, time information cannot be recovered. This property however gives the transform perfect frequency resolution. The Short Term Fourier Transform (STFT) [71] was an attempt to solve this problem by using windows of finite size. This way a non-stationary signal can be broken into discrete parts, which are approximately locally stationary, and analysed using an appropriate window size. This does not however solve the problem completely, as larger windows give good frequency resolution but poor time resolution, and smaller windows give good time resolution but poor frequency resolution (fig. 2.6). Natural images are generally described by non-stationary signals. This is why the wavelet transform, which overcomes the windowing problem, is mostly used in preference to the Fourier transform in texture analysis. The wavelet transform [72], like the STFT, uses windows of finite size, but rather than using a finite window size at different points in time, a window size for each frequency component is used at all points in time. This is known as multiresolution analysis. A high response results when the window size fits with a frequency component in the signal and otherwise a low response is given. This way frequency components can be localised in time and textures with intensity changes in more complicated spatial relations can be identified (fig. 2.7).

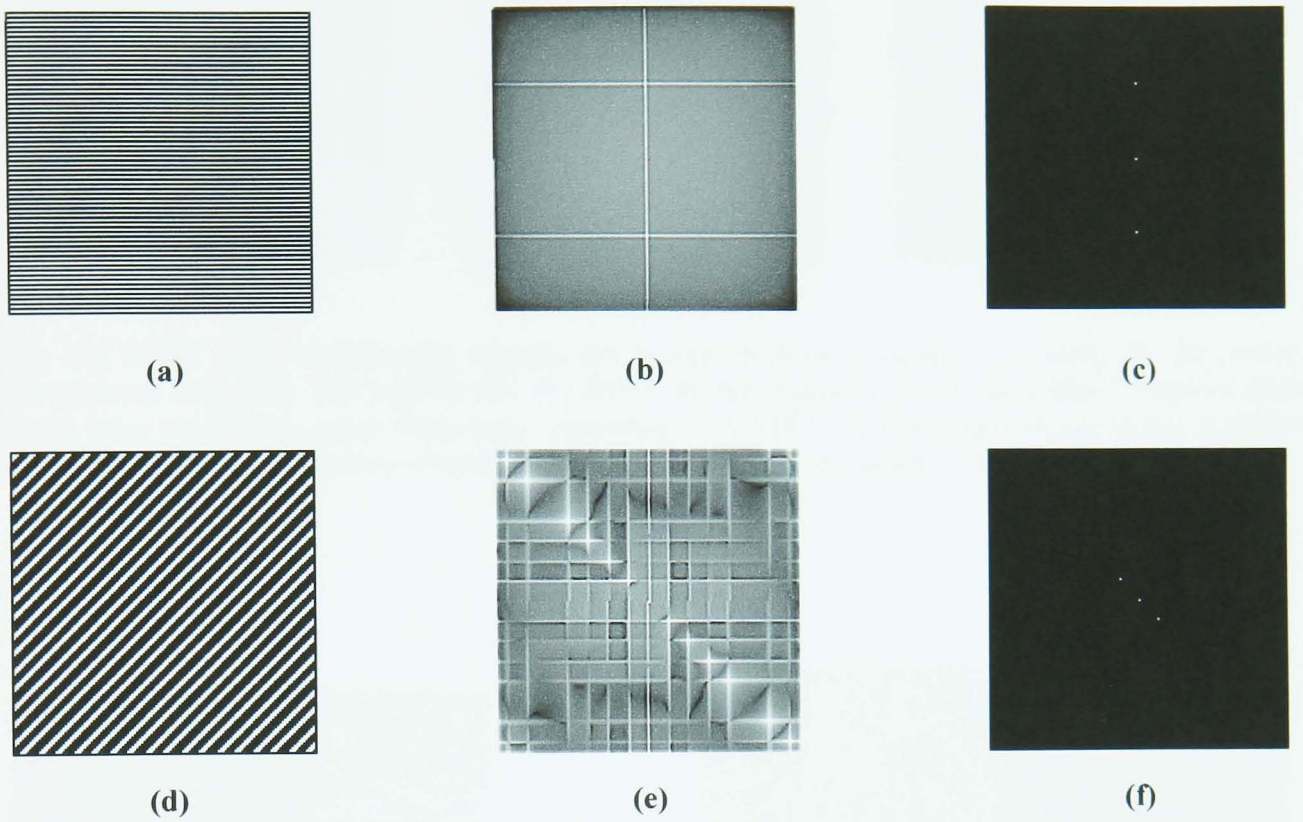


Fig. 2.4. The Fourier transform and regular textures. (a) Regular horizontal grating. (b) The logarithmic Fourier transform of (a). (c) Thresholding of (b) to reveal only the major frequencies. (d) Regular diagonal grating. (e) The logarithmic Fourier transform of (d). (f) Thresholding of (e) to reveal only the major frequencies. Images created with Vision XL from Impulse Imaging [73].

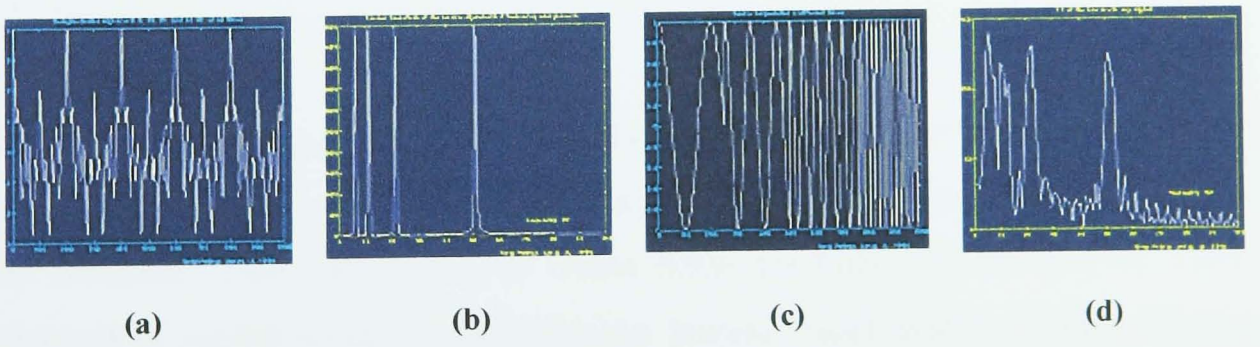
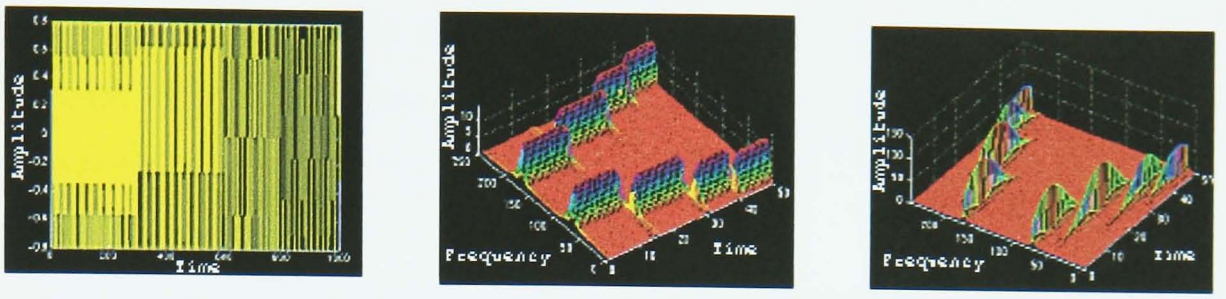
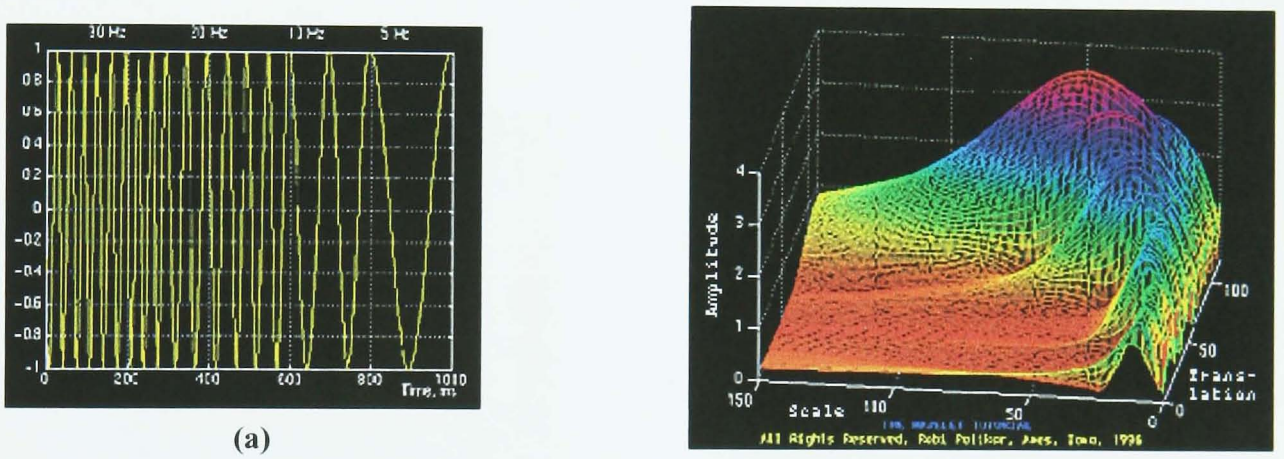


Fig. 2.5. FT of stationary and non-stationary signals. (a) A stationary signal containing the components 5, 10, 20 and 50 Hz. (b) FT of the signal in (a), clearly showing the four frequency components. (c) A non-stationary signal containing the same components, but at different times. (d) FT of the signal in (c) showing many frequencies other than the four real components, which are however still distinguishable. Time information is not available. Images reproduced from [74].



(a) (b) (c)

Fig. 2.6. STFT of non-stationary signals. (a) A non-stationary signal containing the frequency components 300, 200, 100 and 50 Hz. (b) STFT of the signal in (a) using a small window size: Good time resolution, poor frequency resolution. (c) STFT of the signal in (a) using a larger window size: Good frequency resolution, poor time resolution. Images reproduced from [74].



(a) (b)

Fig. 2.7. A non-stationary signal and its Continuous Wavelet Transform (CWT). (a) A non-stationary signal containing the frequency components 30, 20, 10 and 5 Hz. (b) CWT of the signal in (a). Translation represents time and scale represents frequency (high scale is low frequency and vice versa). Images reproduced from [74].

In image processing, wavelets have particularly been used for Gabor time-frequency analysis [75]. In this application wavelets are based on a Gaussian function modulated by sinusoids. These are known as Gabor filters (or Gabor wavelets) [76]. They are claimed to model receptive fields in the human visual system. Gabor filters give maximum response at a particular scale and orientation. A set of Gabor filters (normally referred to as a filter bank) corresponding to different scales and orientations may be used to identify different types of textures. The filters are normally selected manually for the filter bank, based on prior knowledge about the textures to segment. There have however been attempts to automate the selection of appropriate Gabor filters [77]. Classification is achieved by convolving an image with

the filter bank. The texture class is determined based on the magnitude of response from each filter. It is obvious that Gabor filters are best suited to the segmentation of homogeneous and regular textures.

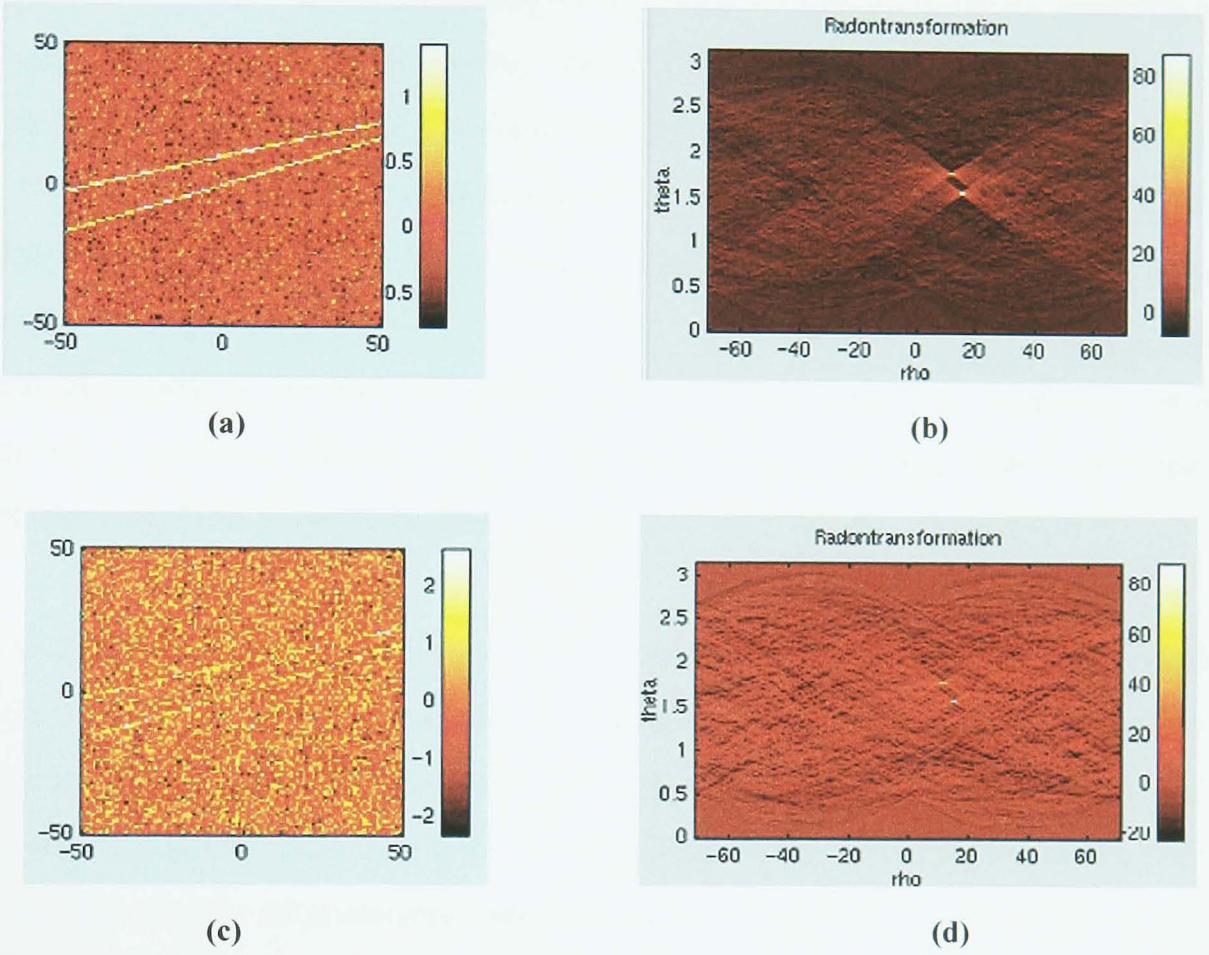


Fig. 2.8. The Radon transform for line detection. (a) Two lines with moderate random noise. (b) The Radon transform clearly showing the two lines in (a) as two bright peaks. (c) The same two lines with a high level of random noise. (d) The Radon transform still detects the lines in (c). Rho is the angle and theta is the distance from the centre. Images reproduced from [83].

Two other transforms often used for feature extraction in images are the Radon and the Hough transforms. The Radon transform was first described in 1917 [78] as a means of characterising 2D and 3D objects oriented at different angles. In 1968 IBM patented the Hough transform [79] for finding straight lines in images. The Hough transform is essentially just a special case of the Radon transform and the latter has since been applied in exactly the same way for finding lines in images [80]. The transform is particularly interesting for highly noisy images. It works by integrating image intensity along all possible lines in an image. A mapping from the image space to a parameter space is carried out, the parameters being the angle of a line and the

distance from the centre of the image. This means that a point in image space is a sinusoidal curve in parameter space. A curve in parameter space represents all possible lines through a point in image space and co-linear points in image space cause an intersection of curves in parameter space. If intensity represents the level of overlap, even highly fragmented lines can be seen as bright peaks in intensity in parameter space (due to many curves intersecting at that point). See fig. 2.8. A problem is that the inverse Radon transform maps back to lines of infinite length. Lines that were originally short thus appear long. This problem was addressed by Copeland et al with their Localized Radon Transform [81]. The transform has also been used to find other parameterised curves than lines (e.g. circles [82]). The problem obviously is that the generalisation of image features into a smaller set of parameterised curves is not always practically possible. For simple feature extraction though, the transform is very effective.

Many naturally occurring patterns and textures can be constructed using fractals [84]. Because fractals describe both the smallest building blocks and the largest structures, they are an obvious choice for analysis at multiple scales. This type of texture analysis known as multifractal classification was described by Voss [85]. Where the mass represents the number of primitives used to build a structure and the length represents the size of the structure, the mass dimension is the relationship between the mass and the length. Local mass dimension can be used to describe the type of structures or in other words the geometrical properties of an image. A classification can be based on local mass dimension, such as in the study of ultrasound liver images by Evertsz et al [86]. While this technique can be useful for detecting the presence of anomalies in images, geometrical structure is often not sufficient to differentiate between individual segment classes in complex images.

2.1.5. Shape based analysis.

Edge detection is a useful tool for finding boundaries, but has a number of drawbacks as mentioned in section 2.1.1. Too many edges may be found automatically and edges may have gaps that need to be closed. A closing operation may not produce a

desirable result. Active contours [87,88] (also known as "snakes") can overcome these problems, but require some interaction with the user. A coarse boundary around the actual boundary must first be given. This coarse boundary is a parameterised curve upon which internal and external forces work. The external forces translate, bend and stretch the curve, attracting it to areas of high gradient magnitude. This is achieved as an energy minimisation function, i.e. a minimum is reached when the curve describes the contour of a real boundary after a number of iterations. At the same time internal forces apply smoothing to the curve in order to prevent noise from deforming it. Active contours have been used successfully for the interactive segmentation of medical imaging scans [89,90] (see fig. 2.9). While this technique is fine for the segmentation of larger regions, it is unsuitable or at least very time consuming for the segmentation of many smaller structures, such as blood vessels.

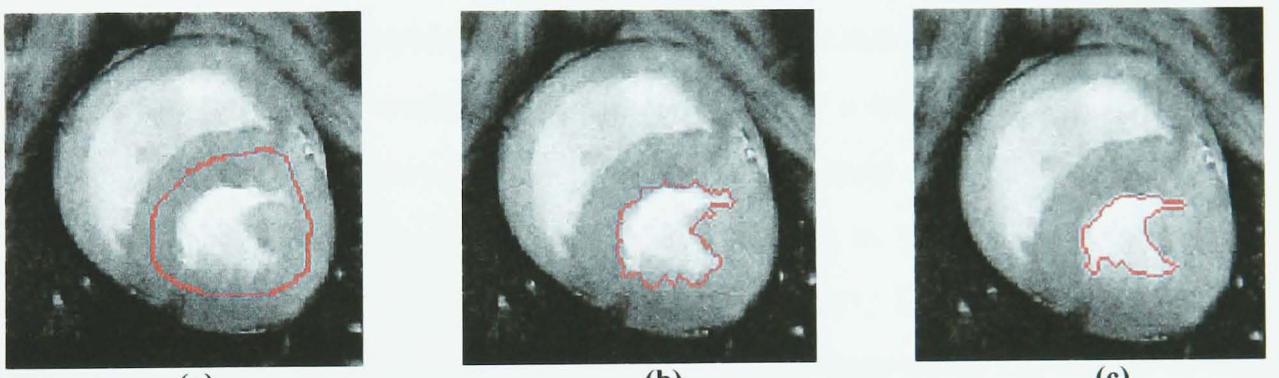


Fig. 2.9. The use of an active contour to find a boundary in a heart MRI image (reproduced from [90]). (a) The manual initialisation. (b) The contour at an intermediate stage after a number of iterations. (c) The final result.

Live-Wire [91] is an approach similar to active contours. The way the user interacts with the segmentation is very different though. Instead of selecting a coarse boundary, a single pixel is selected (start node). Subsequently another pixel (goal node) is selected and the algorithm finds the best path between the start node and the goal node. Finding the best path is, like in active contours, an energy minimisation function, where minimum energy is at the edges within an image. Each pixel has an energy based on gradient magnitude, a Laplacian zero-crossing (for boundary location - precise location of a pixel on a boundary equals low energy) and gradient direction (for smoothing - sharp change in boundary direction equals a high energy). The best

path is thus the one which gives the lowest cumulative energy between the start node and the goal node. It may be argued that Live-Wire gives more control to the user than active contours. On the other hand active contours force the user to provide an initialisation closer to the desired result, reducing the chances of serious errors.

Deformable surfaces have been used in similar ways to active contours for defining 3D boundaries. It is again an energy minimisation approach, only in this case a mesh of 2D primitives rather than a curve is fitted to the 3D object. A substantial amount of work on deformable surfaces, especially for the segmentation of medical volumes, has been done at INRIA in France. See [92] for a comprehensive review of developments in the field.

The introduction of active contours and related methods spawned a number of research projects aimed at making shape analysis more robust and less dependent on user interaction. Most notable is the work of Cootes and Taylor on Active Shape Models and Active Appearance Models [93,94]. From a set of training examples, in which boundaries are defined by sets of manually selected points (at least including points of sharp curvature), Active Shape Models can be trained to recognise and synthesise shapes given arbitrary transformations. Using statistical methods points retain their relative relations and initialisation of the shape may be started at some distance from the modelled structure in a novel image. After a number of iterations the shape is fitted to the desired structure automatically. Active Appearance Models combine the shape models with models of texture. Good results have been achieved in face recognition and medical image segmentation, but problems still remain to be solved, including the difficulty in selecting optimal points for the training phase and the fitting of shapes when initialisation is started too far from the target structure.

2.1.6. Colour.

All the segmentation algorithms described so far have traditionally been applied to greyscale images. They can be applied to colour images by working on a greyscale converted image or on each of the greyscale images that the three colour channels

represent. The advantage of colour is the richer information per point compared to greyscale. Much research has been concerned with how to represent colour in the most suitable way for human users and/or statistical analysis. The RGB colour model [95], which has a representation of colour triples for each point, was designed for light emitting media (monitors and televisions) - one intensity level for each of the three colour channels Red, Green and Blue. These are the primary colours of direct light, making the RGB colour model an intuitive model for display devices. However it is not an intuitive model for human beings, because it is not perceptually linear. Colours which are close in RGB space are not necessarily perceived as being similar by human beings with normal colour vision. This poses a problem for classification systems, which rely on RGB descriptors to model human colour perception. The HSV colour model [96] (and the similar HLS model [97]) was an attempt to create a more intuitive colour model for human users (although still an artificial model). This model's colour triple consists of the Hue (the dominant wavelength), the Saturation (purity of the hue) and the Value (intensity or brightness of the colour). Adding black reduces the brightness and creates different shades, adding white reduces the saturation and creates different tints, like a painter using his palette. Although more intuitive, the HSV model is still not considered perceptually linear. However its separate representation of colour and intensity can be a valuable feature for colour encoding and classification and gives it an advantage over the RGB model.

To devise a truly perceptually linear model, the CIELab colour model [95] was based on psychometric colour matching experiments. The CIELab model is based on four primaries: red, green, blue and yellow. These are also the colours that Ewald Hering described as the fundamentals of human colour vision in his opponent process theory [98,99]. The three types of cone receptors in the retina of the human eye have got three distinct peaks at different wavelengths in their light absorption curves. These correspond to the perceived colours red, green and blue (see fig. 2.10). Cells in the visual neural pathway receive the responses from the cones and perform an operation, which codes the input as relative responses. Green responses inhibit red responses while blue responses inhibit yellow (combined red and green) responses. The brightness is an additive process of all three receptors. These coded responses are transmitted to the visual cortex where colour perception takes place. In artificial

systems, the model has been implemented for different applications with more or less modification of the simple activation/inhibition operations described above. In the unmodified version (referred to hereafter as OPC), starting from the RGB model with 256 intensity levels per channel, opponent processes are modelled by subtracting the green channel from the red channel (R-G) and by taking the average of the red channel plus the green channel and subtracting the blue channel ([R+G]/2-B). A third achromatic channel is produced as ([R+G+B]/3). This model was used by Campbell et al [100] for colour encoding in a study of natural colour image segmentation and in a slightly modified version by Yamaba and Miyake for colour character recognition [101].

For the encoding of colour for a classification system there are clear advantages of models such as HSV, CIELab and OPC over the RGB model. The conversion between RGB and HSV is trivial, while conversion between RGB and CIELab is more complicated. Since the model is dependant on illumination, it also requires a reference white. If this is unknown and has to be estimated, then inaccuracies are inevitable. The much simpler OPC does not have this problem. While OPC does not serve as an alternative to CIELab as a base representation for conversion between other artificial models, it is highly applicable to the task of distinguishing coloured textures.

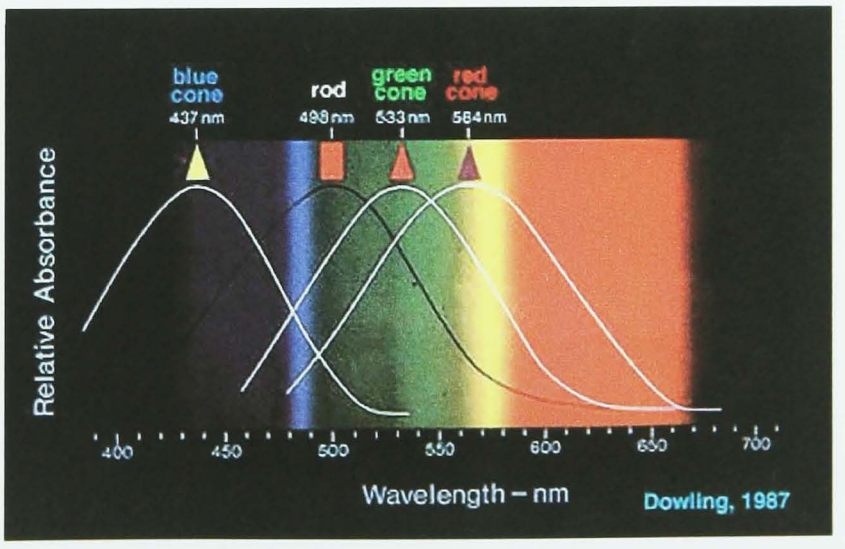


Fig. 2.10. Relative absorbance of light at different wavelengths in the human retina (reproduced from [102]).

In OPC, colours which appear as different shades, are represented similarly except for the achromatic part. These colours are more distinct in others models. For example the two RGB colour triples (200,100,50) and (250,150,100) are the same for R-G and Y-B. They are close in CIELab space but still distinct in all three descriptors and in HSV space the hue is the same, but saturation and brightness is different. This property makes OPC unsuitable for distinguishing between specific sets of uniform colours. It also makes it highly tolerant to changes in intensity and slight variations in colour across images, in a way which is intuitive to a human observer. This is an advantage for natural colour images or cryo sections. It follows that the OPC model is unsuitable for describing artificial images with uniformly coloured segments, where the HSV model is more applicable. In natural images, such uniformity rarely occurs. There is a sufficient spread of intensities in all three of the original RGB colour channels within a logical segment, for the OPC model to distinctly describe a local neighbourhood.

Although there is a substantial literature on colour image segmentation per se, the segmentation of cryo section data has received less attention than medical imaging scans. Most research in the segmentation of cryo section data has been based on data sets from the Visible Human Project. While it can be argued that the Visible Human data sets only need to be segmented once, it seems inevitable that there will be a need for good segmentation methods for cryo section data within the emerging area of virtual pathology. Approaches to the segmentation of colour cryo section data have remained fairly simple, using standard techniques. As part of the VoxelMan project at University of Hamburg a thresholding operation in RGB space was applied [103], essentially transferring the most basic segmentation method for grey scale medical imaging scans to the colour domain. Manual segmentation on a slice to slice basis was required for structures containing composite segment classes where one component was similar to a different class. In [104] Steward et al took a similar approach, although thresholding was applied to a "pseudo-radiographic" greyscale image produced from the original full colour image. Intensities represented the distance from predefined colours of the object to segment. It essentially came down to a point operation, which again poses a problem for structures with segment classes of some similarity in colour space. Edge detection and snakes (see e.g. [105]) as well as region growing (see e.g. [106], which also presents the notion of "virtual staining") have also

been applied to cryo images with the same advantages and disadvantages as described in sections 2.1.1 and 2.1.3.

2.1.7. Statistical methods, neural networks and fuzzy logic.

The techniques described in sections 2.1.1-2.1.4 can be used as low level feature detectors. Although they are used on their own, they are also often integrated in higher level classification systems. This is where the areas of image processing and statistical pattern recognition come together. The goals of such systems are to, as accurately as possible, determine the true probabilities of specific combinations of features leading to specific classifications.

In Bayesian methods a tree structure (Bayesian network or Belief network [107]) is used to connect nodes in a hierarchical manner. Each node is a state and has associated with it a number of probabilities of the state occurring, given all possible combinations of states of the input nodes (parent nodes). Image features might be lower level states leading to a higher level state - a classification. The original probabilities are known as the prior probabilities. The goal is to use new samples to tune the probabilities. When a new sample with a known classification is introduced, adjustments are made throughout the network such that all features of the new sample can point to the correct classification. This is done with many samples. What results is a system, which can take novel samples similar to those in the training set and produce the correct classifications. However more than that, the system describes the data itself. The tuned probabilities are known as the posterior probabilities. Bayesian systems are often implemented using Hidden Markov Models (HMM) [108] combined with an algorithm for parameter estimation. A HMM is a model consisting of chains of stable states and the probabilities of transitions between these states (Markov chains). In a Markov process only the current state and the probabilities of transitions to the next states are necessary to calculate the probability of the final state. Thus the presence of a feature indicating a state can give the probability of a classification. Obviously the probabilities of state transitions may sometimes be well known (tossing coins, throwing dice, drawing cards from a deck, etc.) and they may

sometimes be mostly unknown, especially in large feature spaces (for example shapes and textures in images). Given enough knowledge about the domain, one can obtain highly accurate posterior probabilities. However, mutually exclusive features (states) may not be taken into account, if their exclusiveness is not explicitly modelled. The same can be true for highly correlated features in complex and large feature spaces.

A Markov Random Field (MRF) [109] can be regarded as a 2D (or 3D) version of the HMM (since Markov chains in a HMM are 1-dimensional) and can be used to describe texture. It may be applied either as a pattern recognition tool or for image restoration (restoring an image from incomplete data) or synthesis. A texture is considered as an instantiation of a random field of intensities relating to each other in ways, which can be described by Markov models. While these models are defined a priori, their parameters must be estimated from the data being modelled. The MRF describes the relations of attributes between sets of points in the neighbourhood. These sets are known as cliques, which are clusters of points that are all neighbours of each other (a neighbour is defined in respect of an 8-connection). A clique can consist of 1, 2, 3 or 4 points. Probabilities of patterns in the neighbourhood are calculated from all possible clique potentials. A clique potential is calculated with respect to some criterion, such as homogeneity. A low sum of all clique potentials gives a higher probability. The MRF is tuned from an initial state of prior probabilities. The posterior probabilities are often found using Monte Carlo methods [110], and the combination is referred to as Markov Chain Monte Carlo (MCMC) [111]. The number of relations in the neighbourhood (known as interactions) and the number of desired classes can be varied, but more interactions and classes result in longer processing time. Reversible Jump Markov Chain Monte Carlo (RJMCMC) is a variant of MCMC introduced by Green [112], which allows for parameter spaces of different sizes to be considered within the same model. By jumping between different parameter subspaces it is possible to estimate the number of classes in a fully unsupervised manner. This number would otherwise have to be fixed in traditional MCMC approaches and could require a subsequent split-and-merge operation.

The Expectation Maximization (EM) algorithm [113] is another algorithm often used for parameter estimation in MRF models. It attempts to iteratively recover the correct

parameters from incomplete data. The data realising an MRF is considered to consist of an observed and a missing part [109]. The algorithm iterates between modifying the data from the current parameter estimate and modifying the parameter estimate from the current data until convergence is reached. This is particularly useful for data containing artefacts. RJMCMC or MRF combined with EM are generally extremely demanding methods in terms of processing time, but arguably produce the most accurate unsupervised segmentation possible today. Examples of RJMCMC segmentation of natural colour images are given in chapter 5, while examples of MRF combined with EM for noisy, inhomogeneous MRI data are given in chapter 6.

Because neighbourhoods are analysed in terms of interactions between smaller components of the neighbourhood, the local neighbourhood representation has some adaptability, which is an advantage near boundaries. However, neighbourhoods must be isotropic (symmetrical) and furthermore a neighbourhood system (see fig. 2.11) and interaction function must be chosen to control the desired level of homogeneity and allow for specific types of discontinuity (such as edges). The best choices for maximum performance cannot easily be established without time consuming experimentation.

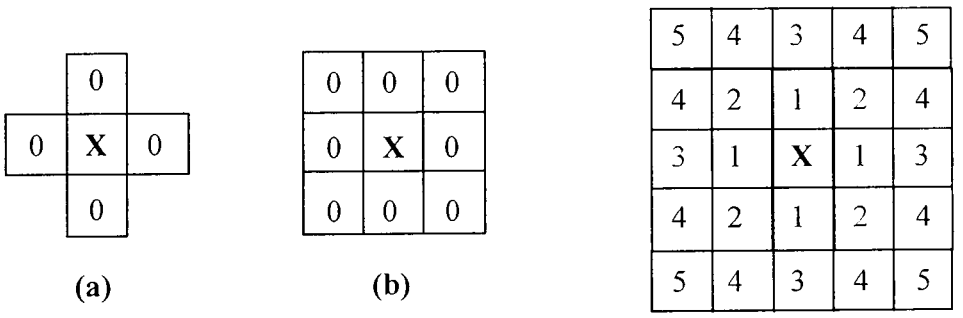


Fig. 2.11. Neighbourhood systems in Markov Random Field models. (a) First order neighbourhood (4-neighbourhood) where 0's are neighbours of X. (b) Second order neighbourhood (8-neighbourhood) where 0's are neighbours of X. (c) *n*-th order neighbourhood for *n* = 1,...,5 showing the outermost neighbours of X (from [109]).

Neural networks are a special type of statistical pattern recognition tools, which employ the idea of artificial neurons as layers of nodes on an interconnected network [114]. In a feed forward network, activation in the input layer represents the input

data, for example pixel intensities. Activation feeds forward through connections with associated weights, which determine the contribution of activation in connected nodes. Activation above a certain threshold may be required for a node to fire. The pattern of activation in the output layer gives the classification of the input data. Connections and weights may be hard-wired to solve specific pattern recognition problems. However for most applications this is not practical, and a learning rule is used to modify the network based on a comparison between the observed and the desired output. Backpropagation networks (also known as multi-layer perceptrons) have an input layer, one or more hidden layers and an output layer. The weights between nodes in the layers are adjusted by comparing an input pattern to the desired output pattern. Using a learning rule (traditionally the delta rule [114] in the case of backpropagation), the hidden layers are modified and a learning effect occurs. When no further changes result from this process, the network is said to have reached convergence. Given the correct network topology, any pattern recognition problem can in theory be solved by a backpropagation network [115]. However depending on the data used, the network topology (number of nodes, number of layers, interconnectivity, etc.) and the learning rule, such neural networks may never reach convergence. A typical problem is when a neural network gets stuck in local minima and appears to have converged, when in fact the global minima has not yet been reached. There are several methods for dealing with this type of problem (such as simulated annealing [116]) and there is a vast number of different network architectures, which will not be covered here. In image segmentation the most widely used neural network architectures are backpropagation networks and the vector quantization networks developed by Kohonen [117] (described at the end of this section). See [118] for a comprehensive review of neural network methods in medical image analysis.

Fuzzy logic [119] is an extension of set theory and models the degree to which memberships of sets are true. It is thus not a theory of *probability* but of *possibility*. In image segmentation it is useful for accommodating the ambiguity between segment classes, and for example determining the true class of a point near a boundary or a point affected by partial volume artefacts. Fuzzy logic was not used directly in the work described in this thesis, but was used to create the BrainWeb MRI data [41] and

in the AFCM segmentation algorithm [120], both described in chapter 6. For an example of a simple fuzzy logic application, please see appendix B, section B.3.

In pattern recognition applied to image segmentation and classification problems, the feature space is often large and complex. The correlations between low level features and higher level classifications may not be well known and even given a priori knowledge of relations between features, these relations may prove to be insignificant. In statistical pattern recognition the aim is generally to approximate the probability density function, i.e. finding the probability of a continuous random variable taking on a value in a specific interval (representing a feature) in a particular domain. This is demonstrated in vector quantization [117], where large vector sets are quantised to a smaller data set of "average" vectors representing the essence of the input data and approximating the probability density function for the represented classes.

The Kalman filter [121] also approximates the probability density function based on previous observations. It takes a large number of samples and their probability density functions (conditioned by the confidence of the particular sample) and combines them into one, approximating the true probability density function. Correlations between many features may be part of this process even if only a subset of these features are used for subsequent processing of new data. The Kalman filter is often used in the segmentation of image sequences in combination with other algorithms to predict properties of the next image from the previous images, tracking features in temporal data (for example combined with the LoG operator and Hough transform in [122]).

Principal Component Analysis (PCA) [123] (also known as the Karhunen Loeve Transform or KLT) is a method, which reduces the dimensionality of vectorial input data and finds its principal components. The co-variance matrix of the input data is formed by the variances and co-variances of all variables in the input data. From the covariance matrix the eigenvalues can be computed using a standard equation. The eigenvectors are found by solving the equation which expresses the eigenvalues as being the variance of the input data from the image matrix projected unto the eigenvectors. Each eigenvector can be computed separately for its corresponding eigenvalue. The eigenvector matrix (transformation matrix) can now be formed by

using the eigenvectors as rows, ranked according to their eigenvalues. Normally only the highest ranked eigenvectors are used, since they contain most of the information necessary to reconstruct the original image [60]. The eigenmatrix can be used to transform the original image data into so-called eigen images. The eigenvectors embody principal features of the input data, which may be used for compression and reconstruction or classification. PCA has been used extensively for recognition of 3D objects, including face recognition [124,125]. Objects are sampled at different orientations and PCA is applied to find the principal features invariant of orientation, such that the object may be recognised in novel orientations. PCA is less applicable to segmentation problems where the structure of a segment class may change significantly across the input data space (for example textures in a sequence of 2D slices through a 3D object of anatomy).

Learning Vector Quantization (LVQ) proposed by Kohonen [117] is a type of neural network, which performs vector quantization and produces a mapping from a higher to a lower dimensional space. The term "Kohonen network" does not generally refer to this particular type, but to the Self-Organizing Map (SOM) [117]. LVQ and SOM are very similar models, the main difference being that LVQ is a supervised approach, while the SOM is unsupervised. They are both 1-layer networks, i.e. the input and output layer is the same. LVQ takes a multidimensional feature vector set and generates a number of codebook (reference) vectors. These represent the best centroids in the input data space. The training data is labelled such that codebook vectors become labelled too. Learning is achieved through a nearest- neighbour rule using a distance metric, usually Euclidean distance (although other metrics such as the city block metric may also be used). Please see appendix B, section B.1 for a more detailed description of the SOM and LVQ algorithms used in this project.

2.2 Segmentation in medical imaging.

The literature review in section 2.1 has described a variety of general algorithms and tools for image segmentation, all of which have been applied to medical images. Most research has concentrated on X-ray, MRI, CT and ultrasound and there is a growing

interest in other modalities of medical imaging scans, such as PET [126] and SPECT [127]. Less work has been done in developing methods for the segmentation of cryo section data. Grey scale medical imaging data is subject to the problems of noise, inhomogeneity and partial volume artefacts. Colour cryo section data can offer a cleaner acquisition process with less artefacts and more information per point, but at the same time the mapping between intensity levels and segment classes is less simple than for example for CT and MRI. All the algorithms presented in section 2.1 have got drawbacks either in terms of accuracy, robustness, boundary location or the balance between accuracy and the level of automation. In summary the segmentation of medical imaging data is a hard problem, which requires highly robust and at the same time flexible solutions, and this problem has been actively researched and produced thousands of publications over the last three decades.

Although the whole area of machine vision has seen a shift in recent years from the application of classical image processing operations in new hybrid methods towards what is described as *model based vision* [128], this has been particularly the case for medical image processing. Conceptually model based vision is the idea of developing precise image models, which can be interpreted within (usually Bayesian) statistical frameworks. Classification is based on fitting novel data to prior knowledge about whole images, regions or segment classes. In medical imaging this involves models of grey level distributions for various modalities, modelling of noise and inhomogeneity and usually the assumption that most distributions can be described by Gaussian functions. At the same time physical models known as phantoms are used extensively in place of real human anatomy for the evaluation of medical image processing techniques. Wires or tubes are used to represent blood vessels while carefully moulded silicone models represent real skulls.

Model based vision allows for the systematic development of methods to handle all possible aspects of an image processing problem in a complete and mathematically tractable way. Since many models can be used to recognise as well as synthesise data, it also makes the simulation of clinical data sets possible, providing a source for systematic evaluation. Unfortunately the assumptions made by these models do not always hold true when methods are applied to real data. There may be other sources

of noise and inhomogeneity than what the model includes and this can prevent repeatability of results on real data (such as the model for noise reduction in [129]). Simulated data and quantitative evaluation is good for development and testing, but the ultimate test of a medical image processing system should be based on real data and evaluated by real people. In all other application areas than simulation, certainly in the area of segmentation, it is important that this ultimate goal is kept in mind from the beginning, rather than the goal of simply creating an attractive model.

Chapter 3

Requirements and methodology

3.1. Overview.

This chapter examines the challenges associated with medical image segmentation. Problems are identified and discussed, and solutions are proposed as a set of conceptual and technical requirements. They provide the foundation for the development of a new segmentation framework in the following chapters. A methodology for the development of this framework is outlined at the end of the chapter.

3.2. Identification of key problems in medical image segmentation.

Medical image segmentation shares a number of common hard problems with any other type of image segmentation and adds to that specific problems related to medical applications and medical data sets. The need for better automatic or semi-automatic segmentation of medical images has grown stronger in recent years, due to higher volumes of data requiring processing and higher quality of imaging across modalities, creating more application areas. Most applications in medical image segmentation require high levels of accuracy, making manual segmentation particularly time consuming. The common hard problems which affect segmentation accuracy were mentioned in chapter 2 and can be summarised as:

- The difficulty of distinguishing between edges within segments and segment boundaries
- The inability for any segmentation method assuming homogeneous regions to successfully deal with composite structures in data sets where intensities do not

directly map to tissue types (such as cryo section data) without a post-processing step for region merging

- The problem of incorrect boundary location
- The often unintuitive nature of optimal parameter estimation in statistical models for the target user

These problems must be solved for medical imaging data corrupted by artefacts of noise, inhomogeneity and partial volume artefacts. A problem in this respect is that different types of data, i.e. different modalities and sometimes even different acquisition modes of the same modality require:

- Different countermeasures to common artefacts
- Different feature detectors to accurately capture the intrinsic features of segment classes
- Different image encoding of detected features to achieve optimal classification

This results in vastly different implementations of the same algorithms for different image modalities. Consequently the user is often faced with different types of initialisation. Algorithms which incorporate countermeasures for specific artefacts (for example inhomogeneity [130,131]) rather than keeping the main segmentation algorithm generic and placing the countermeasures in an optional pre-processing stage, will lack flexibility in applications to new modalities. They may also impose too heavy a generalisation of artefacts, as noted by Likar [132]. Ideally the imaging modality should be transparent to the user when a specific segmentation system is employed.

The initialisation of a segmentation system should require a minimum of interaction from the user, while still adding enough a priori knowledge to the system to ensure accuracy. There is some dispute in the community about whether the way forward is fully automatic approaches, which may one day be as sophisticated as human observers, or whether it is in semi-automatic segmentation, where the user has some direct influence on the segmentation. As previously mentioned, establishing context is important in any machine vision task. However, quite clearly, establishing the nature

of the actual task and its ultimate goal is equally important. After all, *the level of success achieved by any segmentation method applied to data, in which the exact area and location of segments cannot be measured at the source, is inherently subjective*. It can only be measured against the desired goal of the user and in turn this goal should be the driving force of the segmentation. Fully automatic approaches are obviously suitable for applications where vast amounts of data must be processed continuously and not necessarily with the highest level of accuracy. In such applications the goal may be to find distinctly different image areas and to determine the presence of objects in a scene and their approximate location. Medical image segmentation is a very different type of application, where there is a much more specific goal for a specific task. Gerritsen et al [133] identified the following list of problems with fully automatic segmentation in medical imaging (quote from [133]):

> Limiting assumptions - The targeted high level of automation often compels a number of limiting assumptions, e.g. regarding allowable image sizes, image resolution, noise level, contrast extent and the meaning of contrast (imaging protocol). These limitations are not necessarily expected by the clinical user, nor are they necessarily acceptable!
>
> Assumptions often not made explicit - Worse, these and other underlying assumptions are often not made explicit, and the consequences of any unexpected violations of the assumptions are not understood sufficiently, nor are they being inventorized in any details.
>
> Inefficient - Currently, the efficiency of the software implementations of the operators which are crucial to modern segmentation leaves much to be desired indeed. One of the possible causes is that a maximal, albeit non-optimal, accuracy is being used (in several aspects: e.g. in a scale-space environment the data representation, the shape of the scaling operator, and the number of scales). In its turn this might be caused by a lack of understanding of the effects of using less-than-optimal accuracy. It should be noted that the mere use of faster computer hardware and/or multi-processors may make the software faster, but not more efficient. Better speed does improve usability. However, competition with other processing approaches will necessitate efficiency improvements.
>
> Slow, because full-sized images are being processed - The fact that processing is often aimed at the full image stack, with segmentation of more objects than needed, will slow down considerably. This may be improved considerably by using interactive focusing of the algorithms.
>
> Limited testing - Because of the lack of speed mentioned above, testing is often limited to a rather small set of images with a relatively small variation of parameters. Both the parameters determining the image's appearance and the ones which are steering the processing are most often not varied sufficiently, due to the lack of processing speed.
>
> 2D awaiting 3D - The slowness problem sketched above may also necessitate implementations to stay limited to 2D, with an inherently 3D version scheduled for better times in the future. This postpones gaining of insight into the problems arising from a generalisation of the dimensionality.
>
> Expert in the role of corrector - The role of the expert user is often reduced to that of a corrector of an automaton. It is desirable to put and keep the expert much more directly in charge, for defining the goals, for defining the way to get to the goals, and for steering along the way.

> Automation that really helps as advertised (e.g. power steering in a car) is appreciated by the user. On the other hand, users feel frustrated and slowed-down if they have to continuously correct "automatons" which do not perform sufficiently accurately in an independent way (e.g. an imperfect "auto-pilot" in a car).
>
> Vicious circle - Have we, perhaps, made ourselves victims of a vicious circle? Having slow software, we try to avoid directly interactive use by doing massive precalculations and asking the user to afterwards interactively edit the results of automatic processing, thereby preventing the user from intelligently focusing, helping, steering. In its turn, this makes the computational problems much larger or more involved. As a result, the software becomes slow, so that we...

Gerritsen et al emphasise on giving control to the user and also mention the problem of speed. It should be noted that if too much control is given to the user, then the automatic part of the segmentation process may be faster, but the total time taken to complete the task may not improve. The proportion of the total time in which the user is occupied will certainly increase. Achieving a good balance is key. Furthermore if the user is directly involved in the segmentation, such as the case is with active contours and live wire, then the credibility of performance measurements compared to more automated techniques becomes questionable. The ability to focus a segmentation on a ROI (Region Of Interest) is, as Gerritsen et al point out, very important in order to increase the speed of the process. It may be useful for the final visualisation of a segmentation to isolate the ROI before segmentation begins. For example the skull and its contents might be extracted prior to segmentation by detecting the cortical surface in a separate segmentation process (see chapter 6, section 6.4.1). However if the segmentation process would fail to produce good results if this type of prior feature extraction was not used, then there is an obvious problem involved in accurately achieving this feature extraction in a variety of different data sets. The main point is that a crude focussing such as selecting a ROI within a scalable box or cube should be sufficient, rather than having to exactly outline the region.

Olabarriaga and Smeulders produced a list similar to the one by Gerritsen et al in [134] and subsequently identified the following requirements for an Intelligent Interactive Segmentation system (IIS) for medical images (quote from [134]):

> Apart from being robust and predictable in terms of reliability, the following requirements are posed on (IIS) tools:
>
> 1. Data provided by the user should be used to derive model parameters and not directly as a literal part of the output. This guarantees that a uniform process generates segmentation results, with important implications for measurement continuity.

> 2. Segmentation results should be locally invariant to user intervention, being reproducible under predictable limits of variations. This is achieved by searching for a local optimum in the segmentation quality, which is a measured value derived from the image data, the segmentation result, a priori parameter settings, and parameter values derived from the interaction.
>
> 3. Interventions should have high semantics (small action = big impact into the desired direction). It requires the user to be able to predict the impact of her or his actions on the segmentation results, and it guarantees efficiency to the interaction.
>
> 4. User input should be generalizable to allow inference for neighboring or similar elements; that is, it should be possible to "learn" from user input. This requires parameters derived from user input to be used as a priori values in the next segmentation task, guaranteeing efficient interaction.

Olabarriaga and Smeulders underline the importance of not using the manual part of the segmentation process as a literal part of the output. The idea of using input data from the user to achieve learning in a classification system is accepted though. It could be argued that the training data should never be part of the data being segmented. There is however a difference between using an entire data set for training and simply using small user defined representative areas to learn the desired segment classes. With the variety in medical imaging data sets it would hardly be realistic to assume that training data from a few data sets would provide learning capable of producing segmentation of novel data as accurately as if the training data was derived from the data set being segmented. Small samples from a large volume might provide the necessary information about class relationships and artefacts in the volume, but this would most likely not transfer well to a different volume. Olabarriaga and Smeulders also make the point that a segmentation should be reproducible. The implication for semi-automatic systems is that initialisations by different users should produce similar segmentations of the same data. The exact definition of "similar" in this context is as subjective as the quality of segmentation and would depend on the specific application.

Addressing Gerritsen et al's point of "2D awaiting 3D", it should be added that processing speed is not necessarily the main problem. A greater implication could be for the accuracy of the segmentation. The segmentation of 2D slices reconstructed as a 3D segmentation is known as pseudo-3D segmentation, while segmentation using information in all three dimensions is known as isovolume segmentation [135]. Traditionally medical professionals have had to perform pseudo-3D segmentation in

their minds from looking at 2D medical imaging scans. This may be why the approach seems natural. However, given the technology to analyse data in three dimensions, this added information should clearly be used and not thrown away. All available *consistent* information should be used to support any type of automatic classification. This is particularly important for small structures (in a given resolution) or structures which are hard to segment in a 2D slice, but become much more distinct if their features are "followed into the volume". However it should be noted that in order to take advantage of information in neighbouring slices in a 3D volume, a number of constraints on the acquisition process, slice registration and alignment and slice thickness must be satisfied. These issues are discussed in chapter 4, section 4.8 and in chapter 6, section 6.4.1. Failing to do so could produce worse results than pseudo-3D segmentation. Consequently a careful decision based on the data at hand must be made about the extent to which data can be included in local representations to ensure consistency. The goal must be to maximise the utilisation of all available information, while minimising the level of interference caused to the original data in the process.

3.3. Addressing key problems: Conceptual requirements.

The proposed conceptual requirements for a segmentation system for medical imaging can be summarised as follows (addressing the points raised in section 3.2):

1. A segmentation task in medical imaging is goal dependent and subjective. Some a priori knowledge should be transferred from the expert user to the system, making the system semi-automatic.

2. At the same time, following the transfer of knowledge, the system should require a minimum of interaction with the user. Interaction should not be necessary to aid the segmentation if the goal can be explicitly stated first. Thus the actual segmentation stage should be as close to fully automatic as possible.

3. The segmentation algorithm employed should allow for focussing. The accuracy of segmentation should be maintained when applied to a ROI compared to the segmentation of a full image or volume.

4. Accuracy is of the highest importance. The system should minimise distortion of detail, maximise the accuracy of boundary location and be tolerant to common artefacts.

5. The system should adapt to patient specific data without imposing unwanted constraints from any type of generalisation.

6. All stages of the segmentation process in an n-dimensional domain should use information in n dimensions, given that suitable requirements for consistency of information are met and that a possible increase in processing time can be justified by the increase in segmentation accuracy.

7. Any processing needed to reduce artefacts in a specific type of data should be performed in a separate optional pre-processing step. This allows for the segmentation algorithm to be more generic and more easily adapted to a new modality.

A number of technical requirements follow from the conceptual requirements. These are given in the following section. While the conceptual requirements are general recommendations, the technical requirements should be regarded as guidelines for the *implementation* of the conceptual requirements.

3.4. Addressing key problems: Technical requirements.

For the preliminary stages of research a choice had to be made about an initial focus on one imaging modality. Colour cryo section images were chosen as the starting point in favour of greyscale medical imaging scans. Little work has been done in the

area, compared to the vast amounts of research that has concentrated on greyscale medical imaging scans. Initially the investigation also focused on 2D images for less complexity and easier testing. However, an important consideration was the ability to easily extend any developed techniques to greyscale and to 3D colour/greyscale volumes.

In order to facilitate a transfer of knowledge from the user to the segmentation system, the system should be able to learn. This involves the training of classifiers in a statistical framework. The learning should be based on training data, which can be selected in an intuitive way. Some approaches in the literature require the selection of appropriate filters, a long pipeline of image processing operations with manual intervention at several stages, or the selection of model parameters (the effects of which are often not clear to the target user). It seems a reasonable assumption that the most intuitive way of specifying the desired segment classes of an image for a human user, is to simply require the user to select representative areas of each segment class in one or more example images. This allows the user to visually define the goal of the desired segmentation. From the user's selection, what should be considered as segment boundaries and what should remain as edges within segments, is given. Boundaries do not have to be manually traced. This type of image template selection has been used in fast query systems (content based retrieval) for image databases, for example the work by Ratan and Grimson [136]. Conceptually it is a simple case of learning by example without complicated (and complicating) abstractions - simple for human beings, but complicated for an artificial system. However clearly the artificial system should adapt to the human user and not vice versa and in achieving this lies the research described in the remainder of this thesis. The first technical requirement can now be stated:

1. The detection and correlation of features in texture classes should be established automatically from user specified representations.

Due to the complexity of textures in natural (photographed) images, which are often irregular and inhomogeneous, it is extremely difficult to specify any rules, which will clearly define segment classes, without imposing too great a generalisation. Intensity

ranges or values in any colour space are not easily defined for arbitrary real textures. Even if they were defined, any definition would apply to a very limited type of images. In medical cryo section data the same segment class may contain textures and colours which differ widely depending on location. There are low level descriptors such as intensity levels, gradients and relations in neighbourhoods of pixels, but it cannot be made explicit exactly how these come together to yield a particular classification. Homogeneity cannot be assumed and explicit rules cannot be specified. The ability to exactly mathematically describe a texture to the point where it can be accurately reconstructed, is certainly not impossible for natural inhomogeneous textures, but requires a number of assumptions to be made about the nature of the inhomogeneity and how it is affected by external factors. As noted by Bowyer and Phillips [137]:

> Conceptual elegance and sophistication of the mathematics are not necessarily correlated in a positive way with performance of an algorithm in application. If the use of more sophisticated mathematics requires more specific assumptions about the application, and these assumptions are not satisfied by the application, performance could even degrade.

It is also a problem that statistical models are often evaluated on data, which has been synthesised using the same type of image model. Bowyer argues about this problem in [138]:

> The use of purely synthetic data is really a test of whether the implementation of the image analysis algorithm matches the assumptions of the model used to generate the synthetic data.

Fitting a segmentation system for the type of data described above into a strictly Bayesian framework could be limiting. The use of, for example, backpropagation neural networks or fuzzy logic requires a careful selection of network topology or rules. Successfully adapting these for different data sets could be overly complicated for the target user. A vector quantization neural network approach is immediately applicable with a minimum of required network parameter settings, but at the same time highly dependent on the data representation used. It allows for the creation of a feature space consisting of large sets of low level descriptors extracted from selected image areas. From this feature space, correlations within texture classes emerge, and given that convergence is reached, the system can be used for the segmentation and

classification of novel images. Rather than estimating a texture model, a set of average feature descriptors for each texture type is created. The approach itself does not guarantee that classes will be represented in a significant and mutually exclusive manner. It depends on the representational quality of the image encoding being used and the type of network architecture. The area of Artificial Life has seen the introduction of bottom up approaches for robotic systems as embodied situated agents. In Rodney Brooks' subsumption architecture simple interactions between a robot and its environment are used to learn more complicated behavioural patterns without explicitly specifying these [139]. It would be desirable to achieve the same kind of emergent properties [140] as classifications from low level descriptors in a machine vision system.

The second requirement is:

2. Vector quantization allows for a flexible classification system suitable for natural textures. Since such a system is only as strong as the data it is working on, a strong feature vector image encoding should be used.

A feature vector representation must be evaluated within the chosen classification system. A supervised learning technique is implied for the working segmentation system. However, an unsupervised approach can favourably be used for evaluation purposes. If good results are achieved using purely automatic clustering, the correlations extracted from the feature representations are strong. The third technical requirement is thus:

3. An unsupervised clustering technique should be used for evaluation of the feature vector developed according to technical requirement no. 2. A supervised learning technique will however be applied in the working segmentation system.

The problem of correct boundary location and segmentation near edges must be addressed. This is a problem affecting most traditional segmentation algorithms. It will affect a segmentation carried out within the segmentation framework proposed

above, for as long as image neighbourhoods are sampled in a traditional way (see chapter 4, section 4.3). The fourth requirement is:

4. The problem of segmentation near edges and correct boundary location must be addressed in depth. A solution within the given segmentation framework must be found.

3.5. Methodology.

Based on a literature review, a number of key problems in medical image segmentation have been identified. Conceptual and technical requirements have been established, forming the basis for three strands of work towards a robust framework for medical image segmentation with application to multiple modalities:

1. The theoretical development of a framework to satisfy the conceptual requirements:
   - 1.1. Identification of the role of the user and the user's interaction with the segmentation system
   - 1.2. Algorithm development and incorporation of existing algorithms for:
     - 1.2.1. Low level feature detection, encoding and matching
     - 1.2.2. Higher level classification
     - 1.2.3. Pre and post-segmentation processing
   - 1.3. Identification of suitable empirical evaluation techniques for specific types of data

2. The practical implementation of the developed framework adhering to the technical requirements:
   - 2.1. Implementation in software of the required algorithms
   - 2.2. Testing of implemented algorithms
   - 2.3. Generation of results based on suitable data sets

3. The empirical evaluation of the developed framework for multiple modalities:

    3.1. Comparative studies with other segmentation systems
    3.2. Quantitative evaluation on standard test sets using a standard performance metric
    3.3. Qualitative evaluation using human observer experiments and visual ranking

In the following chapters 4 to 6 the developments within the three strands are presented in parallel in approximately chronological order (based on the order of events as they occurred throughout the project). Theory is presented, followed by implementation and results and finally evaluation. Particular considerations for the type of empirical evaluation suitable for specific types of data are discussed in chapter 5. Chapter 7 is dedicated to a final qualitative evaluation of the proposed framework and thus only concerns the third strand of development as listed above.

Chapter 4

The ACSR framework

4.1. Towards a robust framework for medical image segmentation.

This chapter describes how the conceptual and technical requirements set out in chapter 3 were met for 2D colour images. The methods developed are later evolved for 3D volume segmentation of colour cryo section data from the Visible Human Project. Throughout the chapter a framework is built, describing required user interaction and a new concept for adaptable representation of segment classes and their segmentation. Algorithms are introduced to implement the framework and several segmentation pipelines are proposed. Preliminary evaluation is presented visually as segmentation results compared to previously published results on the same images.

4.2. Image encoding and classification.

The SOM architecture was chosen for unsupervised classification, because it is a vector quantization method with an excellent track record in many different areas of pattern recognition [117]. The SOM has been demonstrated as being capable of performing well on large data sets sampled from real life domains. In Kohonen's "Phonetic Typewriter" [141] a SOM was used to recognise speech from phonemes, deriving contextual information from neighbouring phonemes. In the WEBSOM project [142] thousands of newsgroup postings were automatically clustered according to topic based on low level descriptors (words). There has also been numerous studies of the SOM as an image segmentation/classification tool. Iivarinen et al [143] did a study of object recognition in which they showed that a SOM was capable of classifying irregular objects into categories depending on overall shape.

Five simple shape descriptors were used, but none of the categories could be derived from any single descriptor. Thus the classification was based on learning the patterns of descriptors relating to each class. Lawrence et al [144] used a SOM for the classification of face images, based on two different types of feature vectors. In one type each component represented a specific position in the neighbourhood (its point intensity). In the other type one component represented the intensity of the centre pixel and the other components each represented the intensity difference between a specific pixel in the neighbourhood and the centre pixel. Campbell et al [100] used a SOM to segment artificial colour images and natural outdoor colour images. Their feature vector was based on OPC descriptors for the centre pixel and the response of 16 manually selected oriented Gabor filters.

Since the project described in this thesis began, a group at Helsinki University of Technology have developed an image based equivalent of WEBSOM for content based image retrieval. This project is known as PicSOM [145]. At UMIST (University of Manchester Institute of Science and Technology) there is ongoing research in using the SOM as a compression tool for streaming video [146].

4.2.1 Developing a feature vector for image encoding.

The goal of this stage of development was to create a feature vector representation of the local neighbourhood of each pixel in an image. Example areas of images representing target texture classes (texture templates) should be encoded using this feature vector representation and used as training sets for SOM classifiers. The same encoding should subsequently be used on novel images for classification. Campbell et al [100] used oriented filters in their feature sets. This has the advantage of accurately describing oriented, regular textures. The disadvantage is that a filter bank must be created, based on the application. This is trivial within a restricted domain (such as the segmentation of Brodatz textures [147]), but problematic in larger, more complicated domains and requires extensive experimentation to achieve optimal results. Another problem with this approach is that the type of irregular textures found in cryo section images and medical imaging scans are the least suitable for oriented filters as feature

detectors. The representation of pixel neighbourhoods used by Lawrence et al [144] relied solely on pixel intensities, but it was oriented by associating relative position in the neighbourhood with a specific component in the feature vector. A representation for medical images should be invariant to rotation. A solution is to have a neighbourhood representation, which does not represent absolute spatial positions in the neighbourhood, while still maintaining a representation of the neighbourhood relative to the centre pixel.

The idea of Lawrence et al [144] of using intensity differences was employed, but in the form of Average Intensity Difference (AID) between the centre pixel and each of its neighbours. This is equivalent to the average distance using the city block metric. It is a simple measure of spread within a local neighbourhood, which as opposed to the standard deviation puts a dependency on the centre pixel, i.e. the point being represented. Any pixel in a neighbourhood, which is corrupted by noise, will have an effect on the standard deviation, which is less for the AID, as long as it is not the centre pixel being corrupted. This reduces the number of pixels heavily affected by noise in terms of their representation. If the centre pixel is representative of its class, then errors in the neighbourhood can be averaged down with AID. If the centre pixel is corrupted then the AID will give a more extreme deviation from the expected value than the standard deviation. It is then necessary to have additional descriptors to recover the correct classification.

Since a window (neighbourhood representation) may contain pixels from more than one texture class (see fig. 4.1), a centre pixel representative of its true class must strengthen the classification by being explicitly represented. Thus similar to Campbell et al [100] and Lawrence et al [144] descriptors for the centre pixel are included in the feature vector.

An orientation-independent representation of individual neighbourhood pixels is achieved by representing the total number of pixels in the neighbourhood falling within a number of uniform intervals (bins), covering the full range of a colour descriptor (depending on the colour model). All types of components are represented on a per descriptor basis. This means that *for each colour descriptor* of a chosen

colour model, there will be one component to represent the centre pixel and one block of components to represent the neighbourhood. Average intensity difference is tied to the RGB colour model regardless of the colour model used for the other components. This gives three components for average intensity difference, one for each colour channel.

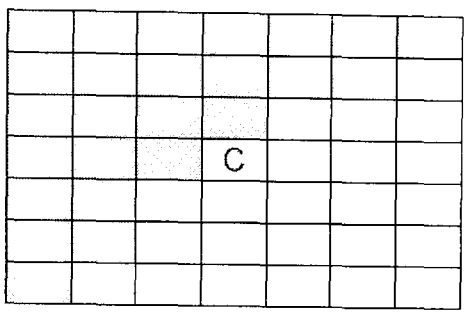


Fig. 4.1. An example of a neighbourhood representation in a 7*7 sampling window with centre pixel C. Window contains two classes (grey and white) but should represent only one (white).

The feature vector representation using OPC is summarised in fig. 4.2. Any colour model could be used with this representation though, depending on the application (see the discussion on colour models in chapter 2, section 2.1.6). This encoding scheme shall hereafter be referred to as the PixelDefine encoding.

In order to test the representational strength of the PixelDefine feature vector encoding, it was used for fully automatic image segmentation with a SOM classifier. The standard SOM_PAK software suite [148] developed by Kohonen and his team at Helsinki University of Technology was used for implementation. First a set of suitable parameters were chosen (number of nodes, dimensionality, lattice type of the network, neighbourhood function, learning rates and number of training cycles). The number of nodes was set to be small - close to the number of desired segments. This was to avoid the need for an extensive post-processing merging operation. The dimensions of the lattice were determined by using the shape of Sammon's mapping [149] as a guideline. Choice of lattice type (rectangular or hexagonal) and neighbourhood function (Gaussian or Bubble [117,148]) did not affect results in any observable way. Learning was carried out in two phases following the guidelines for additional parameter settings suggested by Kohonen as general recommendations in [117]. The first phase lasting 1,000 cycles used a radius close to the largest dimension

of the map and a learning rate parameter of 0.9. The second refinement phase lasting 10,000 cycles used a radius of 2 and a learning rate parameter of 0.02. Different levels of binning were applied to the neighbourhood descriptors in the feature vector and tested on a range of cryo section images to find the maximum level (minimum number of components) producing good segmentation results. This was determined to be 16 components each representing 32 intensity levels for R-G and Y-B and 16 intensity levels for luminance (as described in fig. 4.2). At this early stage evaluation of segmentation accuracy was achieved purely by visual inspection and through the comparison of average quantisation errors for the same data sets.

Fig. 4.3 shows an image from the colon cryo section volume mentioned in chapter 1 and its automatic segmentation using a varying number of nodes. Fig. 4.4 shows automatic segmentations of a cryo section from the Visible Human Project using a constant number of nodes but varying window size. It is evident that the detection of high and low frequency information is highly dependant on the window size. The results showed that even with the large degree of quantisation, with close correspondence between the number of codebook vectors and desired segments, meaningful segments were created fully automatically based on the PixelDefine encoding. The automatic clustering was facilitated by a representation, which was extracting sufficient information from the images.

Following this stage of initial testing of the PixelDefine encoding, the approach was changed to use multiple SOM classifiers. A SOM for each segment class of the colon volume was trained with a learning feature vector set created from representative areas of the gelatine (in which the tissue was frozen), the healthy tissue, tissue of the polyp and blood vessels. Appropriate window sizes for each tissue type were used. Fig. 4.5 shows the segmentation of the image in fig. 4.3(a) into the desired segment classes. It was achieved by calculating the quantisation errors for the feature vector of each pixel in the novel image with the codebook vectors of each individual, trained SOM. The SOM that yielded the lowest quantisation error (best fit) was selected as the winner. As expected this approach gave results much closer to the desired segmentation.

Colour descriptors:

RG: R-G+255
YB: (R+G)/2-B+255
L: (R+G+B)/3

Structure of a 54-dimensional feature vector representing an $(N+1)$-neighbourhood:

Centre pixel:

RG_c	YB_c	L_c
0	1	2

Neighbourhood pixels:

$\sum_n f_n(RG_n \mid 0 \le RG_n \le 31)$	$\sum_n f_n(RG_n \mid 32 \le RG_n \le 63)$	...	$\sum_n f_n(RG_n \mid 480 \le RG_n \le 511)$
3	4		18

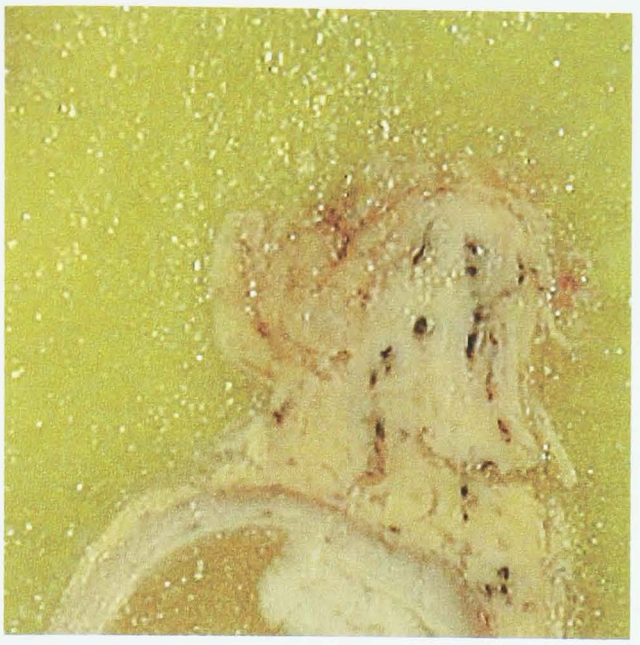
$\sum_n f_n(YB_n \mid 0 \le YB_n \le 31)$	$\sum_n f_n(YB_n \mid 32 \le YB_n \le 63)$	...	$\sum_n f_n(YB_n \mid 480 \le YB_n \le 511)$
19	20		34

$\sum_n f_n(L_n \mid 0 \le L_n \le 15)$	$\sum_n f_n(L_n \mid 16 \le L_n \le 31)$	...	$\sum_n f_n(L_n \mid 240 \le L_n \le 255)$
35	36		50

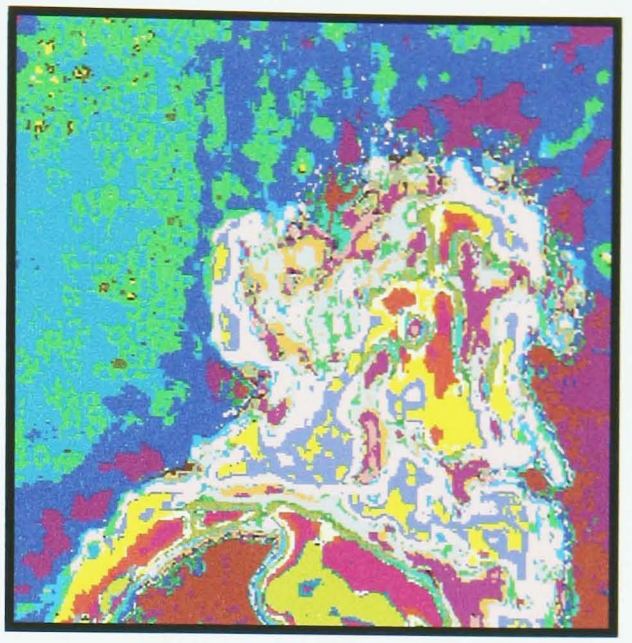
Average intensity differences:

$\left(\sum_n \lvert B_c - B_n \rvert\right)/N$	$\left(\sum_n \lvert G_c - G_n \rvert\right)/N$	$\left(\sum_n \lvert R_c - R_n \rvert\right)/N$
51	52	53

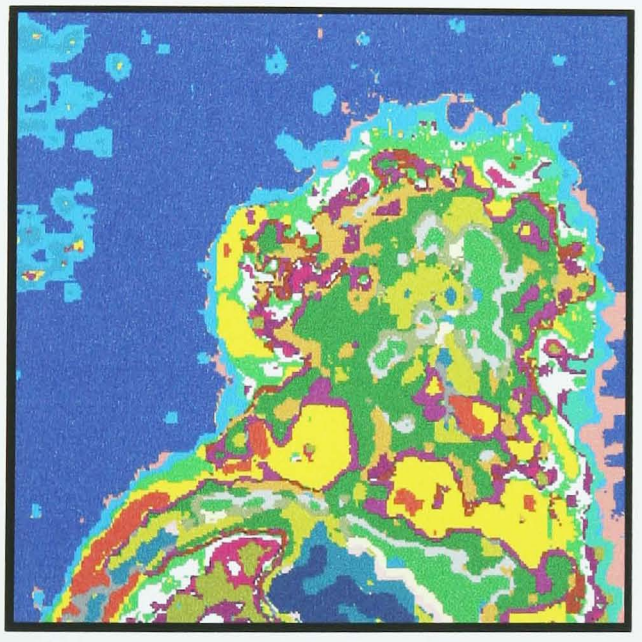
Fig. 4.2. The PixelDefine encoding.



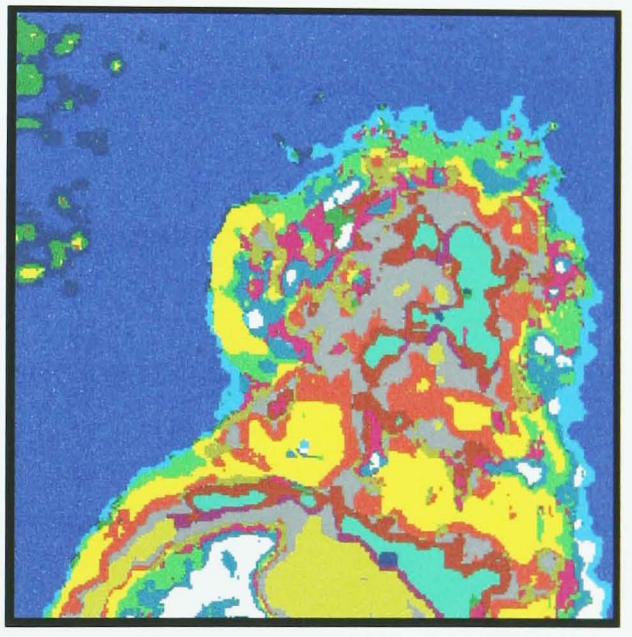
(a)



(b)

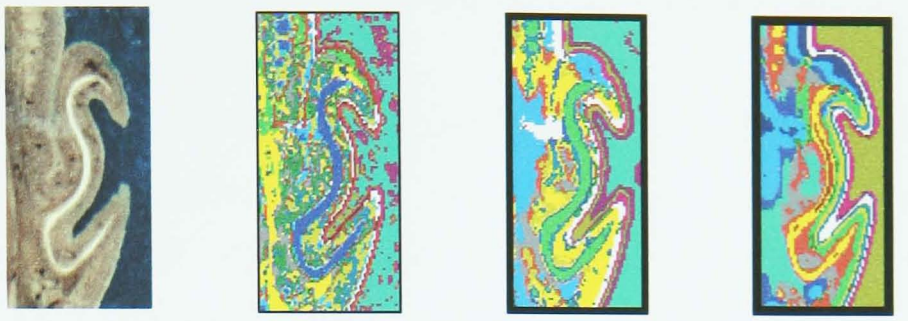


(c)



(d)

Fig. 4.3. Unsupervised SOM segmentation of cryo section based on PixelDefine encoding with varying number of nodes. (a) A cryo section (colon, polyp, gelatine). (b-d) Automatic SOM clustering of correlated image points using b: 31 nodes, c: 25 nodes and d: 16 nodes all with constant window size (9*9).



(a) (b) (c) (d)

Fig. 4.4. Unsupervised SOM segmentation of cryo section based on PixelDefine encoding with varying window size. (a) Cryo section from the Visible Human Project Visible Male data set (bone, muscles, cartilage, gelatine). (b) Window size 3*3: Good representation of high frequency information, poor for low frequency, minor artefacts near boundaries. (c) Window size 7*7: Some high frequency information lost, better for low frequency, larger artefacts near boundaries. (d) Window size 9*9: Poor representation of high frequency information, good for low frequency, serious artefacts near boundaries. Number of nodes constant at 16 for all images.

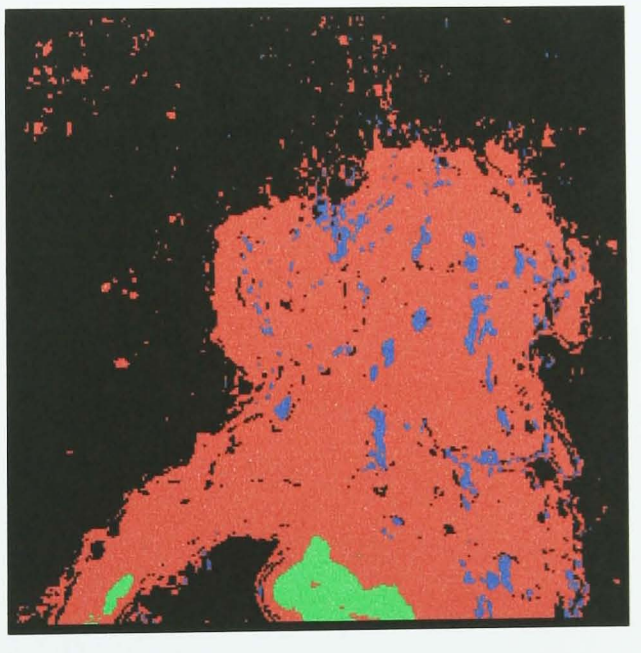


Fig. 4.5. Supervised SOM segmentation using multiple classifiers of source image in fig. 4.3(a). Segments classes are: gelatine (black), healthy tissue (red), polyp (green) and blood vessels (blue).

4.2.2. From unsupervised (SOM) to supervised (LVQ) learning.

According to technical requirement no. 2 a supervised learning approach should be used for the working segmentation system after developing the feature vector. It is clearly possible to continue using the SOM architecture in a supervised manner (demonstrated in fig. 4.5). However, there are a couple of important considerations:

- Although the network architecture of the basic SOM is considerably simpler than for most alternative neural networks, there are still too many network parameters which need to be set depending on the learning data sets. Because there is no definitive way of estimating these parameters, the SOM remains somewhat of a black box. Consistent results were achieved with a heuristic approach to parameter selection, but consistency cannot be guaranteed.

- Codebook vectors are not guaranteed to stay within their class regions if input samples overlap at the class boundaries. Consequently multiple individual classifiers were used to define class borders and simply used the SOMs to extract the best intrinsic features of each texture class. Comparing a novel feature vector with each SOM and finding the lowest quantisation error gave a winner. However, several different variables (such as quantity and quality of training data for each class) affect the relative bias of quantisation errors in each class. Quantisation errors for novel image feature vectors were calibrated according to the quantisation error of each SOM with its original learning data set. This is still not an ideal approach, which may result in misclassifications.

These two points are not problems with the SOM architecture. They only become problems in an attempt to use, what is essentially an unsupervised approach, in a supervised manner. Classes are required to be spatially ordered and some knowledge about the input data distribution and the nature of the data must be given in order to facilitate optimal training. LVQ, the supervised learning equivalent of the SOM, does not have these requirements. Because the LVQ is purely a pattern recognition tool, where classes are not required to be spatially ordered in the same way as in the SOM, fewer parameters are needed. Furthermore LVQ guarantees that codebook vectors

stay within their class regions, making an approach to multiple competing classifiers more feasible. LVQ is a natural choice for the supervised architecture, because of its close similarity with the SOM. Codebook vectors are even interchangeable between the two. As for the SOM, LVQ is well established in the machine vision literature (see e.g. [150,151]).

4.3. Addressing the problem of segmentation near edges.

One of the major sources of inaccuracies in image segmentation is incorrect boundary location. This extends in more general terms to incorrect classification of pixels near edges. Examples were given in chapter 1 of how segment boundaries may not be correctly located using a variety of algorithms. Because of the seriousness of this problem to medical applications and because of the multitude of traditional segmentation methods affected by the problem, this research project has devoted a substantial amount of time to addressing the particular problem of segmentation near edges. A solution to address this problem is presented here, incorporating the use of the type of LVQ classifier detailed in the previous sections and the PixelDefine encoding.

Filtering and region based analysis of sampling points for image segmentation traditionally rely on some form of rigid sampling window. This includes the convolution with kernels and region based representations. Generally high frequency information requires a small sized window, while low frequency information requires a larger size, to capture sets of representative spatial frequencies at a given point. Although tied to the spatial domain, the problem has many similarities to the problems mentioned in section 2.1.4, which sparked the development of the wavelet transform for frequency domain analysis. As explained in section 4.2.1, class-specific window sizes were used when sampling for SOM processing, based on prior knowledge about the textures to segment. A similar solution is to use dynamic adaptation of window size to local image areas (based on frequency information) as suggested by Xiong and Shafer [152]. This may produce better representations of local spatial frequencies (or clusters of point descriptors), but the ideal shape for a

particular texture class at a particular sampling point could still be incompatible with the constraints of a fixed shape window. Different shapes of the sampling window may be desirable for different texture classes at different sampling points. Increasing the size of the sampling window increases the problem of localisation in the spatial domain and reducing the size of the window reduces the amount of information available locally for classification.

Spatial filters with non-uniform kernels, such as the cone and pyramid filters [153] used for image sharpening, give maximum weight to the centre pixel, while the weight of pixels in the neighbourhood decreases with distance to the centre pixel (see fig. 4.6). Using the output of these filters directly to create point descriptors for a classification system or for a weighed neighbourhood sampling may give an improved neighbourhood representation, but it does not eliminate the problem of the rigid window or kernel. Ideally a different kernel should be used at each sampling point depending on the local information, but this brings back the classic dilemma mentioned in chapter 1: Segmentation begs classification and classification begs segmentation.

$$h(x,y) = \frac{1}{81} * \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 4 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 4 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix}$$

(a)

$$h(x,y) = \frac{1}{25} * \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 2 & 2 & 0 \\ 1 & 2 & 5 & 2 & 1 \\ 0 & 2 & 2 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

(b)

Fig. 4.6. Filters with non-uniform kernels giving maximum weight to the centre pixel. (a) The pyramid filter. (b) The cone filter.

For the training of neural networks and the selection of templates for template based methods, boundary pixels are usually avoided [154]. At least this provides representative templates, but problems occur when boundaries in novel images are encountered. In a study by Feng and Shaowei [154], edge points (found using multiscale edge detection and edge linking) were included in the training data to learn expected misrepresentations. Such an approach can be successful for very specific

applications, but lacks the flexibility needed to deal with anomalies, and easy adaptation to new application areas is difficult. Learning every possible type of edge in large data sets is rarely feasible and still makes the exact location of segment boundaries unpredictable due to the rigid window sampling.

4.4. Introducing the ACSR framework.

In traditional image sampling, one sample must represent every possible segment class as accurately as possible. Alternatively, when sampling at multiple scales, constraints can be imposed to increase the probability of a local sample being representative of a specific class, but with varying amounts of information available for classification at various scales. Rather than trying to accommodate all segment classes in a single sample or a scale space of samples, the aim should be to accommodate each segment class individually, creating a topologically different representation for each class for every sampling point with a constant amount of information available for classification. These representations can compete for the winning classification or they may be used for further processing by a higher-level statistical classification system.

Adaptable Class-Specific Representation (ACSR), a segmentation framework for accurate class-specific representation of point neighbourhoods, is introduced.

In Adaptable Class-Specific Representation every sampling point has a unique representation based on neighbouring pixels in n dimensions for every texture class. Every point is considered as a potential candidate for all of the desired segment classes. A representation is created for every class, in each case under the assumption that this class is the correct classification. No bias is present, since all classes are considered individually and independently of each other. Representations are also created independently of already classified points in the neighbourhood (i.e. such a classification is unknown to the system in the next step once it has been made in the previous step). In the same way every point, which forms part of a representation, will itself be represented separately. A local representation for each class is built as an

adaptive sampling window based on the matching of low level descriptors. The representations approximate the optimal representation for each class. This means that regardless of the evaluation criteria or classification system employed, all segment classes are consistently "competing on equal terms" for the final classification of a sampling point. This is in sharp contrast to traditional single representations where a rigid sampling window may produce a representation, which is highly stable for texture class A and highly unstable for texture class B, even if B is the correct classification. Representations are *not* segments. If they were, then the exercise would be pointless, as prior classification would be assumed. Prior classification would assume a classification of the neighbourhood and this would assume a rigid sampling window. Rather than that, low level point descriptors facilitate the construction of representations for each segment class, and based on a comparison of these, a classification is made on a point per point basis. Thus the connectivity of classified points into segments is not explicitly implemented. Instead, segments and boundaries become emergent properties of a dynamic process.

4.5. Introducing the Path Growing Algorithm.

The Path Growing Algorithm (PGA) implementing ACSR is considered here in the first case for 2D colour images. The low level descriptors used are colour descriptors based on a suitable colour model for a given application. For every segment class, a representative texture fragment must be given. This is provided by the user by simply marking out one or more areas of an image with a selection tool and giving it a label. Generalising to artificial (perfect) homogeneous, regular textures, a representative fragment is one, which reflects the spatial frequency of a texture class and the colour space it inhabits. For example an ideal fragment of a texture produced by a single sine wave grating with a single colour offset should cover a minimum of one full cycle. The PGA does not use the spatial frequency information, but this is a means of representing individual point descriptors across the full variation of a texture. Of course in real images most textures are highly irregular and often inhomogeneous. The task for the user then is to maximise the representation of variations across the full texture in the selected fragment(s). The texture fragments are used to create

templates, which are sets of unique combinations of colour descriptors for pixels in a texture. The texture fragments can also be encoded for the training of higher level classifiers, which may process samples created by path growing in novel images.

Relations between neighbourhood pixels are exploited in the co-occurrence matrix, but only selected pairs of intensities in the neighbourhood are considered. This idea could be extended to the co-occurrence of neighbours belonging to the same class, according to a template. Multiple distances and orientations may be used for the co-occurrences. However, the co-occurrence matrix does not provide an underlying structure for a fully connected neighbourhood and the spatial relations are hard-wired for a particular image and do not change dynamically.

If all possible combinations of a centre pixel and its neighbours within a k neighbourhood in 2D were to be considered, the total number of possible representations of the neighbourhood (assuming that at least one neighbourhood pixel be represented along with the centre pixel) would be 2^{k-1}. A 25-neighbourhood would thus give 16.7 million combinations. If this was extended to a 25*5-neighbourhood in 3D, the equivalent number would be $2*10^{37}$. This number would obviously be smaller if connectivity between all represented pixels was assumed, but still computationally unrealistic.

In order to reduce the number of combinations that must be considered, the PGA builds a larger sampling window from smaller components. These components are themselves built from single points. Some constraints on the possible topology of the sampling window apply, but the algorithm allows for a large degree of plasticity of the window shape.

The basic components of the PGA can be described in terms of graph theory as *paths*. Swami and Thulasiraman give a clear definition of a path (quote from [155]):

> A walk in a graph G=(V,E) is a finite altering sequence of vertices and edges v_0, e_1, v_1, e_2, ... v_{k-1}, e_k, v_k beginning and ending with vertices such that v_{i-1} and v_i are the end vertices of the edge e_i, $1 \leq i \leq k$. Alternatively a walk can be considered as a finite sequence of vertices v_0, v_1, v_2, ..., v_k, such that (v_{i-1}, v_i), $1 \leq i \leq k$, is an edge in a the graph G. This walk is usually called a v_0 - v_k walk with v_0 and v_k referred to as the end or terminal vertices of this walk. All other vertices are internal vertices of this walk. Note that

> in a walk, edges and vertices can appear more than once.
>
> A walk is open if its end vertices are distinct; otherwise it is closed.
>
> ...A walk is a trail if all its edges are distinct. A trail is open if its end vertices are distinct; otherwise it is closed...
>
> ...An open trail is a path if all its vertices are distinct.

In the PGA paths are grown from a seed point up to a pre-defined path length M (not including the seed point). A local neighbourhood of pixels (or voxels) may be described as points on an n-dimensional structured grid in a Cartesian coordinate system, where the distance between all neighbouring points is 1. The seed point may be considered as the origin, i.e. the centre of the grid. The path P uses points from the grid as vertices and the seed point is always one of the two end-vertices. Paths are thus grown from the origin, including one new vertex at a time. Two vertices are adjacent if the distance between them is 1. The growth uses a $2n$-connected expansion. This means that all reachable points form a diamond shape around the seed point (fig. 4.7). Paths are ranked in a hierarchical fashion according to their match with a template. This is done for all segment classes. The sampling window is built from a core best path and a number of other closest-match paths until a pre-defined number of points is reached. The path length and number of points per window can be changed depending on resolution and scale, but different path lengths for different classes may be employed if appropriate.

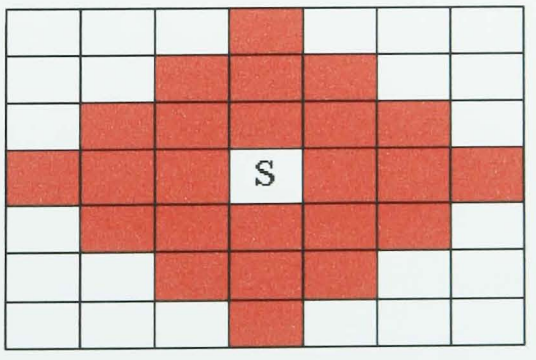


Fig. 4.7. Points reachable at a given path length form a diamond shape around the seed point S in the PGA. In this 2D example the path length is set to 3. Any cluster built using a 4-connected expansion within the diamond containing the seed point can constitute the sampling window for a class.

The PGA may be considered a local region growing algorithm, since points are added to the local neighbourhood representation according to a user-defined criteria. There is however an important difference between the PGA and traditional region growing, in that matching is done at the *region level* rather than the *point level*. An immediate neighbour of the seed point, which itself gives a non-optimal match, may be included, if it provides access to other points (in a connected chain), which gives the overall best match to that particular path. If changes in descriptors throughout a path are considered as state transitions, the next state depends not only on previous states but also on all the potential next states. It must again be stressed that the aim is not to model texture, but merely recognise features, which are known properties of a particular segment class. As opposed to region growing, every image point successively becomes the seed point and a new set of class-specific representations are found. Thus the algorithm does not explicitly facilitate connectivity of points at the *image level* into regions or segments. The following four sections explain the steps of the PGA from single seed points to multiple ACSR sampling windows in 2D colour images.

4.5.1. Single-pixel template matching.

All sets of colour descriptors in every template in turn are compared to the sets of colour descriptors for reachable pixels, including the seed point PX_s. The match value PXM_{ki} for pixel PX_k with template T_i, using j descriptors:

$$PXM_{ki} = \min_{f=1\ldots f_{\max}} \left[\sum_j \left| TXD_{ifj} - PXD_{kj} \right| \right] \qquad (4.1)$$

where TXD_{ifj} is descriptor j for pixel f out of f_{max} pixels in template T_i. PXD_{kj} is descriptor j for pixel PX_k. $PX_k \in K$, where K is the set of all pixels including and reachable from seed point PX_s, using a 4-connected expansion with a length of M pixels.

The Sum of Absolute Distances (SAD) is used in favour of the Sum of Squared Distances (SSD). The goal at this step is to find the best representation of each single class, by calculating the distance to a single template, and not to find the winning

class by calculating the distance to all templates and comparing them. Taking into account that image artefacts may corrupt some descriptors and leave others unaffected (see section 4.2.1), it is beneficial to reward representations, which have a very close match with some descriptors, even if others are far off. The SAD does this, while the SSD penalises such representations, compared to representations where the error is evenly distributed across descriptors.

4.5.2. Path Growing from a seed point.

Two values are calculated at this stage: the path value for every path representing every template and the total path spread for every path. For the path P_l starting from seed point PX_s, the path value PV_{il} for template T_i is:

$$PV_{il} = \left(\sum_{r=1}^{M} PXM_{ri} \right) + PXM_{si} \qquad \textbf{(4.2)}$$

The path P_l is a set of pixels $P_l = (PX_0 ,...,PX_M)$, where $PX_0 = PX_s$ and $P_l \subset K$. Paths with less than $M+1$ elements are illegal due to overlap with themselves.

The total path spread is defined as:

$$\sigma_l = \sum_j \sqrt{\frac{\left(PXD_{sj} - \mu_{lj}\right)^2 + \sum_{r=1}^{M} \left(PXD_{rj} - \mu_{lj}\right)^2}{M+1}} \qquad \textbf{(4.3)}$$

$$\mu_{lj} = \frac{PXD_{sj} + \sum_{r=1}^{M} PXD_{rj}}{M+1}$$

4.5.3. Ranking the paths.

A solution hierarchy is created by ranking the paths for each T_i, so that $P_l \prec P_{l'}$:

(4.4)

if $PV_{il} < PV_{il'}$ or

if ($PV_{il} = PV_{il'}$ and $\sigma_l > \sigma_{l'}$)

Thus the best path at the top of the hierarchy has the highest possible spread for a path with the lowest possible path value for a fixed T_i. A hierarchy exists for every template, for every seed point. In textures with similar local areas and different surrounding areas, maximising the path spread at equal path values provokes a migration from the similar to the different areas. The combined representation of both areas becomes distinct. Since it is harder to obtain a low match value at higher spread, these representations are favoured. Thus the cost of choosing a representation with low spread is higher than that associated with choosing one with higher spread at the same distance to different templates.

4.5.4. Building the sampling window.

The core of the sampling window for template T_i, representing a seed point, consists of every pixel from the path $PWIN_i$ which is at the top of the solution hierarchy. Additional novel pixels ($PX_k \notin PWIN_i$) are added by traversing the hierarchy from the top down until a pre-defined total is reached. The resulting region is the sampling window for template T_i. A region in this context is *not* a segment, it is a representation of the seed point. All other pixels in the region become seed points and are represented by their own unique regions for each T_i.

For a discussion of the complexity and computational overhead of the PGA, please see appendix C.

4.5.5. Classifying representations created by the PGA.

For each sampling point a classification may be given directly from the representations created by the PGA. It is done by selecting the class-specific representation with the highest path spread out of those representations with the lowest path value (similar to the way individual paths are ranked). It is also possible to add a higher level classification system on top of the PGA. In the study described in section 4.6 the PGA was used in combination with LVQ classifiers and the PixelDefine encoding. The study investigated if representations created by the PGA

could favourably replace traditional window sampling for a neural network method, such as LVQ.

From the texture fragments used to define the templates for the PGA, feature vectors are encoded using the PGA on fragments based on their own templates. This means that all encoded paths have the maximum possible path spread and allows sampling of arbitrarily shaped fragments. All vector components are localised and standardised by shifting the mean to zero and dividing by the standard deviation [156]. The learning vectors are used to train individual LVQ classifiers, one for each segment class, using the Optimized Learning Vector Quantization algorithm [117] for fast convergence. Class-specific feature vectors from novel images are encoding from the neighbourhood regions generated by the PGA. These are transformed using the same parameters for the means and standard deviations found in the training sets and the LVQ classifiers produce the final segmentation based on lowest quantisation error.

4.6. Preliminary results for 2D colour images using ACSR and LVQ.

The results in this section were originally presented in the first paper published on ACSR and the PGA [157]. An artificial colour image, a natural colour image and two cryo sections were segmented and compared to other segmentation methods using sampling windows of fixed shape and/or size.

ACSR segmentation was applied to an artificial test image from the study by Campbell et al [100], mentioned several times previously. In summary, the study used a SOM classifier working on a representation of the centre pixel and the response of 16 oriented Gabor filters at each point. For the ACSR segmentation, the HSV colour model was used to produce the low level descriptors for the PGA and the PixelDefine encoding. It was chosen in favour of OPC because the image was artificial with segments of uniform colour components (see chapter 2, section 2.1.6). Feature vectors using the 54-dimensional PixelDefine encoding were used to train the LVQ classifiers using the LVQ_PAK [158] implementation.

The artificial test image has a grey scale low frequency sine wave grating as background. Segments consist of a square of randomly coloured pixels with a uniform distribution, a solid cyan circle and two differently oriented sine wave gratings at higher frequency than the background, with a green colour offset. Gaussian white noise was added [100]. Since the PGA is orientation independent, a modification to the approach had to be made in order to correctly distinguish between the two different orientations of the grating texture. Orientation detection shall not be covered in great depth, since as previously mentioned, it is rarely a desirable feature for medical image segmentation. However, since orientation detection can be important in other application areas, the problem was addressed for the sake of completeness. Orientation was detected in a second sub-classification stage following the primary texture classification. The texture type was classified first and then its orientation. Two LVQ networks were trained, one with and one without orientation information. Orientation was encoded as the major local gradient directions, determined by the topology of the path yielding the highest intensity spread at each image point. As opposed to the path spread used for non-oriented classification (highest spread of all descriptors in a colour model of choice), this spread represented the summed spread of intensities in the red, green and blue channels of each vertex in a path (purely reflecting the spread of intensity). The width and height of each path's bounding box was encoded. For the two orientations used in the test image, orientation would depend on either the width or the height being greater than the other. This encoding was sufficient to automatically encode orientation from selected image fragments without any pre-selection of oriented filters. For the segmentation of the full image, a primary segmentation was first carried out based on the templates and non-oriented classifiers created from the selected texture class fragments. The template of the grating texture was then automatically replaced by a template based on the segments for that texture class found in the primary segmentation. In the second subsequent segmentation step intensity spread was calculated based on the new template (making all points in the segments a perfect match would force the selection of the winning paths to be based on spread only). Orientation encoding was added to the feature vectors created in the primary segmentation step and the oriented classifiers were used to produce the final segmentation. This segmentation pipeline is illustrated in fig. 4.8.

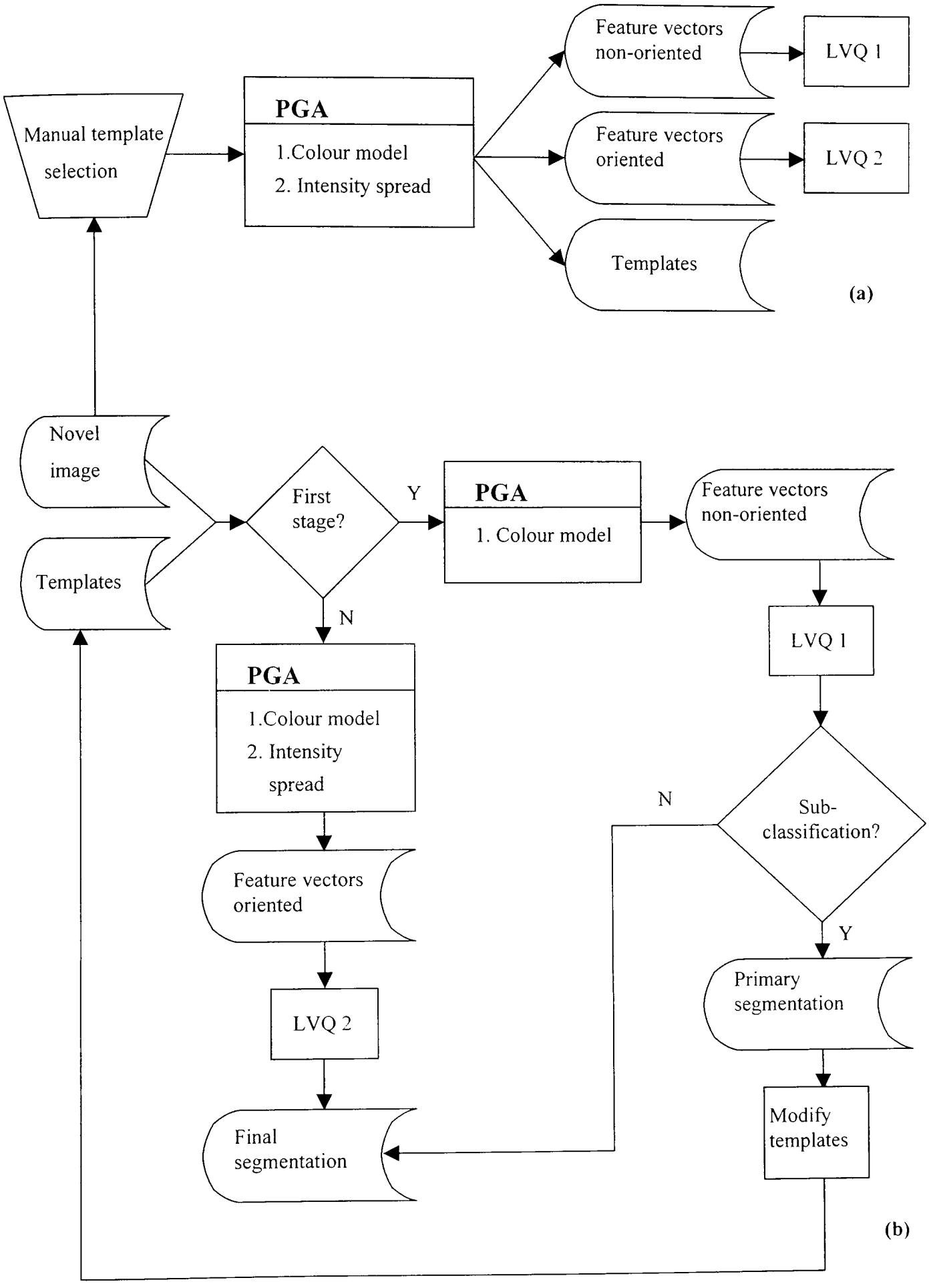
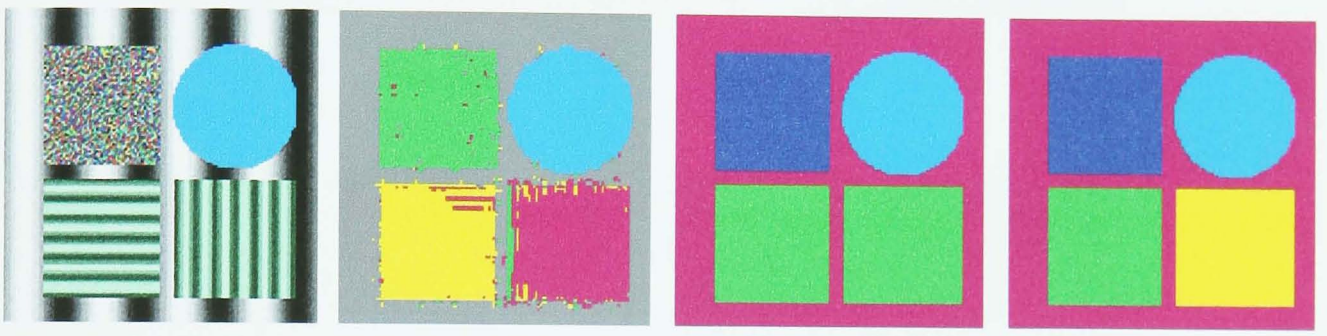


Fig. 4.8. The full ACSR pipeline as proposed in [157]. (a) Training of classifiers and template creation. (b) Segmentation of a novel image. Flow through non-oriented processes. If oriented sub-classification is required: Feedback to the templates, then through the oriented processes to the final segmentation.

Fig. 4.9(a) shows the original test image and its segmentation from [100] (fig. 4.9(b)). Although most pixels inside the segment boundaries have been correctly classified, strong artefacts near the boundaries are present. Fig. 4.9(c) shows the standard non-oriented segmentation using ACSR and the PGA with LVQ classifiers. Fig. 4.9(d) shows the segmentation with added orientation detection. The latter is 100% identical to the ground truth segmentation as specified by Campbell et al [100].

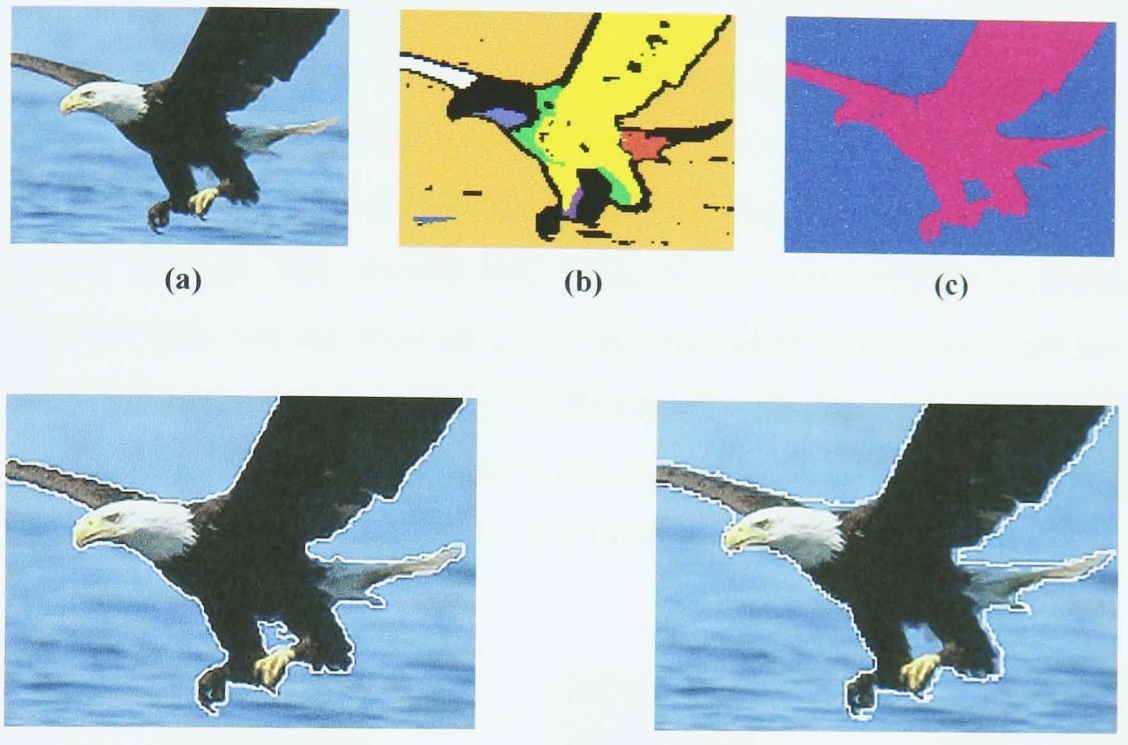


(a) (b) (c) (d)

Fig. 4.9. An artificial image and its segmentation. (a) The source image [100]. (b) Segmentation from [100]. (c) ACSR segmentation without orientation detection. (d) ACSR segmentation with orientation detection.

For the three real colour images, orientation detection was not applied. In favour of the HSV colour model, the OPC model was chosen (see section 4.2.1 and chapter 2, section 2.1.6). The natural colour image was also a previously published test image, used in a study by Williams and Alder [159]. It was a study of a CBIR (Content Based Image Retrieval) approach, using a split-and-merge segmentation, which yielded results that could be compared to ACSR segmentation. Fig. 4.10(a) shows an image of an eagle over water, originally taken from a Corel Photo-CD (reproduced under license). Fig. 4.10(b) shows the segmentation from [159]. Most edge pixels were marked as "unused" (i.e. unclassified) and appeared as black regions. It was noted in [159] about the edge artefacts in the segmentation that "this phenomena is desirable". This was due to the benefits of distinct regions for CBIR. It was however clearly still an artefact rather than a deliberate feature of the segmentation algorithm. Fig. 4.10(c) shows the ACSR segmentation of the original image using path growing and LVQ. Fig. 4.10(d) shows the contour of the eagle emerging from this segmentation. The contour appears as an optimal edge detection. The boundary pixels

in the segmentation presented in [159] include pixels from different segment classes. Based on the outer edge, "water" appears inside the contour for "eagle" in fig. 4.10(e) (notice areas above the left talon and near the tail and head).



(a) (b) (c)

(d) (e)

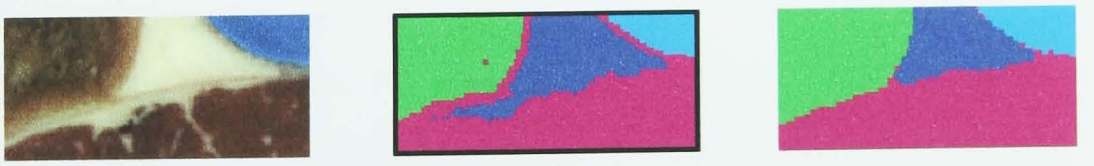
Fig. 4.10. Segmentation of eagle over water. (a) The source image. (b) Segmentation from [159]. (c) ACSR segmentation using PGA and LVQ. (d) Contour from (c) overlaid on the source image. (e) Contour from (b) overlaid on the source image.

While the first two test images were not medically related, their segmentation provided a benchmark test by comparison to previously published segmentation algorithms with fixed window shape. The two remaining test images published in [157] were single 2D cryo section images taken from the Visible Human Project. The ACSR segmentation was compared to a segmentation using the same training data for LVQ classifiers, but sampled with a best fixed size window (the window size producing the best overall segmentation). Fig. 4.11 shows a part of a cryo section and its segmentation. The segmentations of hard bone, bone marrow and muscle for the two approaches are almost identical. There are only two types of boundaries between the desired segments (muscle/hard bone and hard bone/marrow), and their colour components are dissimilar to any individual class. Fig. 4.12 shows another part of a cryo section from the Visible Human Project. The segmentation divides the image

into 4 distinct regions: a piece of the colon, fat, muscle including fascia and blue gelatine (in which the cadaver was frozen). Compared to fig. 4.11, this image has more boundaries and bordering segments are of greater similarity. Again the image was segmented using the best fixed-size sampling window and LVQ classifiers, as well as path growing and LVQ. In this case, the segmentation using the best fixed-size sampling window was clearly poorer than the path grown segmentation. There are strong edge artefacts, and a large part of the muscle fascia is confused with fatty tissue. The path grown segmentation has a few misclassified pixels, but the edges are clearly defined. The muscle and fascia have been correctly segmented into one segment (this was specified as a goal by templating both tissue types under the same label and across their boundary). Although the edge between muscle and fascia is very similar to the boundary between fascia and colon and between colon and fatty tissue, all boundaries visually appear to be precisely located.

(a) (b) (c)

Fig. 4.11. Segmentation of muscle (outer segment), hard bone (middle segment) and bone marrow (inner segment). (a) The source image. (b) Segmentation using best fixed-size sampling window and LVQ. (c) ACSR segmentation using PGA and LVQ.



(a) (b) (c)

Fig. 4.12. Segmentation of colon (upper left segment), fat (upper middle segment), blue gelatine (upper right segment) and muscle including fascia (lower segment). (a) The source image. (b) Segmentation using best fixed-size sampling window and LVQ. (c) ACSR segmentation using PGA and LVQ.

For the images presented in this study no comparison of segmented images with a ground truth segmentation was performed, except for the first image. This image was artificially created and thus boundaries can be exactly located. In natural images however, the exact boundary location is slightly fuzzy and a ground truth segmentation depends on the individual creating it. Visual inspection of boundaries and a visual comparison between different segmentations may therefore be as valid a result as a comparison with a ground truth image on a pixel per pixel basis. This view was particularly strongly expressed by Heath et al [160], who used human observer experiments to determine the quality of five different edge detection algorithms and avoided ground truth images. For a more detailed discussion about the evaluation of segmentation methods including ACSR, see chapter 5, 6 and 7.

The results show that the PGA is capable of producing adaptable regions, which can be encoded for LVQ classification, producing better results than what could be achieved with traditional rigid window sampling. This would be likely to extend to other classification systems traditionally relying on rigid window sampling. However it was found after the publication of the study that when classification was based on the PGA directly, the results produced were virtually identical to those produced by the LVQ classifiers. This finding was repeated on a large number of additional segmentations of cryo section images.

In conclusion to the results presented in this section, ACSR segmentation consistently produced better results compared to three different segmentation algorithms using a form of sampling window with fixed size and shape. The combination of ACSR and LVQ classifiers showed that superior results were achieved with plastic sampling windows, compared to the traditional use of fixed shape windows for LVQ. It also appeared that the PGA alone could produce similar results. This indicated that rather than just being a pre-processing tool for LVQ classification, the PGA could in fact replace the LVQ classification. However a reversed combination of the two is a better option as section 4.7 will explain.

4.7. Focusing ACSR.

One of the conceptual requirements stated in chapter 3 was the ability to focus a segmentation algorithm. Normally this means the selection of a ROI by the user. The PGA works on individual image points and can segment any isolated subsection of an image without significantly degrading the quality of segmentation compared to segmentation of the full image. Therefore the requirement is easily met for manually imposed focusing. This section describes an automatic focusing, which aims to produce the same results as the non-focused algorithm, but faster.

As mentioned in the previous section, experiments with ACSR using the PGA for colour images have shown that the core algorithm is capable of producing accurate results identical to those achieved with the added LVQ classification. It is however still highly processor intensive. LVQ classifiers using the PixelDefine encoding and a fixed size sampling window generally produce accurate results inside segment boundaries, but misclassifications near boundaries. The combination of these two observations is the background for partial ACSR.

Partial ACSR is an automatic focusing of the PGA on points near edges. It may be based on any template based method, where the templates can be used to drive the PGA. A point based nearest-neighbour classifier has been used for discrete 2D images and LVQ for image volumes (see sections 4.8, 4.9 and section 5.3 in chapter 5).

When using LVQ classifiers for partial ACSR, rather than applying them *after* the PGA to create the final segmentation, they are used *prior* to the ACSR segmentation for creating a faster, preliminary segmentation. The smallest possible window size (3*3 in 2D, 3*3*3 in 3D) is used to capture as much detail as possible. The small window size inevitably leads to oversegmentation (too many segments) and artefacts near edges. By applying the PGA in the neighbourhood of boundary points found using the preliminary LVQ segmentation, artefacts are eliminated and segments are merged to produce the final segmentation. Because boundary location in the initial segmentation will be inaccurate, it is not sufficient to simply apply the PGA exactly at boundary points. A dilation operation is used to grow the boundaries of the initial

segmentation by a user defined factor (determining the thickness in pixels on either side of the detected boundary). This will depend on how close the initial segmentation is to the desired segmentation. Assuming a representative template selection, this factor can be small. Values of 2 or 3 are sufficient. This dilated boundary map is subsequently used as a mask. The PGA is applied only at boundary points and the results replace these points in the initial segmentation. The remaining points are unchanged. Fig. 4.13 shows an example of partial ACSR applied to a cryo section brain image. The segment classes are "grey matter", "white matter" and "other". In the binary mask image the grey areas show where the initial LVQ segmentation remains, while the PGA is applied at all points within the black areas. Because of the boundaries taking up a large part of this image, the increase in processing speed is moderate. However in some types of images the initial segmentation may well constitute the largest part of the final segmentation.

Partial ACSR can provide a substantial reduction of processing overhead compared to full ACSR. The actual speed increase varies between images and depends on the dilation factor used and the number of original boundary points in the coarse segmentation. Good results have been achieved, showing identical segmentations to full ACSR (see chapter 5, section 5.3).

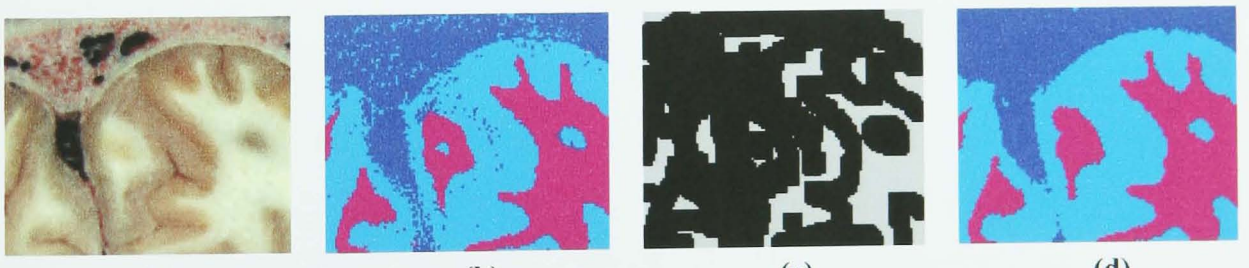


Fig. 4.13. Partial ACSR segmentation of cryo section brain slice. (a) Source Image. (b) LVQ segmentation. (c) Mask image showing dilated boundaries (dilation factor 3) from the LVQ segmentation. (d) Final ACSR segmentation composed of masked initial LVQ + PGA applied at dilated boundary points.

4.8 Isovolume and pseudo-3D segmentation.

It is trivial that the PGA is readily extendable to n dimensions. Similarly the PixelDefine encoding may be based on sampling of local neighbourhoods, which are

cubes of any number of dimensions (a square in 2D, a cube in 3D, a hypercube in 4D, etc.). When applied to 3D volume data, by using information in all three dimensions, true isovolume segmentation can be achieved within the ACSR framework. Volume segmentation means a considerable increase in processing time compared to 2D segmentation (depending on the dimensions of the volume and the path length used by the PGA). To address this problem, partial rather than full ACSR is applied. The only two differences from the 2D case are that the PGA uses a 6-connected rather than a 4-connected expansion and the LVQ classifiers are based on sampling with a 3*3*3 sampling cube rather than the 3*3 square sampling window in 2D.

Partial ACSR and LVQ classifiers have been used for isovolume segmentation of colour cryo section volumes from the Visible Human Project. This study was published in [161].

As mentioned in chapter 2 (section 2.2), isovolume segmentation clearly has the advantage over pseudo-3D segmentation of more information to support a classification at each point. Some approaches such as [162] use a volume segmentation, which relies on a prior segmentation step in 2D. This can obviously reduce the benefits of volume segmentation by excluding data, which would otherwise have been available to a volume segmentation applied to the original data. With 3D partial ACSR and LVQ, information in all three dimensions is used at all stages of the segmentation pipeline. However, to guarantee the benefits of isovolume segmentation of cryo section volumes, a number of assumptions about the data must be made. First of all it must be assumed that all slices are accurately aligned. Secondly it must be assumed that the image acquisition for every slice is done under the same conditions. This means that the slice plane (the photographed surface) is in fact always a plane, i.e. a smooth surface (with a constant tilt angle), that the camera position and direction is unchanged, and that the light sources used to illuminate the slice images are unchanged throughout the whole volume. The PGA and the use of opponent process colour descriptors allows for some changes in intensity across a volume. However, Chandler et al [163] have shown that any directional light source is effectively a directional filter. This means that if the position of a light source changes, illuminated textures change too. Such a change would affect the

correspondence between templates and local image areas. This could have serious effects on segmentation accuracy (although algorithms using oriented filters are more sensitive to this problem) unless changes to light sources are accommodated by an adaptive template scheme with separate template sets for separate illuminations (or possibly using principal component analysis of the same images under a variety of lighting conditions to learn the invariant texture features). Different tilt angles relative to the camera produce perspective effects, which also affect textures in a way, which could render the classification system unreliable, unless countermeasures are taken. Finally the slice thickness should be minimal to avoid large scale partial volume artefacts. What is acceptable depends on the structures being segmented and the resolution of the volume. In the case of the Visible Human Project data sets, accurate alignment of slices, smooth slice surfaces, constant illumination and unchanged perspectives are assumed. The slices are high resolution and for the segmentation of macroscopic anatomy the 1mm slice thickness is considered to be adequate, although an even smaller thickness would be an advantage.

4.9. Extending the ACSR framework to 3D isovolume segmentation.

To illustrate the benefits of using 3D volume information (given that the constraints mentioned in section 4.8 are satisfied), consider an artificial volume, which consists of equally sized yellow rectangles at different positions in the otherwise blue volume. The rectangles are spatially disjoint between slices, except for a chain, which stretches out through the volume, forming a T-shape. All the middle rectangles in the slices (see figs. 4.14(a), 4.14(b) and 4.14(c)) connect with rectangles going across in the middle slices. Consider then three templates, one for blue (background), one for yellow (the rectangles) and one which represents a combination of yellow and a very slightly darker blue then the background (by one unit in the blue channel). This could be a texture *containing* an edge type almost identical to a boundary between two other separate texture classes. ACSR is used to segment the volume using pseudo-3D segmentation (segmenting each 2D slice individually) and volume segmentation (sampling in all three dimensions). It is evident from figs. 4.14(d) and 4.14(e) showing isosurface models of the two segmentations that not all connected rectangles

are identified as the same, using pseudo-3D segmentation. Topologically they form the T-shape, but only connected points in the slice plane have been identified as belonging to the same segment class. The rest have been classified as the blue/yellow combination. However, in the volume segmentation the rectangles connected in a T-shape have been identified as a class of their own. This is because the PGA can follow the chain of voxels into the volume. This model embodies the problem of segmentation of small structures, such as blood vessels, where correct classification depends on the detection of connectivity in all three dimensions.

Fig. 4.15 shows the new volume segmentation pipeline for partial ACSR. A global ROI can be selected from the larger volume for segmentation. This global ROI is broken down into one local ROI after another with the voxel to be classified in the centre as the segmentation progresses. All the local ROIs intersecting a slice plane through the centre voxel produce the classifications needed for the full segmentation of that one slice plane.

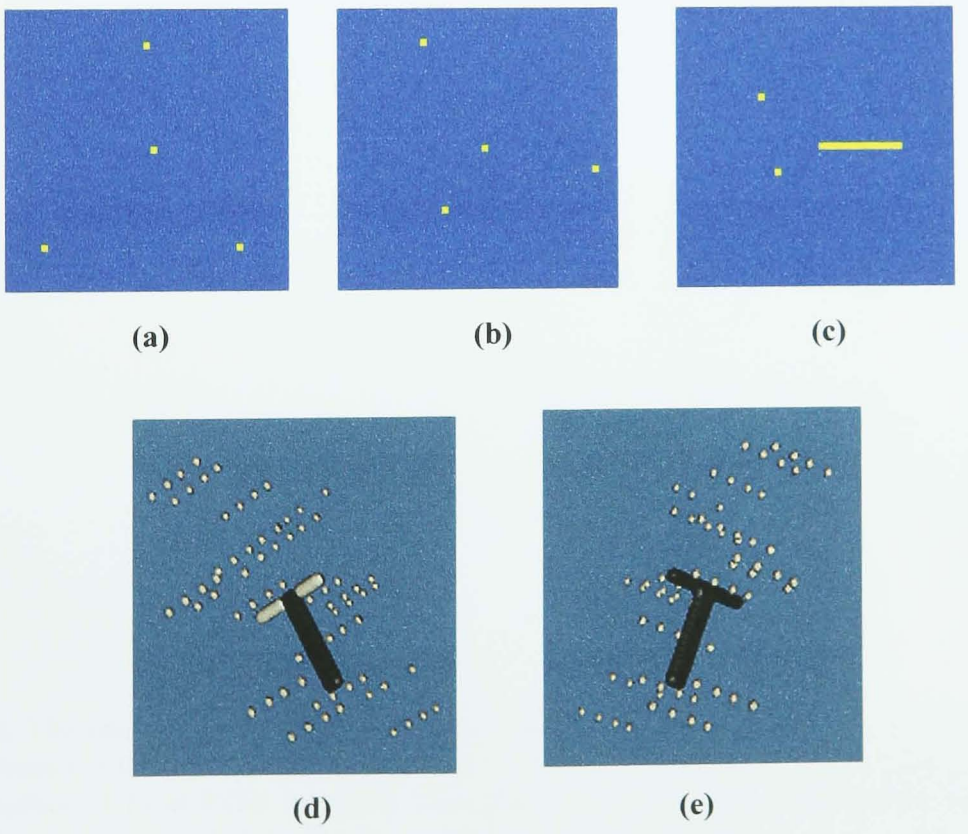


Fig. 4.14. 2D and 3D segmentation of an artificial volume. (a-c) Three slices from the volume, (c) is a middle slice. (d) Pseudo-3D segmentation. (e) Isovolume segmentation.

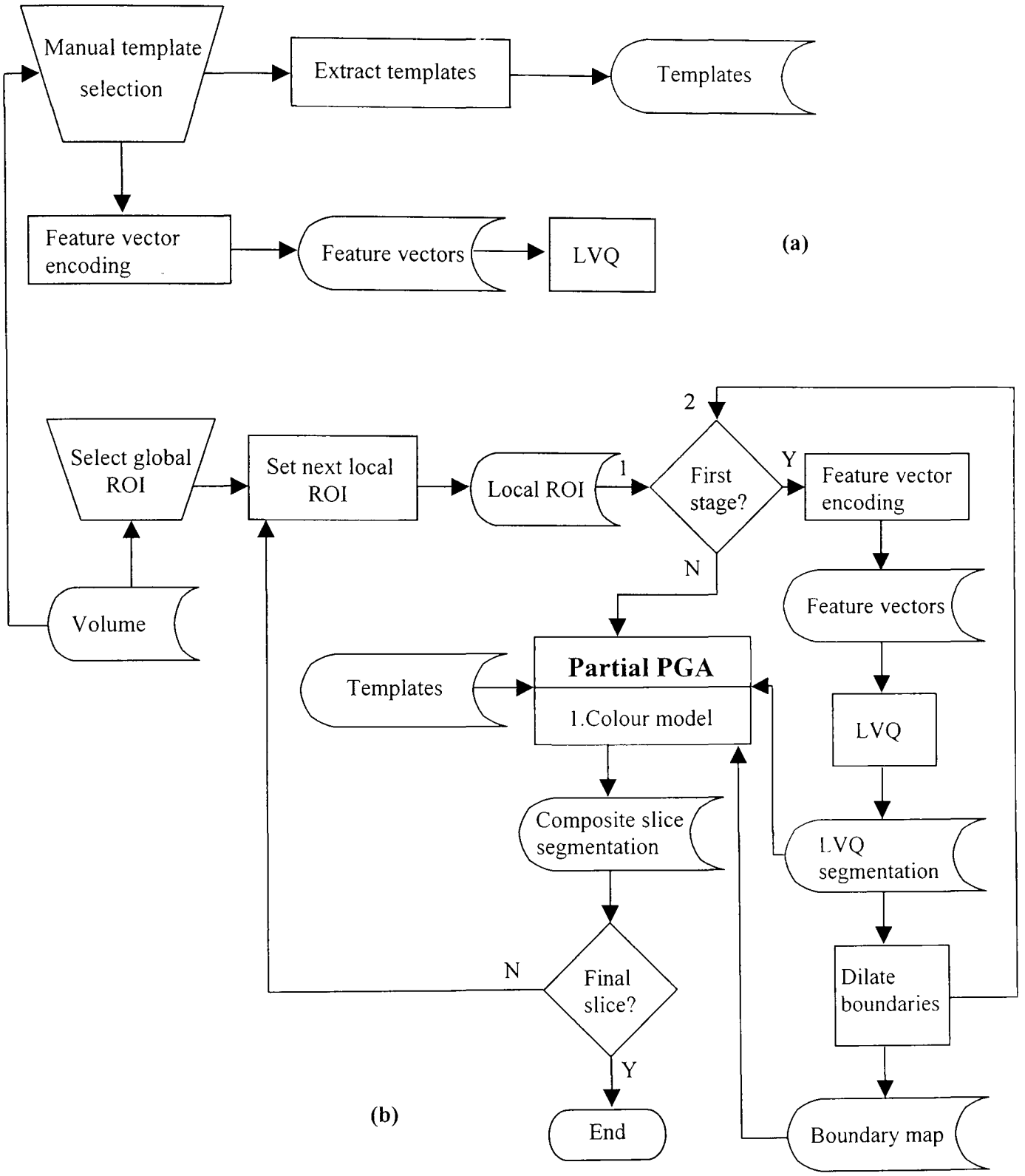


Fig. 4.15. The partial ACSR volume segmentation pipeline. (a) Template selection with extraction of templates for the PGA and training of LVQ classifiers. (b) Selection of global ROI followed by segmentation of local ROIs for each slice plane. Initial LVQ segmentation followed by partial PGA segmentation producing the composite final segmentation.

4.10. Isovolume segmentation of Visible Human Project colour cryo section volumes.

Two colour volumes of cryo section data from the Visible Human Project male data set have been segmented using partial ACSR and LVQ.

4.10.1. Segmenting blood supply to the hip bone.

A volume containing slices of the hip bone was chosen because it contains blood vessels, which are both in and nearly orthogonal to the slice plane. Some of the vessels are suitably small to test the benefits of using 3D information in this 120*147*46 volume. One of the slices from the volume is shown in fig. 4.16(a). Fig. 4.16(e) shows a Sobel filter applied to the slice followed by point thresholding. The contours of the blood vessels are almost identical to those of the structures inside the bone marrow. Two templates and classifiers were used for the partial ACSR segmentation, one for blood vessels and one for bone. Fig. 4.16(b) shows a 2D segmentation of this slice and a comparison with fig. 4.16(d) (the volume segmentation of the same slice) reveals that three small vessels are only visible in the volume segmented slice. Fig. 4.16(c) shows the initial LVQ segmentation. Notice how the vessels are segmented, even if the segment boundaries are not accurate. Following dilation of the segment boundaries and applying the PGA, the shape of the segments are refined. Because no template was used for the surrounding muscle tissue, it is identified as vessels, since out of the two it is closest to that class. The goal of this segmentation was to clearly distinguish between bone and blood vessels and only these two texture classes were templated. The vessel segments were clearly isolated within the solid bone segment and the rest easily removed before boundary dilation and segmentation (fig. 4.16(d)). The partial ACSR produced the same results as full ACSR but 8.5 times faster. Fig. 4.16(f) shows an isosurface model rendered from the segmented volume.

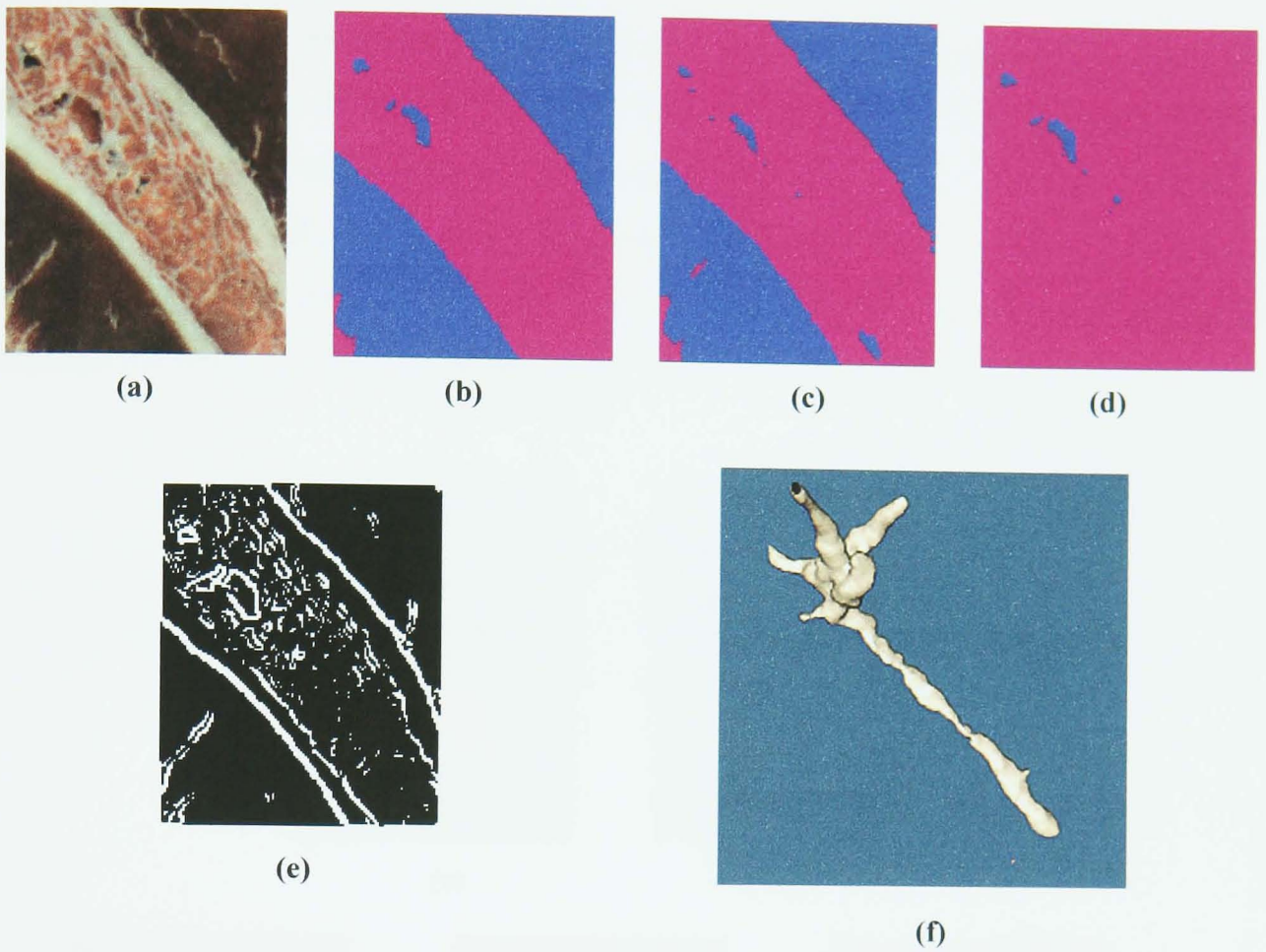


Fig. 4.16. ACSR segmentation of vessels of the hip bone. (a) An original slice. (b) Pseudo-3D segmentation. (c) Initial LVQ segmentation. (d) Isovolume segmentation. (e) Sobel filter and thresholding. (f) Isosurface model of isovolume segmentation (rotated view).

4.10.2. Segmenting the shaft of the radius.

For another volume example (181*230*52) the shaft of the radius was chosen, with the aim of accurately segmenting hard bone and bone marrow. The templates and classifiers were bone, marrow and muscle. A slice from this volume is shown in fig. 4.17(a). In fig. 4.17(b) a Sobel filter is applied to the slice followed by thresholding. The boundary between hard bone and marrow is not well defined. While the vessels in the hip bone had relatively distinct boundaries, they were often too small for psedudo-3D segmentation. The structures are larger in the radius volume, but the boundaries between hard bone and marrow are in some places highly diffuse. This volume thus provided a more difficult segmentation task for the partial ACSR, but one, which would have been successful as pseudo-3D segmentation for some areas of the volume. However, where the boundaries were very diffuse, information from neighbouring slices helped preserve consistency in the segmentation. Figs. 4.17(c) and 4.17(e) show

the pseudo-3D and volume segmentations of a slice, and although similar, some fine detail is only visible on the volume segmentation. The initial LVQ segmentation (fig. 4.17(d)) shows the expected edge artefacts. Partial ACSR again produced the same results as full ACSR, in this case 5.6 times faster. Fig. 4.17(f) shows an isosurface model of the hard bone without the bone marrow. Fig. 4.17(g) shows the marrow through the hard bone (semi-transparent).

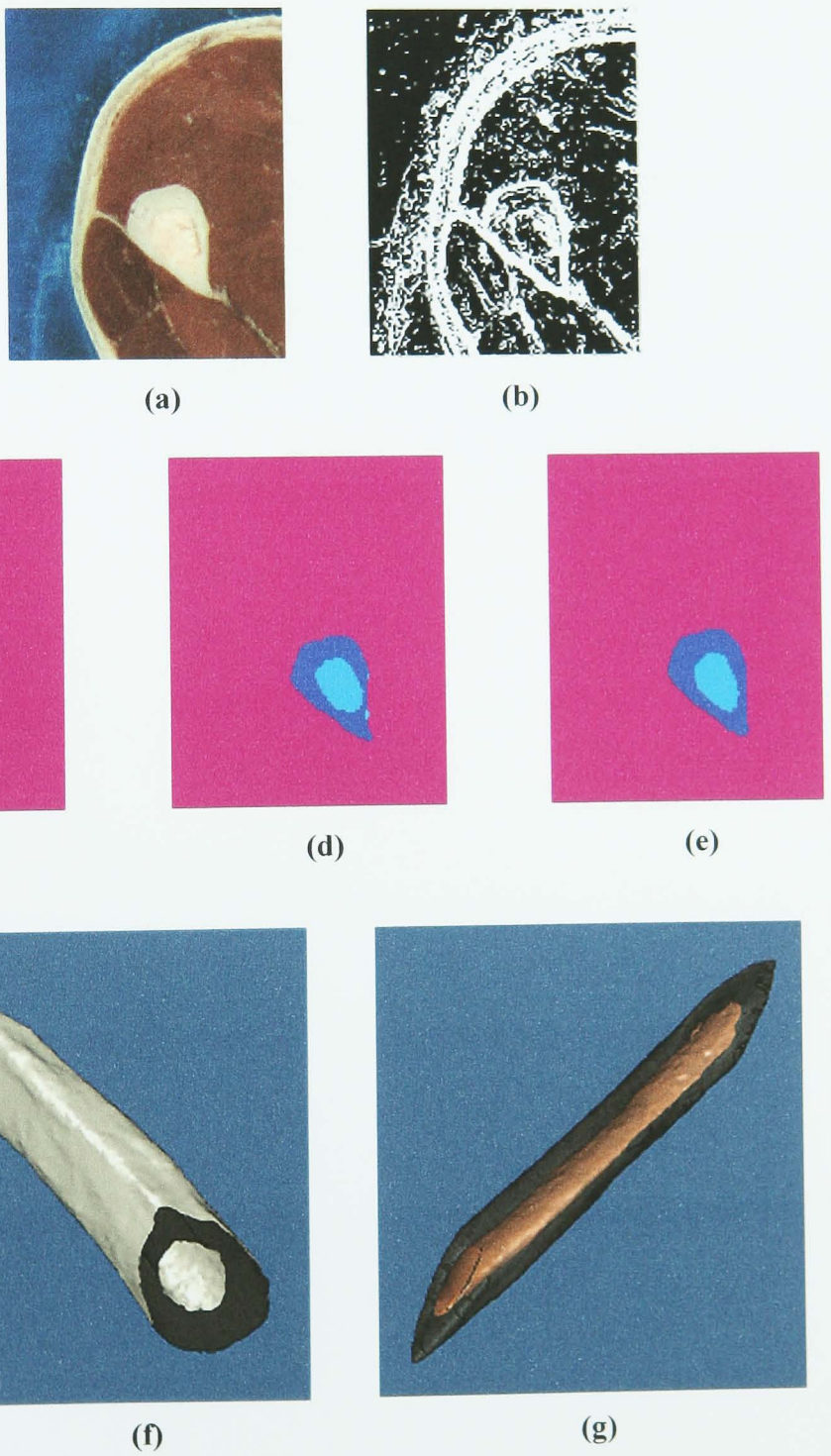


Fig. 4.17. ACSR segmentation of the radius. (a) An original slice. (b) Sobel filter and thresholding. (c) Pseudo-3D segmentation. (d) Initial LVQ segmentation. (e) Isovolume segmentation. (f,g) Isosurface models of isovolume segmentation (rotated views): (f) Hard bone. (g) Marrow and hard bone (semi-transparent).

4.11. Summary.

This chapter has established a new framework for semi-automatic segmentation. In the ACSR framework the user initialises segmentations visually by selecting class templates as representative image fragments. The following segmentation process is fully automatic. All classes are represented individually at every point, using topologically different sampling windows built from paths originating from the point to be classified. This is achieved using the PGA. The best representations for each class compete for the final classification. Connected regions and their boundaries emerge from these single point classifications. The technique was demonstrated for discrete 2D images and 3D volumes, showing that it can take advantage of information in n dimensions. It was shown that the PGA can be used as a spatially adaptable sampling method for encoding of LVQ feature vectors or as the algorithm for the final segmentation combined with an initial standard LVQ segmentation.

Chapter 5

Preliminary empirical evaluation

5.1. Choosing a methodology for empirical evaluation.

Visual presentation of segmentation results next to a source image or selected slices from a volume were used as the means of evaluating the success of the segmentation of cryo section data in chapter 4. This leaves each individual viewer to form their own opinion about the success of a segmentation compared to the source image and possibly alternative segmentations. It is obvious that such evaluation is only acceptable if the purpose is to merely demonstrate that a segmentation method works, without quantifying how well it works. The artificial test image used by Campbell et al (fig. 4.9) was segmented and compared to the ground truth pixel by pixel. Because the image was artificially generated and each segment in the foreground is delimited by a simple geometric shape (squares and a circle), the exact ground truth is known. Segments are placed in the image and their boundaries are perfectly crisp with no ambiguities. In real images the composition of segments is unknown and boundaries are fuzzy (as pointed out in chapter 4, section 4.6). Therefore a ground truth is not given, but has to be manually generated.

The conceptual requirements for a semi-automatic segmentation system outlined in chapter 3, section 3.2 were strongly based on the fact that the success of any segmentation of real data is subjective, because one cannot measure and quantify segment area and location at the source. It can be simulated, but not measured. Bowyer [138] notes that not only do different human observers produce different ground truth images. The same observer may produce a significantly different ground truth from the same source data when asked to complete this task twice on different days. He reported as little as 28% overlap between traced regions in an example using X-ray mammograms [138]. While this is an extreme example involving one of the

most difficult imaging problems today, the problem exists for all imaging modalities and applications. In spite of these problems, the quantitative ground truth evaluation remains the most common type of evaluation of image segmentation. Qualitative evaluation through visual ranking of images, such as the method suggested by Heath et al [160] (see chapter 7, section 7.1), is much harder to use. It requires a group of experienced observers and it has to be assumed that these observers share a sense of "goodness of segmentation" [160]. Heath et al addressed this problem by showing statistically that there was consistency between observers' rankings over a large number of images. In order to show this a substantial number of observers must be used and great care must be taken not to introduce any kind of bias. Reproducing or comparing evaluations between different groups of researchers is not possible based only on the segmentation results from one group. Ground truth comparison on the other hand makes it straightforward. As long as a standard ground truth exists for a standard test image or volume and a standard performance metric is employed, new results can be easily compared to previously published results. The big problem with quantitative evaluation of segmentation images, whether it is based on area or boundary location, is that it does not embody this "goodness of segmentation" which appears to be inherent in human observers. A segmentation with a low error rate compared to a ground truth may have only a small area of misrepresented information, but this information could be more important than any other information in the data set. In other words ground truth comparison for real data does not quantify how well information, which is crucial for a human user in a specific application, is conveyed in the segmentation. Cinque et al expressed this in [164] by observing that:

> Although it would be nice to have a quantitative evaluation of performance given by an analytical expression, or more visually by means of a table or graph, we must remember that the final evaluator is man and that his subjective criteria depend on his practical requirements.

While there is clearly a problem with the use of ground truth segmentation in real data, its use must be considered acceptable in artificial data (created manually or automatically from a finite set of primitives) or simulated data (created using mathematical equations emulating real processes of image construction) where the process of creating the data itself specifies the ground truth exactly.

In this chapter, ground truth evaluation is used for the evaluation of the segmentation of artificially created compositions of real textures in section 5.2. In section 5.3 and its subsections, natural colour images and cryo sections are evaluated through ground truth comparison in a comparative study of several algorithms used for ACSR segmentation, followed by a discussion of the results. The source images and segmentations (including manual ground truth), discussed in section 5.3 are included on the companion CD.

5.2. Evaluating the robustness of ACSR - a pilot study.

Robustness is an important property in any machine vision application. In medical image analysis it is vital. Robustness to common image artefacts is a separate problem which will be explored in chapter 6. The robustness under investigation in this section is defined as the ability to produce consistently accurate segmentation results over a number of different images using the same initialisation. The successful application of edge detection algorithms depends on the selection of parameters (typically kernel size and thresholds). Region growing algorithms are affected by the selection of seed points and the criteria for growth. Bayesian classifiers depend on good priors. ACSR using the PGA depends on the selection of class templates. To achieve robustness several sets of templates selected by different individuals should produce little variation in segmentation accuracy when the PGA is applied to the same image or volume. Because ACSR can accommodate boundary points without having to include them in the template sets, any composition of textures can in theory be segmented equally well using a representative template set.

A pilot study was carried out to test the variability in segmentation accuracy using ACSR with multiple template sets. The study used five participants with no experience in image processing to select templates and compose images from six different texture images. Two texture images had the same texture (plastic with a grating pattern), but in two different orientations (one was rotated by 90 degrees compared to the other). The texture images were original photographs of arbitrary real surfaces: stone, wood, meat sausage, orange (flesh) and plastic (see fig. 5.1). The

photographs were acquired digitally with an Olympus digital camera without special studio lightning. They provided a more realistic, less perfect quality test set than textures from established texture collections, such as Brodatz [147] and VisTex [165]. Textures ranged from homogeneous and near-regular to highly inhomogeneous and irregular.

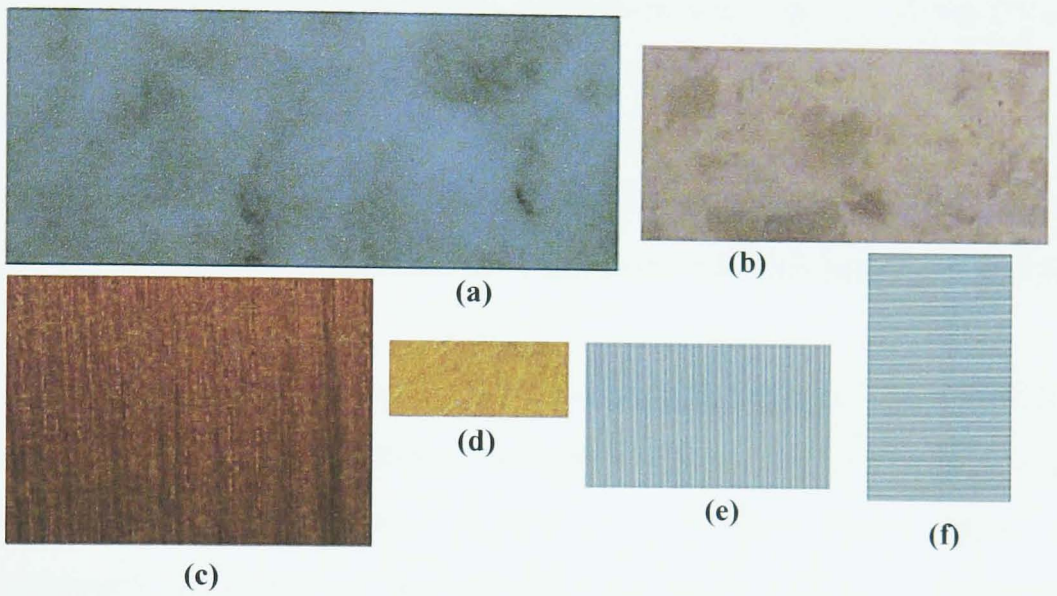


Fig. 5.1. The texture images used in the image composition and template selection experiment [166]. (a) Stone. (b) Meat sausage (c) Wood. (d) Orange (e) Plastic (f) Plastic 90 (same as 5, rotated 90 degrees).

Participants carried out two tasks. In a template selection task they were instructed to select one or more representative areas from each texture image. The total selection could not exceed one quarter of the size of any image. Templates were only selected from one of the two plastic texture images but used to segment both orientations. In an image composition task, participants composed a new image using selections (different from the first task) from the six texture images. They were allowed to choose one texture as background and paste two other textures into the background. One of these had to be one of the two plastic textures in either orientation. Selections from the texture images could be made using either a circular or a rectangular shape. Subsequent to the completion of both tasks for the five participants, all template sets were used to segment all composed images individually. This yielded 25 automatic segmentations. Results were obtained from the PGA directly. A segmentation of all five images using one participant's template set is shown in fig. 5.2. The results

showed that all 25 segmentations produced the exact ground truth segmentations except for two, where in each case only 1 pixel was misclassified. All combinations of textures in participants' template sets corresponding to combinations in the composed images showed a low degree of overlap. Similarly the areas selected as templates for the same classes varied greatly between participants. The plastic texture showed the smallest degree of overlap between template selections with no overlap at all in 7 out of the possible 10 combinations of the 5 template sets. Fig. 5.3 shows the overlap for the stone texture fragment. The smallest template set (which produced the exact ground truth segmentation for all five images) contained only 18.5% of the total number of possible unique pixels (sets of colour descriptors) in all texture images. This set was four times smaller than the largest set, but both performed equally well.

In conclusion the pilot study described in this section indicated that the PGA can offer robustness to different initialisations. Templates selected by different individuals, who had no training in using the system or other types of image segmentation, accurately and consistently facilitated the segmentation of arbitrary image compositions. These were comparable to the test image by Campbell et al (chapter 4, fig. 4.9) but using natural textures. For a more detailed account of this pilot study and analysis of the experimental data, please refer to [166].

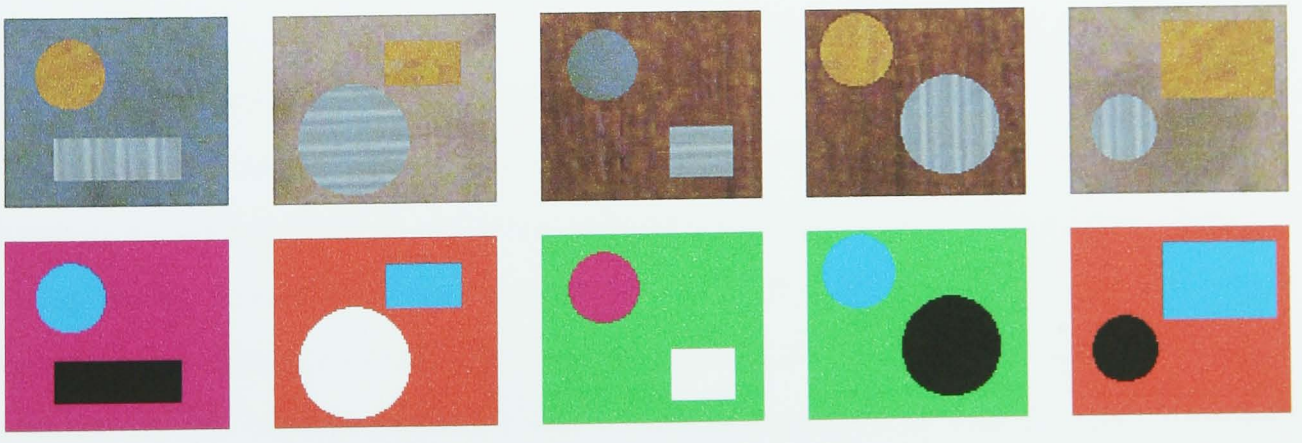


Fig. 5.2. Manual image compositions and their segmentations. Top row: images 1 to 5 from left to right. Bottom row: segmentations of images above (using template set no. 5). Segments are represented as pseudocolours. The two different orientations of the plastic texture are classified differently.

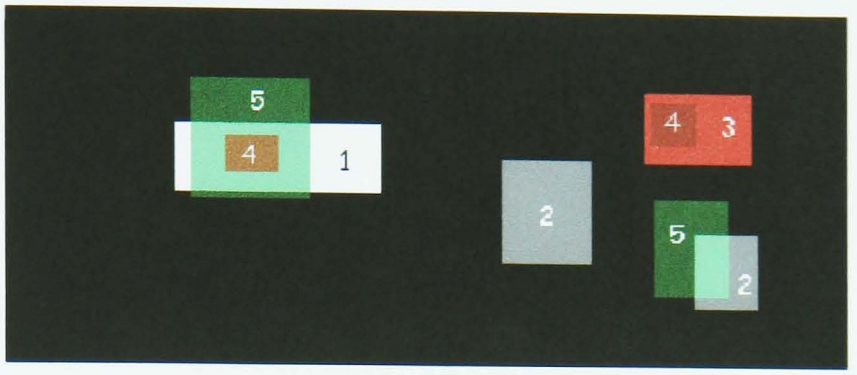


Fig. 5.3. Visualisation of image areas from the stone texture fragment selected as templates by the participants in the image composition and template selection experiment [166]. The number on an area identifies the participant.

5.3. A comparative study of colour image segmentation.

This section describes a comparative study based on two sets of images. The first set consists of 6 large, high resolution colour images originally published by Kato [167]. The images are photographs of natural scenes and objects. The second set is a series of 5 colour cryo section slice images of the brain from the Visible Human Project Visible Male data set. Six different methods for initial segmentation, full and partial ACSR were used to segment the images.

5.3.1. Segmenting six natural colour images.

The Visible Human Male and Female data sets are the standard for the testing of algorithms for colour cryo section segmentation. However due to the amount and diversity of data in the full volumes, only smaller subvolumes are normally used for testing image processing algorithms. Because no specific subvolume has become a standard for empirical evaluation and no standard ground truth data set for the volumes of the Visible Human Project exist, the choice of data to evaluate on is somewhat arbitrary, and one set of results is not easily comparable to those achieved by other researchers.

In chapter 4 an image of an eagle over water was used to demonstrate ACSR segmentation compared to an alternative segmentation method. An image of a polar

bear in an arctic environment was used in [161] for the same purpose. Although these images are non-medical, they are images which have been published previously as sets of source images and their segmented counterparts. The source images are easily available in their original form because they come from an established collection (the Corel photo CD series).

This section presents results based on six natural colour images (see fig. 5.4) previously published by Kato [167] in a paper on automatic segmentation using Reversible Jump Markov Chain Monte Carlo (RJMCMC). The images are standard machine vision test images from Kodak (USA), INRIA (France) and the IEN Computer Vision Research Group (Istituto Elettrotecnico Nazionale, Italy) and their segmentations were presented visually by Kato [167] with information about the speed of the segmentation process. RJMCMC, like the SOM, is capable of fully automatic segmentation where the number of classes does not have to be determined a priori. Although this section will comment briefly on the results achieved by Kato as a representation of a fully automatic method, the comparative study described in this section should not be regarded as a comparison between the quality of segmentation achieved by Kato and that achieved using ACSR segmentation. The six images used by Kato were chosen for a comparison of techniques used for ACSR segmentation because they represent a set of standard test images with previously published results, the equivalent of which currently does not exist for medical colour cryo section images. In spite of complications such as shadows and changing camera focus in these images (which do not apply to cryo section images), their segmentation is merely an abstraction of a 2D cryo section segmentation task.

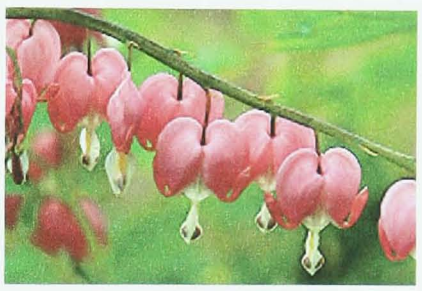
Segmentation of the six natural colour images was performed using the following types of classifiers:

- Point Based Nearest-Neighbour classifier using the RGB colour model (PBNN-RGB)
- Point Based Nearest-Neighbour classifier using Opponent Process Colour descriptors (PBNN-OPC)
- LVQ based on 54-dimensional PixelDefine and OPC
- Full ACSR using the PGA and RGB (PGA-RGB)

- Full ACSR using the PGA and OPC (PGA-OPC)
- Partial ACSR using the PGA and OPC, based on point based nearest-neighbour classification

The Point Based Nearest-Neighbour classifier uses the same type of nearest-neighbour matching with a set of class templates as the PGA, but only on a single point at a time. This is similar to a simple thresholding, but rather than using exact intervals of colour descriptors the shortest distance to a template is used to determine the class of a point. Although fast, this technique lacks the benefits of region based methods and more than one class may give an equally good match, resulting in unclassified points.

Poppy



Hearts

Seagull

Rose

Bird11

Bird12

Fig. 5.4. Six natural colour images used for segmentation with multiple classifiers.

All classifiers were template based and used the same template sets. For the PBNN classifier and the full ACSR, results using the RGB and the OPC colour models were compared. For the LVQ classification 100 ± 1 nodes per class were used. Partial ACSR was based on the PBNN classification using OPC (similar to the encoding used

for LVQ). Results are presented for visual comparison on the companion CD, including the results of the RJMCMC segmentation. It is apparent that the goal oriented nature of ACSR segmentation allowed for a more refined segmentation. Clearly the results achieved with RJMCMC are impressive for a fully automatic method, but the extra time taken for initialisation of ACSR segmentation is compensated for by the much faster processing time. Although the efficiency of the implementation of each of the two techniques in software is an obvious issue when comparing processing time, it should be noted that the RJMCMC implementation was run on a far superior architecture compared to the ACSR implementation.

Table 5.1 shows the image sizes and the window sizes used by the PGA and LVQ and the dilation factor employed before partial PGA. For the PGA a path length of 5 was used at all times. Processing time for each type of classifier for every image is shown in table 5.2 including the processing time for the same images using the RJMCMC algorithm in [167].

All processing except for RJMCMC was carried out on a Dell Inspirion PC with a single Pentium-III processor running at 500 MHz and 64 MB of physical memory.

RJMCMC processing was carried out on a Silicon Graphics Origin 2000 server with 16 R10000 processors each running at 250 MHz and 8 GB of physical memory.

The segment classes were chosen for the template based methods to resemble the classes found in [167]. Fig. 5.5 shows an example of template selection for the Poppy image.

Table 5.1. Sizes of natural colour test images, sampling windows for the PGA and LVQ and the dilation factor used for partial ACSR. All values (except the dilation factor) are in pixels.

	Poppy	**Hearts**	**Seagull**	**Rose**	**Bird11**	**Bird12**
Image size	512*512	736*492	458*381	734*486	498*332	498*332
PGA window	11	11	8	11	8	8
LVQ window	5*5	5*5	3*3	5*5	3*3	3*3
Dilation	2	3	2	3	2	2

Table 5.2. Processing time for each type of classifier on each natural colour test image.

	Poppy	Hearts	Seagull	Rose	Bird11	Bird12
PBNN-RGB	45 sec	107 sec	9 sec	54 sec	21 sec	4 sec
PBNN-OPC	88 sec	107 sec	13 sec	62 sec	24 sec	7 sec
PGA-RGB	423 sec	639 sec	542 sec	823 sec	248 sec	247 sec
PGA-OPC	505 sec	707 sec	606 sec	932 sec	299 sec	278 sec
LVQ	135 sec	191 sec	67 sec	183 sec	36 sec	32 sec
Partial PGA	54 sec	137 sec	16 sec	127 sec	19 sec	14 sec
Partial ACSR, total	142 sec	244 sec	29 sec	189 sec	43 sec	21 sec
RJMCMC	36 min (1)*	253 min (2)	109 min (1)	211 min (2)	150 min (2)	87 min (1)

(1) LUV colour space (2) LHS colour space * The image used was half the resolution compared to the one used with the other segmentation methods

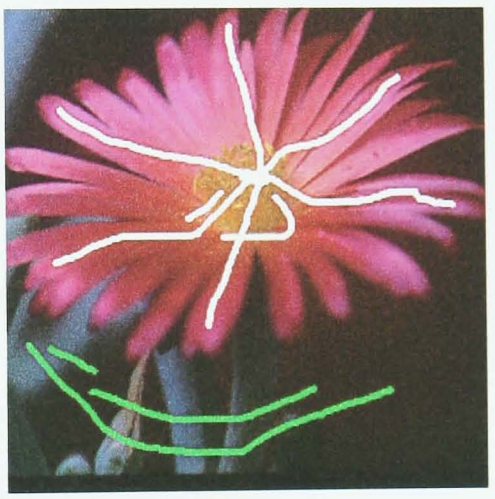


Fig. 5.5. Template selection for the Poppy image. Templates are shown in solid white (head of flower) and green (surround).

A number of observations can be made by studying the images and the entries in table 5.2. The PGA based segmentation clearly produced the best results for all images. For these images the PBNN classifier consistently produced equal or better results compared to LVQ and at speeds up to 5 times faster (see table 5.3). Partial ACSR used PBNN-OPC for the initial segmentation step, producing results 100% identical to the full PGA-OPC up to 20 times faster (Seagull image). The dilation factors given in table 5.1 were the minimum factors required to achieve a 100% correspondence with the full PGA-OPC in each image. Visual inspection of the segmented images shows a slight advantage of OPC compared to RGB both for the PBNN classifier and the PGA. The difference is small but particularly distinct in the images Bird11 and Bird12. These observations are of course highly subjective and are merely the

opinions of the author. Consequently they must be backed up by a more solid evaluation.

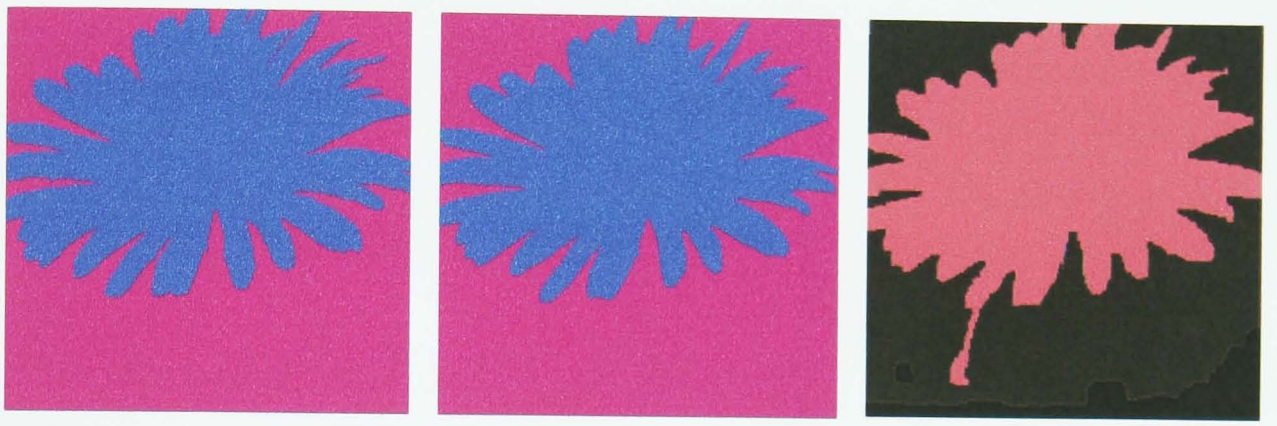
Table 5.3. PBNN and LVQ compared to full ACSR segmentation of natural colour test images. Each percentage shows the similarity between the segmentation produced by PBNN or LVQ and the full ACSR for each image.

	Poppy	Hearts	Seagull	Rose	Bird11	Bird12
PBNN-OPC	98.68%	96.68%	99.93%	98.22%	99.13%	99.56%
LVQ	98.71%	91.25%	99.86%	97.22%	97.34%	99.55%

A ground truth segmentation of the six images was created by a human observer manually tracing the boundaries of the desired segments in each of the source images. A segmentation goal in the form of the number of classes and what they should represent was determined for each image in advance. While the ground truth for the Seagull image is clearly identifiable, an image such as Hearts is highly likely to produce different ground truth images when boundaries are traced by different individuals, since the exact delineation of some segments is not obvious. Even in the Poppy image the fuzziness of the boundaries is prone to produce different ground truths by different observers. Table 5.4 shows the results of each of the five classifiers compared to the manually traced ground truth. The percentages show the ratio of correctly classified points according to the ground truth. It is evident that on a pixel per pixel basis the PGA performed well compared to the ground truth with a high of 99.58% and a low of 94.83% for PGA-OPC. It is not surprising that the Hearts image would give the lowest match, since parts of the ground truth for this image is slightly arbitrary. The Mann-Whitney U-test performed on PBNN-RGB vs. PBNN-OPC and PGA-RGB vs. PGA-OPC (based on number of correctly classified pixels) showed that there was no significant difference (at the $p = 0.05$ level) between results based on RGB and those based on OPC. Fig. 5.6 shows an example from the Poppy image with ground truth, PGA-OPC and RJMCMC [167].

Table 5.4. Segmentation of natural colour test images compared to a manual ground truth.

	Poppy	Hearts	Seagull	Rose	Bird11	Bird12
PBNN-RGB	96.71%	93.23%	99.56%	97.10%	98.46%	98.94%
PBNN-OPC	96.57%	93.28%	99.60%	97.15%	98.52%	99.06%
PGA-RGB	96.89%	94.84%	99.55%	97.56%	98.98%	99.07%
PGA-OPC	96.82%	94.83%	99.58%	97.56%	98.94%	99.09%
LVQ	96.58%	90.82%	99.56%	96.47%	96.90%	98.96%



(a) (b) (c)

Fig. 5.6. Segmentation of the Poppy image. (a) Manually selected ground truth. (b) PGA-OPC. (c) RJMCMC [167]. The goal for the ground truth and the ACSR segmentation was to achieve two segments, one representing the top of the flower and another representing everything else.

To compare the performance of the classifiers on corrupted image areas, 10% random impulse noise was added to each of the six source images. The images where then segmented again based on the same templates as the clean images and compared to the ground truth. The results in table 5.5 show that the added noise had a maximum of 1% effect on the PGA-OPC based segmentation with up to 7.02% effect for PBNN-OPC and up to 5.77% effect for LVQ. Visually the images show a dramatic difference between the PGA based segmentations and the two other types of classifiers.

Table 5.5. The effect of 10% random noise on the segmentation of the natural colour test images. Each percentage shows the difference in classified pixels between the segmentation based on the noisy image and that of the clean image.

	Poppy	Hearts	Seagull	Rose	Bird11	Bird12
PBNN-OPC	3.17%	4.44%	5.01%	4.15%	3.47%	7.02%
LVQ	3.00%	3.32%	4.92%	4.18%	4.09%	5.77%
PGA-OPC	0.27%	1.00%	0.06%	0.54%	0.29%	0.23%

5.3.2. Segmenting a brain cryo section series.

Five consecutive images from the Visible Human Project Visible Male data set from the brain region were segmented using PBNN-OPC, PBNN-RGB, LVQ, PGA-OPC, PGA-RGB and partial ACSR. Each image was 120*100 in size. The PGA used the path length 5 and the window size 11. The LVQ used a window size of 5*5 and 100±2 nodes per class. Template sets and learning vectors were created mainly from the second image in the sequence (Brain1) with a few minor areas selected from Brain3, to include features not present in Brain1. All images were segmented based on those templates and the codebooks based on those learning vectors. The dilation factor 3 was used to achieve 100% identical results for the partial ACSR compared to the full ACSR. Table 5.6 shows the processing time for each type of classifier.

Due to the extensive boundary lengths in the images, the partial ACSR did not provide speed-ups at the levels found in section 5.3.1, but it still produced a speed increase of approximately 30% compared to full ACSR. The results for one slice are shown is fig. 5.7. The results for all five slices can be found on the companion CD.

Table 5.6. Processing time for each type of classifier on each cryo section brain image.

	Brain0	Brain1	Brain2	Brain3	Brain4
PBNN-RGB	<1 sec.	<1 sec.	<1 sec.	<1 sec.	<1 sec.
PBNN-OPC	<1 sec.	<1 sec.	<1 sec.	<1 sec.	<1 sec.
LVQ	7 sec.	7 sec.	7 sec.	7 sec.	7 sec.
PGA-OPC	30 sec.	30 sec.	30 sec.	29 sec.	30 sec.
PGA-RGB	27 sec.	27 sec.	26 sec.	27 sec.	27 sec.
Partial PGA	20 sec.	22 sec.	20 sec.	21 sec.	22 sec.
Partial ACSR, total	21 sec.	23 sec.	21 sec.	22 sec.	23 sec.

The five brain images must be considered a tough segmentation task due to the similarity between the segment classes. Three classes were templated: white matter, grey matter and "all other tissue". The latter class was extremely inhomogeneous with areas highly similar to both the grey matter and white matter classes . The PGA still produced good results, while the PBNN and LVQ performed poorly in comparison both to the PGA and to the results seen in section 5.3.1. Table 5.7 shows the similarity

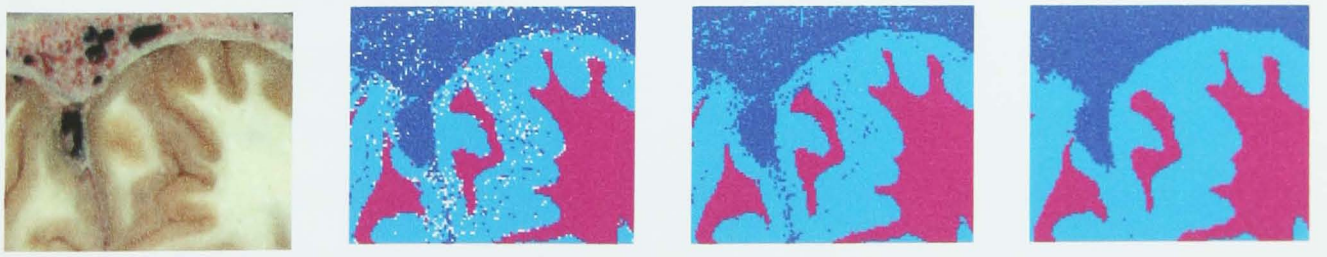
between the segmentations produced by the PBNN and the LVQ compared to the full ACSR. It can be seen that the similarity for PBNN with full ACSR is inversely proportional to the distance from Brain1 (which provided most of the areas used for template creation). The similarity of LVQ with full ACSR does not show such an effect and was consistently higher than for PBNN. This is not surprising, given that LVQ has the ability to generalise over the data. Thus for the brain series partial ACSR was based on LVQ for the initial segmentation step. In comparison to a manually generated ground truth, the LVQ also shows better results than PBNN (see table 5.8). Similarly to the natural colour images in section 5.3.1 OPC visually appeared to facilitate slightly better segmentation than RGB, but again the U test did not show a significant difference.

Table 5.7. PBNN and LVQ compared to full ACSR segmentation of cryo section brain images. Each percentage shows the similarity (overlap) between the segmentation produced by PBNN or LVQ and the full ACSR for each image.

	Brain0	Brain1	Brain2	Brain3	Brain4
PBNN-OPC	90.50%	91.32%	89.33%	89.03%	88.41%
LVQ	93.53%	91.73%	92.88%	91.98%	92.00%

Table 5.8. Segmentation of cryo section brain images compared to a manual ground truth.

	Brain0	Brain1	Brain2	Brain3	Brain4
PBNN-RGB	88.66%	90.07%	86.66%	84.94%	85.88%
PBNN-OPC	88.93%	89.79%	86.68%	85.13%	86.14%
PGA-RGB	93.49%	93.03%	92.53%	91.90%	92.13%
PGA-OPC	93.75%	92.88%	92.88%	91.95%	91.72%
LVQ	92.14%	90.59%	90.21%	88.33%	89.99%



(a) (b) (c) (d)

Fig. 5.7. Segmentation of a cryo section brain slice. (a) Source image. (b) PBNN-OPC (white points are unclassified). (c) LVQ. (d) PGA-OPC.

The effect of noise on the brain images was studied by adding 10% random noise to the source images and comparing the segmentation results for PBNN-OPC, LVQ and PGA-OPC based on the original templates to the segmentations of the clean source images. The results are shown in table 5.9. The maximum effect for PGA-OPC was 3.25%.

Table 5.9. The effect of 10% random noise on the segmentation of the cryo section brain images. Each percentage shows the difference in classified pixels between the segmentation based on the noisy image and that of the clean image.

	Brain0	**Brain1**	**Brain2**	**Brain3**	**Brain4**
PBNN-OPC	9.25%	8.35%	9.51%	9.59%	9.72%
LVQ	5.81%	5.52%	5.42%	5.71%	5.59%
PGA-OPC	2.84%	2.29%	2.92%	3.02%	3.25%

5.4. Summary.

The preliminary conclusions that can be drawn from the results in section 5.3.1 and 5.3.2 are first of all recommendations to use a fast point based nearest-neighbour classifier for discrete 2D image segmentation using the partial ACSR, and LVQ for image sequences (volumes). There is obviously a trade-off, which must always be considered: A fast initial segmentation step with a high degree of oversegmentation requires a slower partial PGA processing. A slower, more accurate initial step allows for a faster partial PGA processing. Based on the ground truth evaluation there appears to be no advantage of using OPC rather than RGB, which is counterintuitive to expectations. However as pointed out in the beginning of this chapter, ground truth comparison may not quantify how well essential visual information for a human user is preserved. In chapter 7 the images from section 5.3.1 and 5.3.2 will be revisited in a series of human observer experiments. The qualitative results obtained will be compared to the results presented in this chapter. The robustness of ACSR segmentation to multiple initialisations, which showed good results on artificially composed images in section 5.2, will also be investigated further in chapter 7 using real cryo section data.

Chapter 6

Extending the ACSR framework to greyscale MRI segmentation

6.1. Developing ACSR segmentation for greyscale medical imaging scans.

The application of ACSR segmentation to a new imaging modality requires modifications of the algorithm used for implementation. For any type of intensity based segmentation, the information density (the richness of information at the atomic level) is crucial. Cryo section images offer three channels of typically 256 intensity levels per point, and the combination of the three channels in tried and tested colour models give a representation, which helps a classification system to perform well. MRI and CT scans typically offer only one channel of 12-bit information (4096 intensity levels) per point. As opposed to cryo section images though, the grey levels in ideal CT and MRI scans directly correspond to tissue types. Grey levels for different tissue types represent the different levels of absorption of X-rays for CT or the emitted frequencies of radio waves following exposure to a pulse of radio frequency magnetic field oscillations for MRI scans respectively. Unfortunately partial voluming, noise and inhomogeneity can cause severe artefacts. These problems call for higher robustness and specific countermeasures. The sole dependency on point descriptors from unfiltered data in the PGA used for cryo section segmentation is not sufficient.

Additional point descriptors may be obtained from filtered versions of the original source images. This technique could be applied to use local high and low frequency information for additional descriptors. It could for example be achieved using a high-pass Butterworth filter to enhance high frequency information and a median filter to suppress high frequency information, while still preserving edges. Point descriptors might thus be derived from the original image, the high-pass and the low-pass version. Such an approach does not, however, conform to the requirement of

minimising the distortion of detail. Any type of filtering will enhance some features while attenuating others. The median filter for example effectively reduces noise without removing connected edges, but fine detail is washed out in the process.

To take advantage of the combination of low and high frequency information, while largely avoiding the distortion of detail, filtering can be applied *locally* rather than globally within a neighbourhood. In the ACSR framework this can be achieved using *path descriptors* in addition to the type of point descriptors used in colour data. Two different path descriptors have been implemented and tested with the PGA. The median path descriptor represents the median of the intensities at each vertex of a path. The AID (Average Intensity Difference) descriptor represents the average intensity difference between the point being classified (seed point) and all other points in the path. This descriptor is also used in the PixelDefine encoding. Kato et al [168] used a similar method (intensity mean and variance) to describe classes in an MRF model. MRF models are strong contenders in the field of greyscale medical image segmentation with their ability to model segment classes and artefacts which affect them.

The remainder of this chapter will focus on the MRI imaging modality. The development of the PGA for MRI segmentation under the ACSR framework is described, starting with a comparative study of three different PGA algorithms using path descriptors. Results are based on simulated data and ground truth evaluation throughout. Subsequently an optimisation of the developed method is proposed, including countermeasures for noise and inhomogeneity artefacts. Comparisons with previously published results on the same data are given. Finally multispectral MRI segmentation is described and the quantitative versus qualitative interpretation of the results are briefly discussed.

6.2 Evolving the PGA for MRI segmentation.

The initialisation of ACSR segmentation for MRI images is performed in the same way as for cryo sections. The user selects areas representing the desired segment

classes. These areas are the templates, which are encoded for processing by LVQ and the PGA.

The introduction of path descriptors means that paths are evaluated not only as sets of points, but as regions. The notation for the standard PGA given in chapter 4 (equations 4.1-4.4) is extended:

The path median for descriptor j is defined as:

$$PMED_{lj}=(PXD_{(M+2)/2})_j \quad \text{if } M+1 \text{ is odd} \tag{6.1}$$

$$PMED_{lj}=((PXD_{(M+1)/2}+PXD_{(M+1)/2+1})/2)_j \quad \text{if } M+1 \text{ is even} \tag{6.2}$$

The path AID is defined as:

$$PAID_l = \sum_j \sum_k \left| PXD_{kj} - PXD_{sj} \right| / M \tag{6.3}$$

For the purpose of single-channel MRI data j is equal to 1 in (6.3).

Equations (6.1), (6.2) and (6.3) are used for template creation and for the processing of novel images. We can now repeat (4.1) on a path level with three descriptors: point intensity of seed point, path median and path AID. This match value is denoted as PM_i and (4.2) can now be extended to:

$$PV_{il} = c\left(\left(\sum_{r=1}^{M} PXM_{ri}\right) + PXM_{si}\right) + PM_i \tag{6.4}$$

The combination of (4.2) and PM_i can be regarded as a multiple classifier problem. It was found experimentally that neither the use of (4.2) nor PM_i alone produce optimal results. However a combination of the two produced better results and optimal performance was found for $c=3$.

The introduction of equation (6.4) forms the basis of the basic version of the PGA with path descriptors (PGA-PD). Sampling windows consisting of single paths or several full paths are assumed (rather than the gradual addition of points from the path

hierarchy) to avoid expensive optimisation of the path median and AID, and an increase in processing overhead. Using a single path per window this algorithm is denoted as *PGA-SPD*.

While the PGA attempts to solve a windowing problem, the path descriptors with AID are vulnerable to noise corrupting the seed point. This could in itself result in a local windowing problem. The correct AID for the true class of the seed point might be found using a different point in the path from which to calculate the distance to the remaining points. Given that we do not assume a particular type of distribution for the noise, it is considered to be random and could affect any point within any path originating from any seed point. To increase the probability of seed points being correctly classified as part of their neighbourhood in this situation, a second version of the PGA-PD is introduced. The seed point with regards to the calculation of AID is shifted one point in the direction of growth. The actual seed point however remains the same for all other calculations. This reduces the effect of noise corrupting the seed point. If the point being shifted to is affected by noise then the representational strength of the actual seed point and the median will still be sufficient, given that enough points representable of the seed point's true class can be reached. This variation denoted as *PGA-SPDS* changes (6.3) to:

$$PAID_l = \sum_j \sum_{k \neq s} \left| PXD_{kj} - PXD_{(s+1)j} \right| / (M-1) \qquad \textbf{(6.5)}$$

Finally a third variation is introduced, which builds on PGA-SPDS, but uses two paths per window. The criterion for selection of the second path is as follows: The second path must have only the seed point in common with the first path and it must have the best match with its own class. Although the optimal path excluding the points contained in the first path is approximated to for the class in question and the template currently being evaluated, this is no guarantee that the path is actually more representable of its own class than of another. The segment boundaries may be such that a second path is forced to cross a boundary. To avoid this representational problem an extra check is performed. The path is evaluated against the competing templates and if a better match is found then only the representation based on the first path is used for the sampling window representation of the class in question. The

larger window size could be an advantage, particularly for very noisy images. This variation with double paths is denoted *PGA-DPDS*.

6.3. Standard MRI test sets.

Standard MRI test volumes are important for comparing results from different segmentation systems. Not only are results based on such volumes immediately comparable to previously published results by other researchers. Standard test sets also to a large extent make it unnecessary to re-implement algorithms developed by other researchers to compare results. This removes a classical source of error in comparative studies [137]. Standard ground truth sets allow different researchers to compare performance, given that the same performance metric is employed. This project has used MRI test sets from two standard collections described below: BrainWeb (section 6.3.1) and the Internet Brain Segmentation Repository (section 6.3.2).

6.3.1. BrainWeb.

The Brain Imaging Centre at the Montreal Neurological Institute of McGill University, Canada, provide an image database known as BrainWeb [41] containing simulated MRI data. Unlike phantom data sets in the traditional sense, BrainWeb is based on actual brain scans. Voxels were manually labelled and fuzzy tissue classes were created. The simulated data was modelled from these classes with different levels of RF inhomogeneity and noise. Because the robustness to these artefacts can be systematically tested for and because the exact ground truth is known (as opposed to in real data), it makes the BrainWeb data highly suitable for comparative evaluation. Volumes are available as T1, T2 and PD (Proton Density) images with 1, 2 and 3mm. slice thickness. Currently two types of simulated clinical data are offered: a healthy volume and a volume with multiple sclerosis. BrainWeb volumes have been used in a large number of previous studies of inhomogeneity correction and segmentation (see e.g. [130,169,170]).

6.3.2. The Internet Brain Segmentation Repository.

A diverse collection of real clinical MRI volumes is offered by the Internet Brain Segmentation Repository (IBSR) [42] provided by the Center for Morphometric Analysis at Massachusetts General Hospital, U.S.A. The collection comprises healthy as well as diseased brain volumes from individuals of different sexes and ages, acquired with different models of MRI scanners. T1, T2 and PD images are available and most volumes come with manually selected expert ground truth. As with BrainWeb the use of IBSR volumes is established in the literature (see e.g. [169,120,171])

6.4. Results on simulated MRI data I.

In order to produce and compare baseline results from the three proposed PGA-PD algorithms, three volumes from the BrainWeb database with varying levels of noise and inhomogeneity were selected. Partial ACSR using each of the three algorithms was applied without any pre-processing of the data. The parameters for the BrainWeb volumes were selected to comply with the volumes used by Pham and Prince [130] in a study of MRF segmentation using a standard Expectation Maximization (EM) algorithm [113] and the Adaptive Generalized EM (AGEM) algorithm [130], which models inhomogeneities. Two of these three volumes were also used in a previous study by Pham and Prince [120] on a fuzzy algorithm known as the Adaptive Fuzzy C-Means Algorithm (AFCM), which did not produce as good results as AGEM.

MRF based approaches traditionally require the manual selection of model parameters, which may be less intuitive to the target user than the visual initialisation process used in ACSR. MRF models have however become established as a robust tool for MRI segmentation, their main advantage being the dynamic adaptation to local image features and parameter optimisation. This generally reduces the effects of partial volume artefacts, noise and inhomogeneity. Although these artefacts may be modelled individually, their combination in a volume can severely reduce segmentation accuracy. Another drawback is the excessive processing overhead often

associated with MRF model computation. Although there is no gold standard of MRI segmentation, the encouraging results which have been achieved with MRF based methods such as AGEM must be regarded as a desirable initial goal for ACSR. In the studies by Pham and Prince [120,130] results were reported based on a standard performance metric (see section 6.4.1) and incorporating all of the three major tissue types of the brain (CSF, grey matter, white matter).

6.4.1. Experimental methodology.

The BrainWeb volumes used were all T1-weighed and had 1mm slice thickness. The parameter settings were:

- 3% noise with 20% inhomogeneity
- 3% noise with 40% inhomogeneity
- 7% noise with 20% inhomogeneity

Extracranial tissue was removed in an initial segmentation step based on the LVQ segmentation used for partial ACSR. Similar results could have been achieved with a tool such as BSE (automated Brain Surface Extraction program) from University of Southern California [172,173]. Three classes were templated: grey matter, white matter and cerebrospinal fluid (CSF). 127 transverse slices at a resolution of 181*217 were segmented in each volume, based on templates from 8 slices with a distance of 18 slices between them. Templates were based on the entire ground truth segmentation for each class in the 8 images (ground truth sets supplied by BrainWeb), simulating the manual template selection that a human user would produce. Template selections were used for the encoding of PGA point and path templates. LVQ encoding was achieved using a 3*3 sampling window and a very simple feature vector representing the same descriptors as for the PGA, i.e. centre point, window median and AID. PGA template encoding used a path length of 5. Interpolated templates and LVQ codebook vectors were produced for every 18 slices, using an offset of 9 from the first templated slice. They were based on quantisation of the two selected template sets on each side. Template interpolation was also achieved using LVQ. In all cases

the Optimized Learning Vector Quantization algorithm [117] was used. Initial experiments showed increased performance for slices between templated images, using this technique, compared to only using the original templates and LVQ learning sets. Segmentation of novel images always used the actual or interpolated PGA templates and LVQ codebook vectors closest to the current slice in the volume.

Because all seed points are independently represented and classified, artefacts caused by high levels of noise in ACSR segmented images are often restricted to single points. This means that the actual shapes of boundaries are preserved, but artefacts appear in a similar way to salt and pepper noise (impulse noise) in the segmented image. Cleaning up images using a median filter may improve results at high noise levels. A median filter with a kernel size of 3*3 was applied to the segmented images as an optional post-processing step for cleaning up the final segmentation images.

The LVQ segmentation isolated the brain in most slices and manual correction was used to remove any points connected to extracranial tissue. The brain was then detected using a flooding operation. Manual correction was based on the ground truth images, again simulating the human user. Following the removal of extracranial tissue, ACSR segmentation was performed. Dilated boundary maps were created from the LVQ segmentation using dilation factors of 1 and 3. The three volumes with varying levels of noise and inhomogeneity were segmented using PGA-SPD, PGA-SPDS and PGA-DPDS with full and partial ACSR. A comparison with the ground truth was carried out, producing an error rate defined as the ratio of misclassified pixels over the total number of pixels pertaining to the three segment classes (same metric was used by Pham and Prince in the AGEM [130] and AFCM [120] studies).

6.4.2. Results.

Table 6.1 shows the error rates and processing time for each variation of the PGA-PD and different levels of partial ACSR. Error rates for the initial LVQ segmentation are also shown. PGA-SPD with full ACSR is used as the baseline for processing time, while other variations and levels of partial ACSR are shown relative to the baseline.

PGA-SPD and PGA-SPDS consistently performed better than PGA-DPDS and faster. Error rates did not change much from full ACSR to partial ACSR with a dilation factor of 1. However the speed increase was significant. Counterintuitively partial ACSR with a dilation factor of 3 consistently performed better than full ACSR for all three variations of the PGA-PD and for all volumes. This is in spite of the LVQ segmentation error rates being consistently high. This phenomenon can be attributed to the fact that the LVQ segmentation produces good segmentation inside segments, sometimes better than ACSR, but introduces artefacts near the boundaries. Because of the gyri and sulci, boundaries account for a particularly large area in brain volumes. Therefore the speed increase using partial ACSR with a factor of 3 compared to full ACSR was small (similar to the cryo brain segmentations in chapter 5), but it still improved the accuracy of the segmentation.

Table 6.1. Error rates for segmentation of BrainWeb volumes with 3% Noise, 20% RF Inhomogeneity (3N 20RFI); 3% Noise, 40% RF Inhomogeneity (3N 40RFI); 7% Noise, 20% RF Inhomogeneity (7N 20RFI). ACSR segmentation using PGA-SPD, PGA-SPDS and PGA-DPDS with full ACSR and partial ACSR, dilation factors 1 (f1) and 3 (f3) and post-processing using median filter, compared to LVQ, AFCM (from [120]), EM and AGEM (from [130]).

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI	Rel. proc. time
SPD full	5.707%	7.733%	9.051%	1.00
SPD partial f3	5.257%	7.366%	8.739%	0.96
SPD partial f1	5.822%	7.798%	8.968%	0.67
SPDS full	5.754%	7.700%	9.168%	1.06
SPDS partial f3	5.294%	7.327%	8.858%	1.02
SPDS partial f1	5.871%	7.775%	9.072%	0.71
DPDS full	6.130%	8.018%	9.165%	1.13
DPDS partial f3	5.722%	7.686%	8.872%	1.07
DPDS partial f1	6.233%	8.081%	9.079%	0.75
SPD partial f3 median	6.565%	8.486%	8.522%	0.96
SPDS partial f3 median	6.532%	8.387%	8.407%	1.02
LVQ	9.720%	10.926	10.346%	N/A
AFCM	4.322%	4.938%	N/A	N/A
EM	5.487%	8.986%	10.699%	N/A
AGEM	4.144%	4.759%	8.414%	N/A

Post-processing using a median filter was performed for the two most successful variants (SPD and SPDS with partial ACSR, factor 3) and had a slight positive effect on error rates for the volume with the highest level of noise, while the drawbacks of losing image detail outweighed the benefits for volumes with higher Signal to Noise Ratio (SNR).

Table 6.1 also shows a comparison of ACSR with the results reported by Pham and Prince on the same volumes, using MRF segmentation with a standard EM algorithm and the Adaptive Generalized EM (AGEM) algorithm [130] as well as the fuzzy based AFCM [120]. It is apparent that ACSR performed consistently better than EM. However compared to AGEM and AFCM the error rates were higher for the volumes with 3% noise (AGEM/AFCM) but similar to the volume with 7% noise (AGEM). Fig. 6.1 shows an example of an original slice with the three different BrainWeb parameter settings used, its ground truth segmentation and the ACSR segmented equivalents. The results presented in this section were published in [174,175].

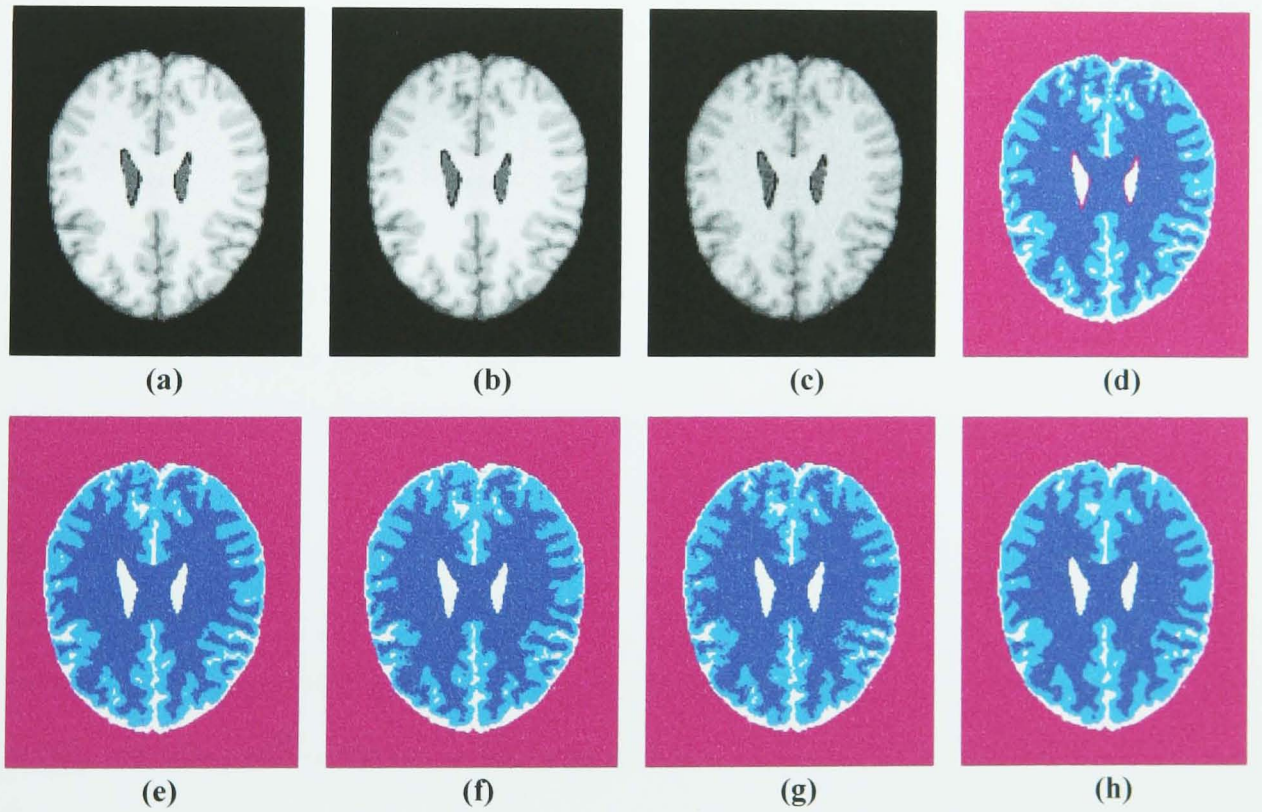


Fig. 6.1. A Slice from the BrainWeb volumes. (a) 3N 20RFI. (b) 3N 40RFI. (c) 7N 20RFI. (d) Ground truth image. Magenta areas outside the cortical surface and around the ventricles represent tissue not pertaining to any of the three classes CSF, grey matter and white matter. (e-g) Partial PGA-SPD with factor 3 dilation, segmentation of source images directly above into CSF, grey matter and white matter. (h) Same as (g) following post-processing with median filter.

6.4.3. Conclusion.

The results of comparison with the ground truth suggested that the two variations of the PGA with path descriptors using a small sampling window size (single path) consistently performed better and faster than the one using a larger size (double path). It is concluded from this that the single path representation is sufficiently rich and that the constraints on topology of the double path representation has a negative effect on accuracy. These constraints may be overcome using local optimisation, but clearly at the price of higher processing overhead. Shifting the seed point for the purpose of AID calculation gave marginally better results in some cases, but not all. The results show that the combination of the PGA with LVQ (partial ACSR) not only speeded up segmentation without significant loss of accuracy, but actually increased accuracy. The two best variations of the PGA combined with LVQ consistently performed better than MRF segmentation using a standard EM algorithm and similarly to AGEM at a high level of noise. A post-processing step using a median filter proved useful, but only at low SNR. The error rates at lower levels of noise are clearly outperformed by AGEM. This is not surprising given that the data is simulated and AGEM models intensity distributions in similar ways. However the increase in error rate from high to low SNR is relatively smaller for the PGA, which indicates robustness. Less robustness compared to AGEM is shown for increase of the level of RF inhomogeneity. Section 6.6 describes how inhomogeneity correction can be applied as a pre-processing step to significantly increase the robustness to this artefact. Similarly higher robustness to noise is desirable. Higher levels of noise call for a closer match of the templates with local image intensities, which can be achieved by including variations. This is possible in ACSR by explicitly templating classes across variations. It could not be achieved by the simulated user, because the template selection was simply based on all points labelled as the desired classes in slices at regular intervals. Section 6.5 describes how higher robustness to noise can be achieved through automatic template creation at every slice.

6.5. Introducing automatic template creation.

Although ACSR is capable of accommodating high levels of variation within segment classes, these variations have to be explicitly templated by the user. In cryo section data this is relatively easy, assuming that all images were acquired under the same conditions (lighting, camera settings, etc.). Variations within tissue types are clearly visible and can be selected and incorporated into class templates by the user. In MRI volumes however, RF inhomogeneities and noise make it extremely difficult to visually identify areas of significantly different visual representation within the same tissue type (even using colour look-up tables). This reduces the user's ability to select templates, which will generalise over subvolumes and increases error rates.

To counter these problems *automatic template creation* is introduced. The user still selects class templates, which are used to generate the learning data for the preliminary LVQ segmentation. In addition to a fully LVQ segmented image for each slice, an additional segmentation is now generated. Minimum quantisation errors for all image points are calculated and sorted *individually for each segment class* and all points in the upper 40% are discarded. The remaining 60% of all points for each class are encoded as class and slice-specific templates for the PGA and these templates are used in place of the original user defined templates in the final segmentation. This ensures a higher level of consistency and reduces overall error rates (see section 6.7 and fig. 6.4).

6.6. Incorporating inhomogeneity correction.

In an MRI scanner a strong magnetic field causes the spinning protons in the patient's body atoms to align [176]. A pulse of radio frequency magnetic field oscillations is injected using RF (Radio Frequency) transmitter coils in the scanner, the injected frequency corresponding to the frequency of the target nuclei (usually of Hydrogen atoms). The nuclei resonate and absorb energy. This energy is released again as a radio frequency signal when the injected pulse ends. The emitted signal is detected by RF receiver coils from which an image can be constructed. Due to non-uniformities in

the RF field produced by the transmitter coils and/or the RF field detected by the receiver coils and due to non-uniform loading of the coils by the patient's body (this can be reduced by using special coils constructed for the imaging of specific parts of the body), inhomogeneity artefacts are introduced in scan images. These intensity variations in the volume within the same type of tissue visually appear as smooth gradients across individual slice images. Inhomogeneity correction or modelling is essential to avoid serious errors in intensity based segmentation.

The results of two different inhomogeneity correction algorithms used with ACSR segmentation of MRI volumes have been employed and compared. The N3 algorithm [44,45] from the Brain Imaging Centre at McGill University was first published in 1997 and its implementation has since become part of a suite of software tools for MRI image processing available from the centre. The EQ algorithm developed by Marc Cohen et al [43] at the Brain Mapping Division, UCLA, is more recent and considerably simpler than N3. Cohen et al have however demonstrated significant improvements in segmentation accuracy for a number of intensity based algorithms.

6.6.1. The EQ inhomogeneity correction algorithm.

The plasticity of local neighbourhood representations in the PGA enables it to work well on unfiltered data, thus enabling the system to preserve original image features. Pre-filtering could result in the attenuation of intrinsic class features, which could take away the benefits of reducing the inhomogeneity artefacts. However an indirect filtering operation in the form of intensity equalisation applied to inhomogeneous MRI data could be highly beneficial. The EQ intensity equalisation algorithm for MRI volume data seemed a promising candidate for achieving this goal. An aggressive smoothing using a large Gaussian kernel (3/8 of the volume size) is applied to the volume using Fourier methods after the background has been filled with the average signal intensity (threshold for the background is automatically estimated based on histogram analysis). The smoothed volume is subsequently used to normalise the original volume, preserving the same average intensity. This operation is feature preserving and the background fill reduces boundary artefacts. The source code

implementing the algorithm is freely available for download from http://porkpie.loni.ucla.edu/BMD_ HTML/SharedCode/EQ/index.html.

The implementation used was based on the source code released 16th March 2001 (v. 1.13). A number of bugs had to be fixed in order to work on raw data rather than the commercial Analyze format and to get the Fourier transform to work properly with volumes of dimensions not of the order 2^n. The bugs were reported to and acknowledged by Marc Cohen.

6.6.2. The N3 inhomogeneity correction algorithm.

N3 (the Non-parametric intensity Non-uniformity Normalization algorithm) is based on a simple MRI image model:

$$v(x) = u(x)f(x) + n(x) \qquad \textbf{(6.5)}$$

At location x the measured signal is v, the true signal is u, f is an unknown bias field (causing inhomogeneity) and n is Gaussian white noise, considered to be independent of the true signal. High frequencies in the true signal are attenuated by the bias field, which is effectively causing a blurring in the measured signal. The kernel producing this blurring is considered to be approximately Gaussian. The true signal can then be estimated using a de-convolution and this produces a mapping from the measured signal to the estimated bias field. This estimate is smoothed by fitting a b-spline to the curve in order to iron out sharp jumps where tissue classes overlap. This is not achieved in a single iteration, so the measured signal is updated according to the estimated bias field and the process is repeated until convergence is reached. Alternatively since this process is computationally expensive, it may be stopped after a predefined number of iterations. The source code for N3 is freely available from http://www.bic.mni.mcgill.ca/software/N3.

The implementation of N3 was based on v. 1.05 of the source code with a few minor corrections for compatibility with Solaris. The following prerequisite libraries were also built/installed: netCDF v. 3.5.0, MINC v. 0.8 and the MNI perllib v. 0.05.

6.7. Results on simulated MRI data II.

The same BrainWeb volumes which were used for the initial study of the PGA with path descriptors were segmented with automatic template creation and inhomogeneity correction using either EQ or N3. The new segmentation pipeline is shown in fig. 6.2. The results presented in this and the following section (6.8) were published in [177].

Table 6.2 shows the error rates for the three volumes using PGA-SPD and PGA-SPDS segmentation with automatic template creation and inhomogeneity correction using EQ. Table 6.3 shows the equivalent error rates using N3. It is apparent that PGA-SPDS consistently performed better than PGA-SPD and N3 consistently resulted in more accurate results than EQ. In the remainder of this section "PGA" will refer to the stand-alone PGA-SPDS, while "PGA auto" will refer to PGA-SPDS using automatic template creation. Fig. 6.3 shows an example of PGA auto on a slice from the volume with 3% noise and 40% inhomogeneity.

Table 6.2. Error rates for BrainWeb volumes using EQ inhomogeneity correction and automatic template creation.

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI
PGA-SPD EQ	5.155%	5.655%	7.963%
PGA-SPDS EQ	5.104%	5.583%	7.901%

Table 6.3. Error rates for BrainWeb volumes using N3 inhomogeneity correction and automatic template creation.

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI
PGA-SPD N3	4.697%	4.413%	7.721%
PGA-SPDS N3	4.661%	4.396%	7.670%

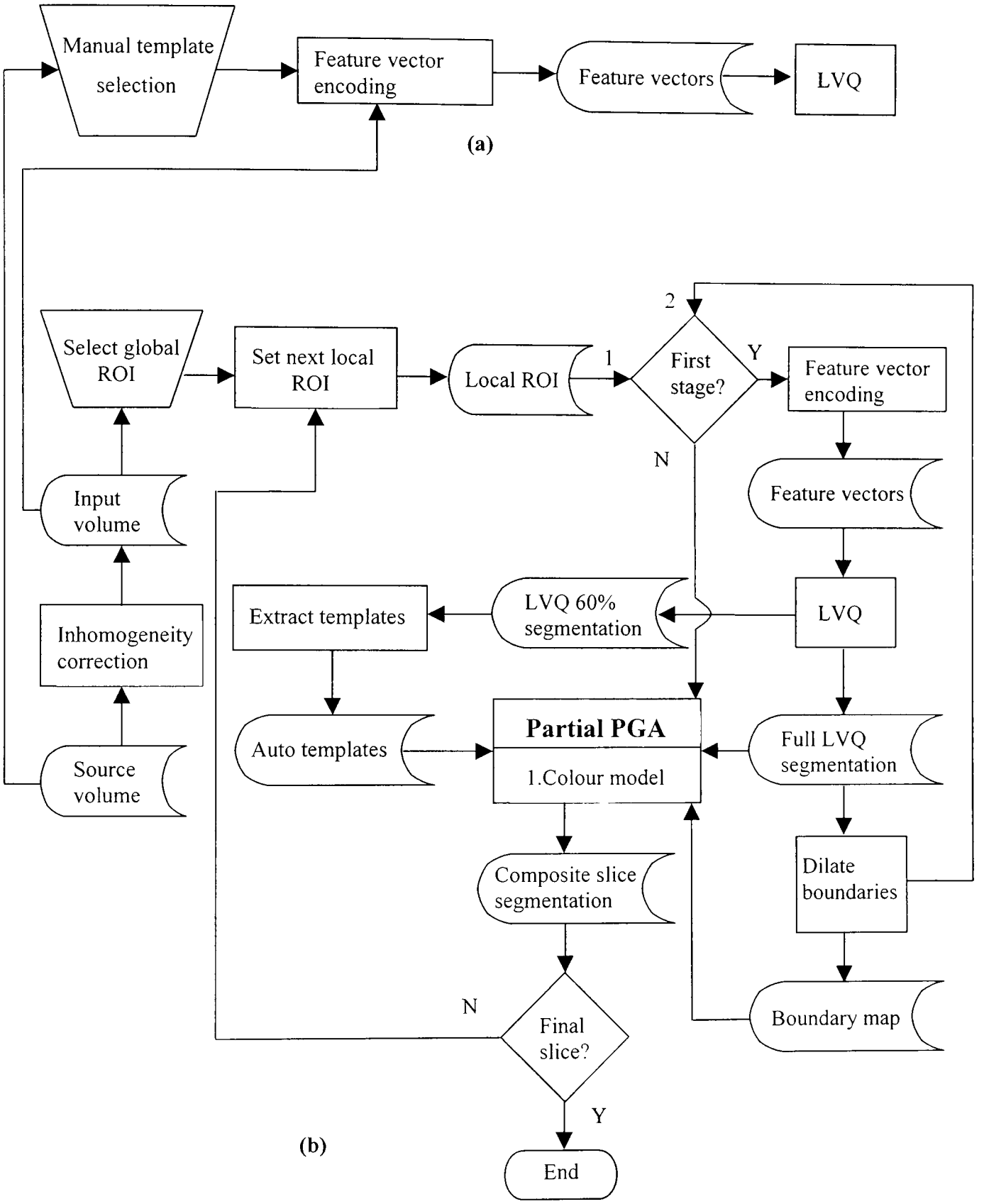


Fig. 6.2. The partial ACSR volume segmentation pipeline for MRI data. (a) Template selection from source volume and training of LVQ classifiers. (b) Selection of global ROI followed by segmentation of local ROIs from inhomogeneity corrected volume for each slice plane. Local ROIs are 2D in pseudo-3D segmentation. Initial full LVQ segmentation and extraction of automatic templates from 60% best classification for each class. Followed by partial PGA segmentation, using automatic templates, producing the composite final segmentation.

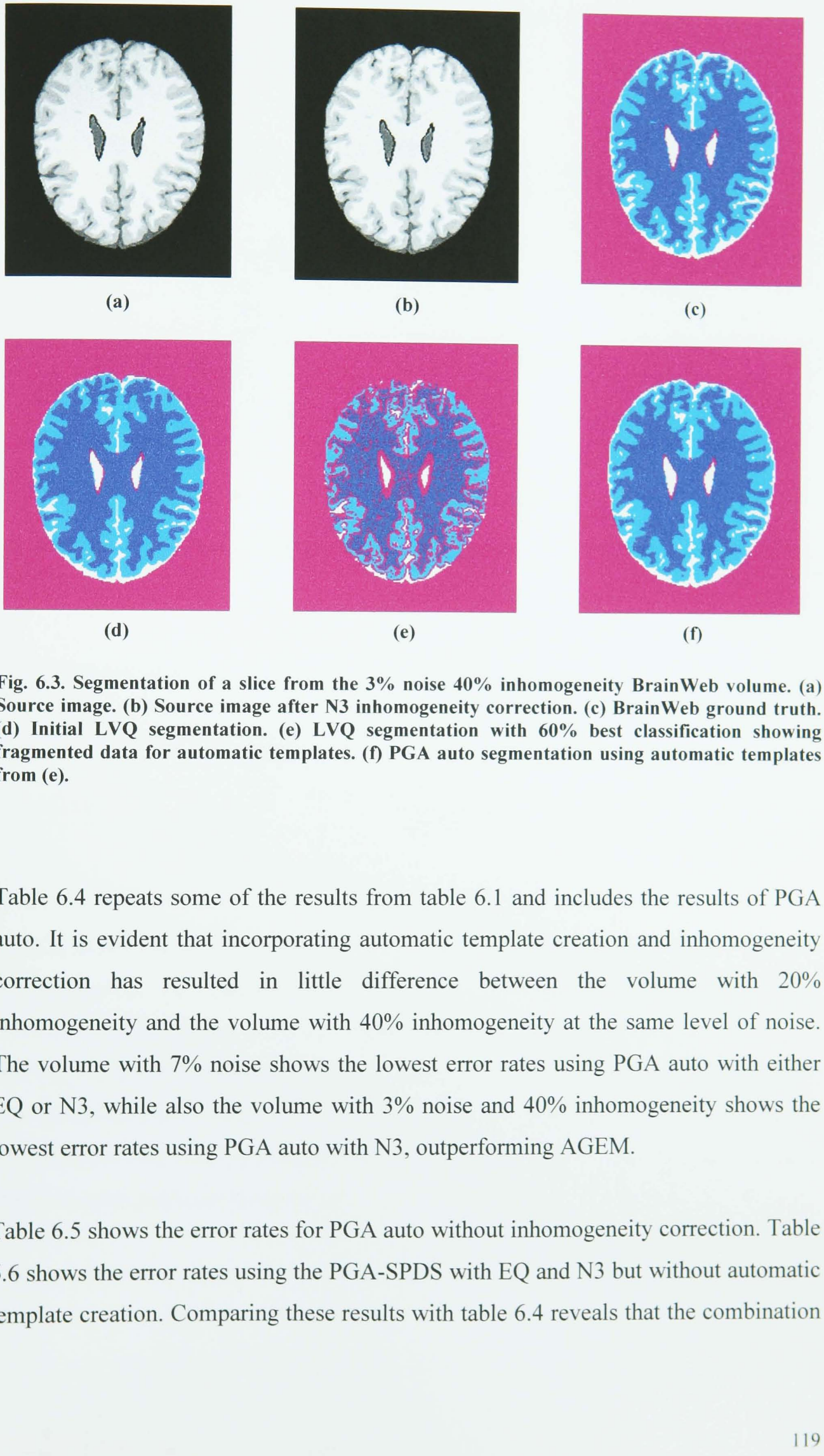


Fig. 6.3. Segmentation of a slice from the 3% noise 40% inhomogeneity BrainWeb volume. (a) Source image. (b) Source image after N3 inhomogeneity correction. (c) BrainWeb ground truth. (d) Initial LVQ segmentation. (e) LVQ segmentation with 60% best classification showing fragmented data for automatic templates. (f) PGA auto segmentation using automatic templates from (e).

Table 6.4 repeats some of the results from table 6.1 and includes the results of PGA auto. It is evident that incorporating automatic template creation and inhomogeneity correction has resulted in little difference between the volume with 20% inhomogeneity and the volume with 40% inhomogeneity at the same level of noise. The volume with 7% noise shows the lowest error rates using PGA auto with either EQ or N3, while also the volume with 3% noise and 40% inhomogeneity shows the lowest error rates using PGA auto with N3, outperforming AGEM.

Table 6.5 shows the error rates for PGA auto without inhomogeneity correction. Table 6.6 shows the error rates using the PGA-SPDS with EQ and N3 but without automatic template creation. Comparing these results with table 6.4 reveals that the combination

of inhomogeneity correction and automatic template creation always performed better than either one or none of the two optimisations.

Table 6.4. Summary of results: EM and AGEM [130] compared to PGA-SPDS with manual templates and no inhomogeneity correction and PGA auto with EQ and N3 inhomogeneity correction.

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI
PGA-SPDS	5.294%	7.327%	8.858%
PGA auto EQ	5.104%	5.583%	7.901%
PGA auto N3	4.661%	4.396%	7.670%
EM	5.487%	8.986%	10.699%
AGEM	4.144%	4.759%	8.414%

Table 6.5. Error rates of PGA auto without inhomogeneity correction.

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI
PGA auto	5.161%	6.725%	7.972%

Table 6.6. Error rates of PGA-SPDS based on manual templates with EQ and N3 inhomogeneity correction.

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI
PGA-SPDS EQ	5.156%	5.684%	8.874%
PGA-SPDS N3	4.742%	4.483%	8.750%

Fig. 6.4 shows the accuracy (inverse of the error rate) per individual slice plotted for segmentation of the volume with 7% noise and 20% inhomogeneity, using PGA-SPDS (manual templates), PGA auto with no inhomogeneity correction and PGA auto with EQ and N3. While the segmentation using manual templates directly has distinct peaks at the templated slices, the level of accuracy is considerably more consistent for the segmentations using automatic template creation. The mean accuracy is also plotted in each graph for comparison.

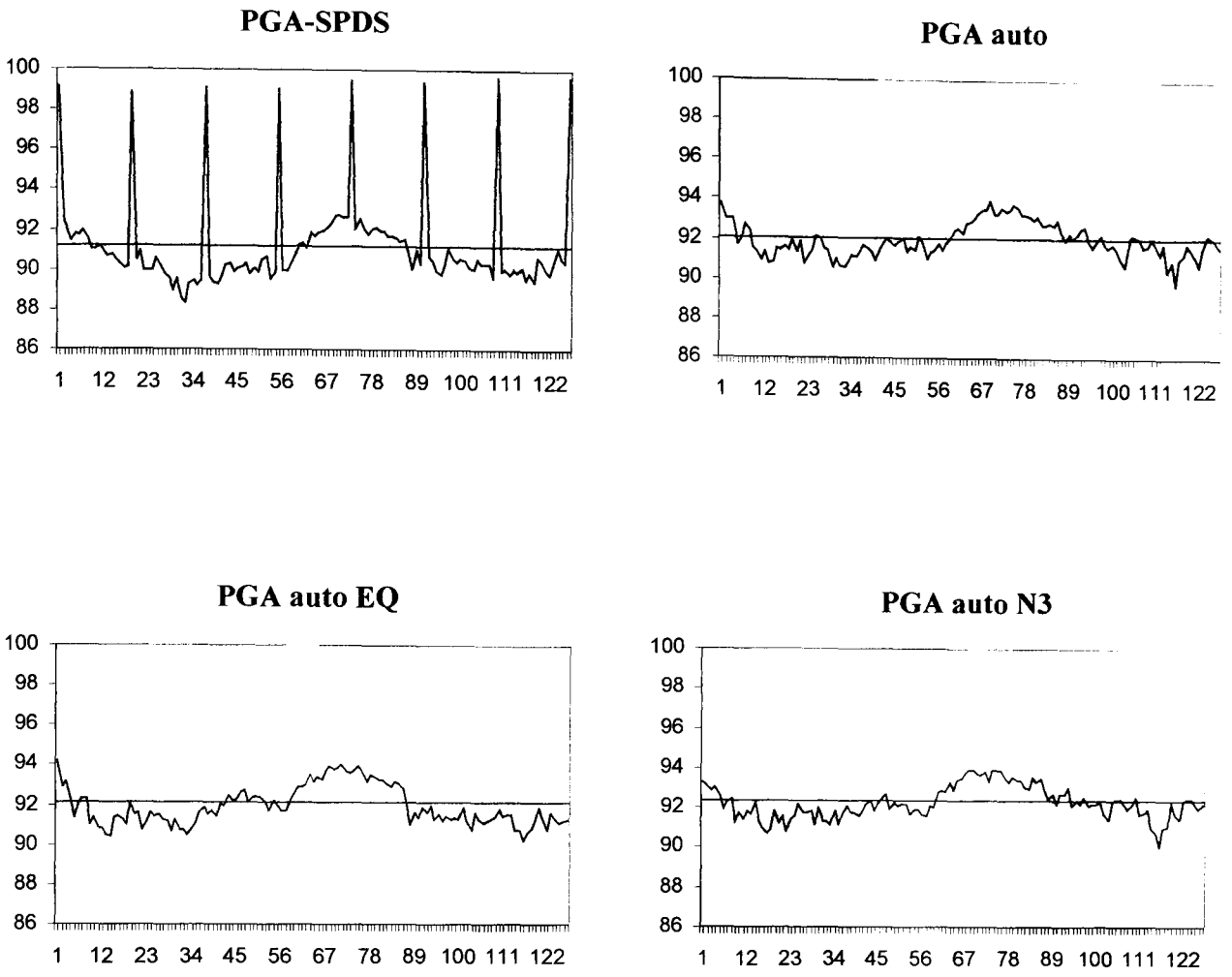


Fig. 6.4. Graphs of segmentation accuracy for the 7N 20RFI volume, expressed in % overlap (y-axis) with the ground truth for CSF, grey matter and white matter. The blue line shows segmentation accuracy per slice (x-axis) using manual templates directly (PGA-SPDS), automatic template creation with no inhomogeneity correction (PGA auto) and with EQ and N3 inhomogeneity correction (PGA auto EQ and PGA auto N3). The red line shows the mean accuracy for each segmentation. The segmentation using manual templates directly shows best segmentation for the slices immediately surrounding the eight templated slices. In the segmentations using automatic template creation with or without inhomogeneity correction, the level of accuracy is more constant throughout.

Due to longer acquisition times, it is often not practical in a clinical application to obtain volumes with 1mm slice thickness. 3mm or 5mm is more commonly used. This could prohibit volume segmentation from being an advantage over pseudo-3D segmentation. All previous results in this chapter are based on per slice pseudo-3D segmentation, but for the sake of completeness, volume segmentation of the three BrainWeb volumes was performed. The slice thickness was only 1mm in these volumes, similarly to the cryo sections described in chapter 4. In order to keep processing time at a reasonable level the volume segmentation was restricted to growth into only the two neighbouring slice planes, allowing paths to include voxels

in the plane as well as at a distance of 1 voxel away along the Z axis. Table 6.7 shows the error rates for PGA auto, which are only very marginally lower than the error rates using pseudo-3D segmentation. Processing time however was quadrupled.

Table 6.7. Error rates of restricted isovolume PGA auto with EQ and N3 inhomogeneity correction.

Algorithm	3N 20RFI	3N 40RFI	7N 20RFI
PGA auto EQ	5.045%	5.526%	7.866%
PGA auto N3	4.573%	4.296%	7.630%

A surprising result was that while both N3 and EQ reduce the error rates for the volume with 40% inhomogeneity to the same level as that of the volume with 20% inhomogeneity, the use of N3 resulted in very slightly better results for the volume with 40% inhomogeneity (using pseudo-3D as well as isovolume segmentation). The opposite, which was indeed the case when using EQ, would have been expected. Table 6.8 compares the results of PGA auto with no inhomogeneity correction, using EQ and using N3 on a BrainWeb volume with 3% noise and 0% RF inhomogeneity. It is evident that the volume pre-processed with N3 actually performed very marginally better than the volume with no pre-processing, but worse than the N3 pre-processed volume with 40% inhomogeneity. The volume pre-processed with EQ showed the opposite, as expected. A possible explanation for the effect of N3 could be that the bias field estimate is more accurate at higher levels of inhomogeneity, at least in simulated data, resulting in an overall improvement in segmentation accuracy. In real data it would be unlikely to see such an effect.

Table 6.8. Error rates of PGA auto on a BrainWeb volume with 3% noise and 0% RF inhomogeneity using EQ, N3 or no inhomogeneity correction.

Algorithm	No correction	EQ	N3
PGA auto	4.356%	4.838%	4.352%

6.8. Multispectral MRI segmentation.

Following successful applications of ACSR to colour images and single channel MRI images, the obvious next step would be an investigation into multispectral MRI segmentation. T1, T2 and Proton Density images all have their advantages and drawbacks for different tissue types. The combination of two or more acquisition modes could potentially improve segmentation results [130,178] by providing richer point descriptors. Path representations for the PGA can use any number of descriptors. However while descriptors produced by a suitable colour model in natural colour images relate to each other in ways which are well understood, the relations between multiple channels in multispectral MRI data sets are less well defined. The combination of descriptors from multiple acquisition modes and the possible development of a multispectral appearance model for MRI data is an interesting problem, but outside the scope of this project. To provide some preliminary results a more simple type of multispectral MRI segmentation will be considered here.

Table 6.9. Class error rates for PGA auto using EQ inhomogeneity correction. Multispectral segmentation based on T1 and T2 images.

Class	3N 20RFI	3N 40RFI	7N 20RFI
CSF	1.390%	1.422%	3.659%
Grey matter	8.514%	9.267%	10.636%
White matter	4.657%	5.321%	7.374%
All	5.889%	6.485%	8.251%

Table 6.10. Class error rates for PGA auto using N3 inhomogeneity correction. Multispectral segmentation based on T1 and T2 images.

Class	3N 20RFI	3N 40RFI	7N 20RFI
CSF	1.984%	1.838%	3.782%
Grey matter	8.111%	8.339%	10.676%
White matter	2.811%	2.296%	6.326%
All	5.134%	5.028%	7.912%

Table 6.11. Class error rates for PGA auto using EQ inhomogeneity correction. Single-channel segmentation based on T1 images.

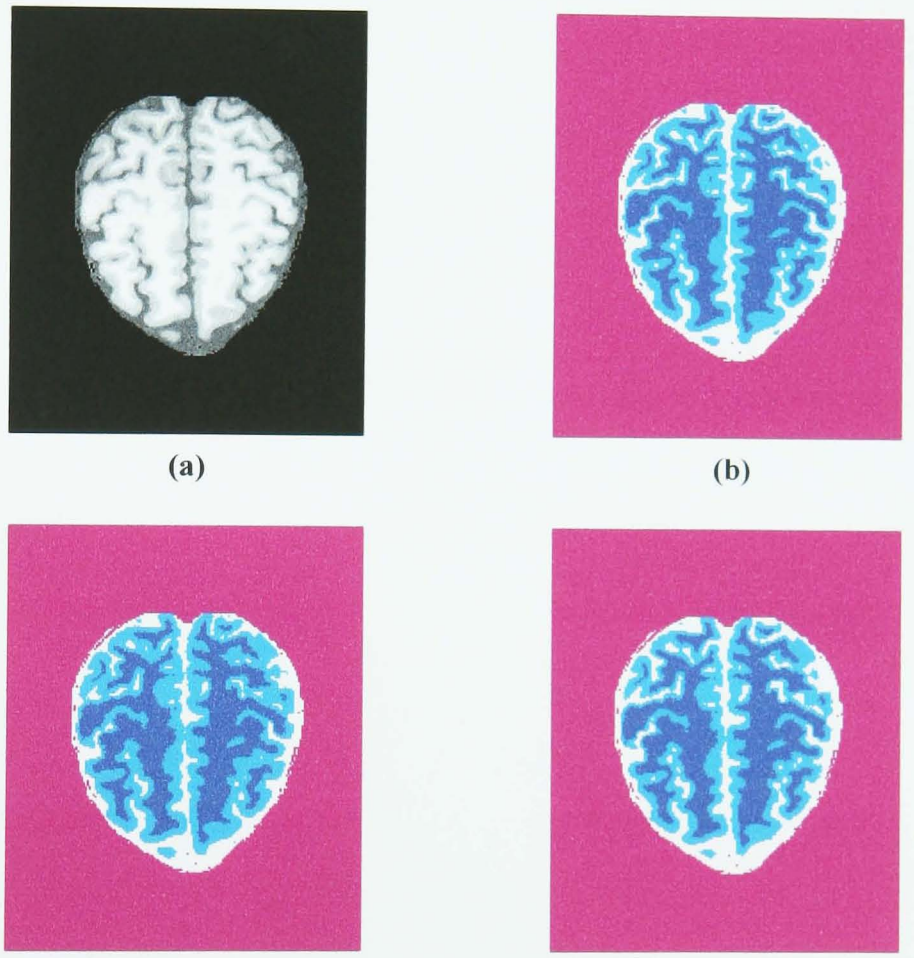
Class	3N 20RFI	3N 40RFI	7N 20RFI
CSF	8.154%	8.668%	11.426%
Grey matter	3.744%	4.048%	6.637%
White matter	5.401%	6.086%	7.850%
All	5.104%	5.583%	7.901%

Table 6.12. Class error rates for PGA auto using N3 inhomogeneity correction. Single-channel segmentation based on T1 images.

Class	3N 20RFI	3N 40RFI	7N 20RFI
CSF	9.361%	9.450%	11.922%
Grey matter	3.748%	3.640%	6.749%
White matter	3.599%	2.965%	6.832%
All	4.661%	4.396%	7.670%

T1-weighed images show good contrast between grey and white matter, while fluids (including CSF) are very well defined on T2-weighed images. The three BrainWeb volumes were segmented using LVQ learning vectors and PGA templates generated from T2 images for CSF and T1 images for grey and white matter. Using a specific mode for a specific class, different modes can simply be regarded as extra dimensions in the sampling space. Both modes were thus used during segmentation, but were transparent to the algorithm. Tables 6.9 (using EQ) and 6.10 (using N3) show the error rates for individual classes based on T1 and T2, while tables 6.11 (using EQ) and 6.12 (using N3) show the error rates based only on T1 images. From this purely quantitative type of evaluation, it appears that error rates for CSF are remarkably low using multispectral segmentation, while there is a substantial increase in error rate for grey matter. The explanation is that CSF is well represented when T2 images are used, but invades grey matter regions according to the ground truth segmentation in these data sets, significantly affecting error rates. Because CSF accounts for the smallest volume of the three tissue classes, the overall error rate is roughly the same as for single-channel segmentation. However when visually comparing individual slices (such as in fig. 6.3) as a qualitative measure, the more accurate segmentation of CSF clearly appears as an improvement and grey matter still appears to be well represented. In chapter 5 it was established that while it is reasonable to use ground

truth evaluation in simulated data, it does not necessarily reflect the *effectiveness* of the segmentation as perceived by a human user. Other performance metrics than overlap in area and volume could be employed, for example the Hausdorff distance [179] might be used to quantify the accuracy of boundary location in a segmented volume compared to the ground truth. However although similarity can be objectively quantified, the loss of crucial visual information versus the loss of redundant visual information in a specific task for a specific target user cannot be.



(a) (b)

(c) (d)

Fig. 6.5. Single-channel and multispectral segmentation of a slice from the 3% noise 40% RF inhomogeneity volume. (a) Source image. (b) BrainWeb ground truth. (c) PGA auto segmentation with N3 inhomogeneity correction, single-channel T1. (d) PGA auto segmentation with N3 inhomogeneity correction, multi-channel T1 and T2.

6.9. Summary.

This chapter has introduced a PGA suitable for ACSR segmentation of MRI volumes, and two optimisations significantly increasing the robustness of the method to common artefacts. It was shown that an automatic template creation generated from

the initial LVQ segmentation resulted in higher robustness to noise. Inhomogeneity correction as a pre-processing stage, using the EQ and N3 algorithms, increased the accuracy of segmentation for volumes with high levels of inhomogeneity. Results compared favourably to previously published results on MRF segmentation of the same volumes. Better results were achieved for N3 corrected data than for EQ. However it cannot be concluded that these results on simulated data would transfer to real data. This as well as the issue of single-channel and multispectral segmentation is investigated further using human observer experiments described in the following chapter.

Chapter 7

Evaluating the quality and robustness of ACSR segmentation

7.1. Empirical evaluation of ACSR segmentation through human observer experiments.

The two previous chapters have presented results on natural colour image segmentation and MRI segmentation based on ground truth evaluation. For the natural scenes and cryo sections in chapter 5, the ground truth was manually generated by tracing the boundaries of segments in the source images. Segmentations were based on one set of templates for each image or image series. In chapter 6 the source data itself was generated from the ground truth with varying parameters. Segmentations were based on automatically generated templates from an original single set of templates for each volume, after one of two different types of pre-processing for inhomogeneity correction had been applied to the source data.

For ground truth evaluation of real data segmentation, a ground truth must be manually generated by expert observers. In model-based simulated data, the image construction itself specifies the ground truth exactly and allows for the ability to vary parameters (such as noise and inhomogeneity in the BrainWeb data). Ground truth comparison as a standard performance metric used on such data gives perfect reproducibility, but the evaluation of model-based algorithms on model-based data may not reflect the performance on real data. Testing on real data on the other hand is complicated by the subjectivity of a manually generated ground truth. The computer based ground truth comparison is an ideal observer, when the goal is to quantify the level of artefacts present in a segmentation, and an artefact is defined as a single point in the automatic segmentation labelled differently from the corresponding point in the ground truth segmentation. However this is not always an ideal measure of segmentation quality in applications where the overall visual representation conveyed by a segmentation is key. A human expert observer (such as a radiologist) is not

capable of the exact quantification of artefacts due to the limitations of the human visual system, but is capable of a higher level interpretation, grounded in expert knowledge about the expected anatomy, the specific application domain and an acquired robustness to common image artefacts.

Heath et al [160] conducted a study in which edge detections of images of natural scenes were ranked by a group of experienced observers (the observers all had a computer vision background). The edge detections were performed using five different algorithms. The ideal parameter settings for each algorithm were determined by a smaller number of observers in an initial ranking task. The Intra-Class Correlation coefficient (ICC) was used to demonstrate that observers ranked highly consistently between them. Ranking used an ordinal scale from 1 to 7 and was based on participants arranging printed edge detection images on a table in front of them in order of preference and giving each image a grade. The study demonstrated significant differences between the observed performance of the algorithms when applied to different categories of images, without the use of a manually generated gold standard ground truth.

In this chapter a series of human observer experiments for the qualitative evaluation of ACSR segmentation are described. They involved visual ranking of segmentations by naïve as well as specialised expert observers. Comparison between the qualitative results obtained and previous results on ground truth evaluation are given. It is demonstrated how changing some parameters caused the observed quality of segmentation to follow the level of artefacts detected using ground truth comparison. Changing other parameters resulted in a significant difference in observed segmentation quality, while ground truth evaluation showed no such difference or even favoured an alternative parameter setting. The significance of these results for different applications is discussed. The issues of RGB vs. OPC in colour segmentation, single-channel vs. multispectral segmentation of MRI volumes and EQ vs. N3 for inhomogeneity correction are investigated along with the effect of different initialisations of the same segmentation task. Written material given to participants and an overview of the computer based experiments with screen shots can be found in

appendix D. The images and segmentations described in this chapter are included on the companion CD.

The segmentation tasks investigated in the experiments, like the tasks in the previous chapters, were segmentations into general segment classes. Without knowing exactly what the application is, it can be an ambiguous task for observers to appreciate what a perfect representation of a given segment class is in a given image. Therefore segmentation quality was defined to participants as a *minimisation of obvious representational errors*. Robustness in the context of these experiments was defined as the ability for ACSR segmentation to produce consistent results given multiple initialisations with the same segmentation parameters.

In every experiment a *relative ranking* was produced by observers arranging segmentations in order of preference according to the observed quality. The relative positions were converted to grades on an ordinal scale from 1 (lowest) to n (highest), where n was equivalent to the total number of segmentations. Subsequently an *absolute ranking* was given to each segmentation by observers using an ordinal scale from 1 to 7 (similar to the study by Heath et al [160]). A grade of 1 meant *"image shows no coherent representation of your perceived ideal segmentation"* while a grade of 7 meant *"image matches your perceived ideal segmentation"* (see appendix D). Several segmentations could be given the same grade even if they were distinguished in the relative ranking task. The idea was not for participants to necessarily use the full absolute scale, but to pinpoint objectively how well each individual image represented the desired segments.

ICC and chi-square tests were performed using SPSS v. 0.9.1 from SPSS Incorporated. U tests and H tests were performed using Minitab v. 13.1 from Minitab Incorporated.

In the following, "DSOT St. Mary's" will refer to the Department of Surgical Oncology and Technology, St. Mary's Hospital, London, U.K. "RSU Hammersmith" will refer to the Radiological Services Unit, Hammersmith Hospital, London, U.K.

7.2. Qualitative evaluation of natural colour image segmentation.

The six natural colour images shown in fig. 5.4 and their segmentations using PBNN-RGB, PBNN-OPC, PGA-RGB, PGA-OPC and LVQ were used in an experiment to evaluate segmentation quality through visual ranking by observers with no experience in image segmentation. The aim was to show if a group of inexperienced observers would be able to consistently rank segmentations of ordinary scenes in standard colour test images, and whether or not the results would be consistent with those obtained using ground truth comparison (chapter 5, section 5.3.1). It was of particular interest to establish if the subtle differences in segmentations based on OPC and RGB would have a significant influence on the observed quality by human observers.

7.2.1. Methods.

A group of 11 computing science students from the School of Computing Science, Middlesex University, London, U.K. (3 final year BSc and 8 MSc students) was used. All participants had to pass a short version (using six patterns) of Ishihara's standard test for colour blindness [180] before proceeding to the main task. Participants were first shown two examples of ACSR segmentation using the eagle image from [157] and the polar bear image from [161] on an LCD display. The position of the screen was adjusted for optimal viewing. Each of the six images to be evaluated were then presented in turn in random order on the screen. For each image the participant was told what the segmentation task was and what the desired segments were. Printed and laminated versions of the five segmentations were then laid out in front of the participant in random order. Individual images could be identified and their order recorded after the experiment through a mark on the back, which was not visible on the front. The participant was required to arrange the segmentation images in order of perceived segmentation quality, minimising the error of representation of the desired segments. *Unclassified* points in the PBNN images were to be regarded as errors no better or worse than *misclassified* points in any of the segmentations. Images were to be placed with the lowest quality to the far left and progressively better to the right with the best segmentation to the far right. The participant was allowed to place two images in the same position if the quality of segmentation could not be distinguished.

Subsequently an absolute ranking was given to each segmentation. After the relative and absolute ranking tasks the next source image would be displayed on the screen and its five segmentations presented. The same display device and the same printed segmentation images were used for all participants.

7.2.2. Results.

The results of the relative and the absolute rankings were analysed separately. In the relative ranking each image was given a value according to its relative position. The value 1 corresponded to the least good segmentation and the value 5 to the best segmentation. Sets of segmentations, which were indicated by a participant as not being distinguishable, were each given a value equal to the average of the current and the next position. Table 7.1 shows the summed values of the relative rankings by all participants for the six images. Overall LVQ was rated as the least good segmentation, followed by PBNN-RGB, PBNN-OPC, PGA-RGB and finally PGA-OPC as the best segmentation. On the image level the OPC based segmentation was rated better than the RGB based segmentation in 5 out of 6 images for PBNN and in 4 out of 6 images for PGA (one tie). The largest difference could be found for the Bird11 and Bird12 images as predicted in chapter 5, section 5.3.1.

Table 7.1. Summed relative rankings of segmented natural colour test images for all subjects.

Image	PBNN-RGB	PBNN-OPC	PGA-RGB	PGA-OPC	LVQ
Poppy	21.5	27.5	46.5	46.5	23
Hearts	21	26	50.5	48.5	19
Seagull	32	25.5	44	48	15.5
Rose	25	30	49	50	11
Bird11	26	29	46.5	52.5	11
Bird12	11	27	44	55	28
All images	**136.5**	**165**	**280.5**	**300.5**	**107.5**

Table 7.2 shows the summed values of the relative ranking for all images for each participant. From these values the ICC was calculated to determine if participants were consistent in their rankings. The ICC(3,*k*) version of the statistic was used because a single group of subjects rated all images. The number of subjects *k* was 11.

The ICC(3,11) for the relative rankings was found to be 0.996, which means that the participants were highly consistent in their rankings and shared a "goodness of segmentation" similar to what Heath et al [160] found.

Table 7.2. Summed relative rankings of segmented images for all natural colour test images (rows correspond to subjects).

Sample	PBNN-RGB	PBNN-OPC	PGA-RGB	PGA-OPC	LVQ
1	11.5	13	26	27.5	12
2	12	14	22	29	13
3	16	17	22.5	25.5	9
4	12	15	27	27	9
5	11	14	27	27	11
6	11	16	26	28	9
7	13	13	26.5	27.5	10
8	12	15.5	25	27	10.5
9	13	16	27	26.5	7.5
10	12	16	27	27	8
11	13	15.5	24.5	28.5	8.5

The absolute rankings were treated in the same way as the relative rankings. Table 7.3 shows the mean summed values of the absolute rankings from all participants for the six images. The total mean of 5.9 for PGA-OPC indicates a clear correspondence between the observed segmentations and the perceived ideal segmentations.

Table 7.3. Mean summed absolute rankings of segmented natural colour test images for all subjects.

Image	PBNN-RGB	PBNN-OPC	PGA-RGB	PGA-OPC	LVQ
Poppy	3.3	3.6	5.8	6.0	3.6
Hearts	2.7	2.8	5.8	5.6	2.4
Seagull	4.9	4.5	6.1	6.3	3.8
Rose	3.5	4.0	6.1	6.0	1.8
Bird11	3.1	3.5	5.1	5.4	1.9
Bird12	2.5	3.3	5.5	6.2	3.6
All images	**3.3**	**3.6**	**5.7**	**5.9**	**2.9**

Table 7.4 shows the summed values of the absolute ranking for all images for each participant. The ICC(3,11) was calculated as 0.991, again showing a very high consistency in rankings by the participants. The ICC(3,11) was also calculated for each individual image over all five classifiers for both the relative and the absolute ranking. Segmentations of the Seagull image were very close to each other while in

comparison they were vastly different for the Hearts image. It was therefore possible that a low correlation of rankings would show for some of the images. It is evident from table 7.5 that this was not the case. All images showed a high correlation, although as expected the Seagull image was below average.

Table 7.4. Summed absolute rankings of segmented images for all natural colour test images (rows correspond to subjects).

Sample	PBNN-RGB	PBNN-OPC	PGA-RGB	PGA-OPC	LVQ
1	21	23	34	37	20
2	17	20	33	38	18
3	22	23	30	33	17
4	12	17	34	34	12
5	21	23	37	37	20
6	14	17	31	34	10
7	34	34	42	42	34
8	20	22	31	30	19
9	20	22	37	37	13
10	19	18	35	35	13
11	19	21	31	33	13

Table 7.5. ICC(3,11) for each natural colour test image over all classifiers based on relative and absolute ranking.

Image	ICC(3,11) relative ranking	ICC(3,11) absolute ranking
Poppy	0.917	0.912
Hearts	0.974	0.978
Seagull	0.933	0.899
Rose	0.992	0.986
Bird11	0.991	0.978
Bird12	0.995	0.984

Because the rankings could not be guaranteed to be normally distributed, the non-parametric Mann-Whitney U test was used to test individual segmentations against each other. A number of two-tailed tests were carried out. The null-hypothesis was no difference between the RGB and the OPC based segmentations. Table 7.6 shows the results of these tests. Based on the relative rankings the null-hypothesis could be rejected and the results were highly significant at the $p = 0.05$ level, showing that PBNN-OPC and PGA-OPC produced superior results to PBNN-RGB and PGA-RGB. Based on the absolute rankings the results were not significant. LVQ (which used OPC) was compared to PBNN-OPC, and PBNN-OPC was found to be significantly better based on both the relative and the absolute rankings.

Table 7.6. Results of two-tailed Mann-Whitney U tests for relative and absolute rankings of RGB and OPC based segmentations of natural colour test images. Significance at p = 0.05.

Classes	Relative rankings	Absolute rankings
PBNN-RGB vs. PBNN-OPC	p = 0.0013 **significant**	p = 0.1756 not significant
PGA-RGB vs. PGA-OPC	p = 0.0064 **significant**	p = 0.3031 not significant
PBNN-OPC vs. LVQ	p = 0.0001 **significant**	p = 0.0175 **significant**

7.2.3. Discussion.

The effect of varying the segmentation algorithm was the same for the visual ranking as it was for the ground truth evaluation. The PBNN classifier outperformed LVQ because it uses more undistorted local information, but was in turn outperformed by the adaptable region based PGA. Effectively changing the algorithm in this experiment produced a scaling of the amount of artefacts distributed across segmentations, similar to adding increasing levels of impulse noise. Thus in agreement with the ground truth evaluation, the level of artefacts was inversely proportional to the observed quality. Based on ground truth comparison there was no significant effect of changing the colour model. However the visual ranking showed significantly better results using OPC. This is in spite of the ground truth comparison actually showing higher error rates for some of the OPC based segmentations. It is clear that in this case, changing the parameter changed not the level of artefacts, but the nature of the artefacts and how they affected the visual information conveyed in the segmentations. This was missed by the ground truth evaluation. If several such parameters are present, the effect on a segmentation can be significant, particularly in applications for accurate 3D reconstruction.

The relative ranking was needed to show that the difference between RGB and OPC was significant. It could be argued that the absolute ranking would have achieved the same goal, had a higher resolution of the scale been used. However more grades to choose from would complicate the rating task for participants. On the other hand the absolute ranking was useful to quantify not the order of preferred segmentations but a rating of their independent quality. This would not be possible if the full scale was used. The results obtained from the absolute ranking showed a very large difference

between LVQ and PGA-OPC (PGA-OPC on average rated twice as high as LVQ) while the difference based on ground truth evaluation was small.

7.3. Qualitative evaluation of a brain cryo section series segmentation.

Having established in the experiment described in the previous section that inexperienced human observers were able to highly consistently rank segmentations of natural colour images, the next step was to investigate if the same could be shown for specialised observers ranking medical images. In the first instance the five brain cryo slices from chapter 5, section 5.3.2 were used. Again the aim was to test for consistency of the rankings and to compare results to those obtained using ground truth comparison, as well as to obtain an overall measure of observed segmentation quality.

7.3.1. Methods.

The observers were 11 surgeons from St. Mary's Hospital. These participants were selected because of their theoretical and practical experience in assessing macroscopic human anatomy and as expert users of medical images. Since these surgeons regularly carry out keyhole surgery, they had a possible additional advantage of more experience in assessing 2D close-up colour images of anatomical structures. Although the type of images viewed in keyhole surgery procedures may not be directly comparable to cross-sections, they still present the surgeon with the task of determining the delineation of tissue types based on natural colour and texture.

The images were presented in sequence on a CRT display. The source images appeared in the top row and each of the five segmentations of the image sequence based on PBNN-RGB, PBNN-OPC, LVQ, PGA-RGB and PGA-OPC appeared in rows underneath. The starting order of the rows for each segmentation was randomly selected for each participant. Each column corresponded to a specific image in the sequence. Using an interactive selection tool each participant was able to arrange the

rows in order of perceived segmentation quality, placing the least good sequence as the bottom row with sequences getting progressively better towards the top of the screen. Participants were thus performing a ranking based on the overall quality of a sequence of five images and not on individual images. Following this relative ranking, participants graded each row using the same scale from 1 to 7 as employed for the images in section 7.2.

7.3.2. Results.

The ICC(3,11) was calculated as 0.920 for the relative ranking and 0.926 for the absolute ranking, showing a high correlation between subjects.

The summed relative rankings are shown in table 7.7 and the mean absolute rankings in table 7.8. Table 7.9 shows the results of Mann-Whitney U tests between PBNN-RGB and PBNN-OPC, PGA-RGB and PGA-OPC and finally PBNN-OPC and LVQ.

Table 7.7. Summed relative rankings of the segmented cryo brain volume for all subjects.

Volume	PBNN-RGB	PBNN-OPC	PGA-RGB	PGA-OPC	LVQ
Cryo brain	25	18	45	48	29

Table 7.8. Mean absolute rankings of the segmented cryo brain volume for all subjects.

Volume	PBNN-RGB	PBNN-OPC	PGA-RGB	PGA-OPC	LVQ
Cryo brain	3.1	2.9	5.0	4.8	3.3

Table 7.9. Results of two-tailed Mann-Whitney U tests for relative and absolute rankings of RGB and OPC based segmentations. Significance at p = 0.05.

Classes	Relative rankings	Absolute rankings
PBNN-RGB vs. PBNN-OPC	p = 0.0725 not significant	p = 0.7678 not significant
PGA-RGB vs. PGA-OPC	p = 0.1666 not significant	p = 0.6550 not significant
PBNN-OPC vs. LVQ	p = 0.0403 **significant**	p = 0.4710 not significant

7.3.3. Discussion.

Similar to the experiment on natural colour images, the order of preferred segmentation algorithms in the visual ranking was predicted by the ground truth evaluation. In this case the LVQ classifier produced better results than PBNN, because it generalises over the data, while matching purely local data does not perform well on multiple slices using PBNN. LVQ still showed artefacts due to the rigid sampling window, while the adaptive representations of the PGA produced superior results. Contrary to the results on the natural colour images, this experiment showed no significant benefit of using OPC over RGB. The average absolute grade given to PGA-OPC for this volume was 4.8, which is considerably lower than the 5.9 given for the natural colour images. However this was to be expected, considering the inhomogeneity of the "other" class (classes were grey matter, white mater and "other") and the fact that the evaluation was based on only one small volume.

7.4. Qualitative evaluation of a cryo volume segmentation with multiple initialisations.

In order to investigate the robustness of ACSR segmentation to variations in template selection, an experiment was carried out, in which a group of expert observers ranked segmentations of the same cryo section volume, based on template sets selected by four different individuals. The aim was to determine whether or not participants would rank the four segmentations in a consistent and significantly different way. If they did, it would be an indication that different individuals were not capable of facilitating equally good segmentation through their template selection.

7.4.1. Methods.

A subvolume of 19 slices from the hip bone cryo section volume described in chapter 4, section 4.10.1 was selected for this experiment. The four participants chosen for template selection consisted of the author as well as two surgeons (DSOT St. Mary's)

and one radiologist (RSU Hammersmith) with no previous experience in using ACSR segmentation. As mentioned in section 7.3.1, the surgeons are used to assessing anatomy based on natural colour and texture, although typically not on cross sections. Radiologists are highly skilled in interpreting cross-sections, but are not used to basing their assessment on natural colour textures. The author (who holds an exam in macroscopic and neuro anatomy) was included because of his experience with ACSR segmentation.

A group of 11 surgeons (DSOT St. Mary's) carried out relative and absolute ranking of the four segmentations. This group was the same one used to rank the cryo section brain images described in section 7.3. In fact the two experiments were carried out in one session for each participant, but the order of the two tasks was randomly chosen in each session to even out any bias or effects of learning.

After receiving instructions about the desired segment classes, participants in the template selection task were free to select any of the slices (a minimum of two) from the hip bone volume. Although participants could select as many slices as they wanted to (in order to achieve templates representative of the whole volume) none of the four selected more than two slices. From each of their selected slices, participants selected templates for two classes: blood vessels and bone marrow. Templates were marked in a pseudo-colour using a commercial paint package (Paint Shop Pro v. 5.1 from JASC), where participants also had the option of using a zooming tool. All the slices of the volume could be viewed as thumbnails simultaneously with their slice number. At any time participants could switch to another custom made program, which displayed the volume as an image stack with the current slice number displayed. In this program it was possible to do an automatic fly-through of the volume or browse interactively forwards or backwards through the volume slice by slice.

Participants in the ranking task were presented with the source volume and its four segmentations as image stacks on a CRT display. The individual viewports of each image stack were arranged with the source in the centre and the segmentations above, below, to the right and to the left. The starting position of each segmentation was randomly chosen for each participant. The current slice number was synchronised for

all volumes, but the starting slice number was randomly selected in each session. Participants could select a fly-through or browse through the individual slices. The task was to arrange the volumes in order of perceived segmentation quality, based on an overall assessment of each image sequence. Subsequently a grade from 1 to 7 was given to each segmentation.

7.4.2. Results.

The rankings for each participant are shown in table 7.10 (relative) and table 7.11 (absolute).

Table 7.10. Relative rankings of all segmentations of the hip bone volume. Rows correspond to observers, columns to the segmentation being ranked, each based on a different template set.

Sample	Volume 1	Volume 2	Volume 3	Volume 4
1	3	1	4	2
2	3	1	4	2
3	1	4	2	3
4	3	1	4	2
5	2	1	4	3
6	1	4	2	3
7	1	4	3	2
8	4	3	1	2
9	1	4	2	3
10	4	1	3	2
11	3	1	2	4
All samples	**26**	**25**	**31**	**28**

Table 7.11. Absolute rankings of all segmentations of the hip bone volume. Rows correspond to observers, columns to the segmentation being ranked, each based on a different template set. The last row shows the mean grade for each segmentation.

Sample	Volume 1	Volume 2	Volume 3	Volume 4
1	5	3	5	4
2	4	2	6	4
3	3	6	4	4
4	4	2	5	2
5	4	2	6	5
6	2	5	3	3
7	4	5	5	4
8	6	4	1	3
9	2	5	3	4
10	4	2	3	3
11	6	5	6	6
Mean	**4**	**3.7**	**4.3**	**3.8**

The ICC(3,11) was calculated as -1.781 for the relative ranking and -1.467 for the absolute ranking.

Looking at table 7.10, it is obvious that the rankings were highly inconsistent, with volumes having a high relative ranking by one observer and a low ranking by another. The segmentations referred to as volume 1, 2 and 3 in table 7.10 all have several rankings of both 1 and 4 by different observers, while volume 4 is slightly more consistent across observers. In the absolute ranking this trend is even more exaggerated for all four segmentations. Table 7.12 shows the result of the Kruskal-Wallis H test performed on the four segmentations. The difference between segmentations was non-significant for both the relative and absolute ranking.

Table 7.12. Results of Kruskal-Wallis H tests for relative and absolute rankings of the four segmentations of the hip bone volume. Significance at p = 0.05.

Volume	Relative rankings	Absolute rankings
Cryo hip bone	p = 0.684 not significant	p = 0.776 not significant

7.4.3. Discussion.

There are two possible interpretations of the results. It is possible that observers did not a share a "goodness of segmentation". Therefore although they could distinguish between the four segmentations, they did so based on different undefined subjective criteria. This would be supported by the absolute rankings following the relative rankings in terms of spread for each observer. However it is highly unlikely that the same group who ranked the cryo section brain segmentations with an ICC above 0.9 would be unable to rank in a mutually consistent way if the four segmentations of the hip bone volume had been clearly distinguishable. The negative ICC and the highly non-significant H test can be accounted for by the four segmentations being too similar to distinguish, leading to a random relative ranking for each participant. The difficulty in distinguishing the segmentations was reported during the experiment by most of the observers. The absolute rankings are not consistent with this interpretation, since each observer would have been expected to rank all

segmentations similarly. However it is likely that observers felt that they had to demonstrate the difference between segmentations, having specified a relative order. While the absolute rankings for the cryo brain volume were consistent with the comments given by participants in terms of observed difference, several participants commented that there was only a minute difference between the best and the least good segmentation for the hip bone volume. One participant reporting this still gave a 1 to the least good segmentation and a 6 to the best one. It is therefore highly probable that the observers were not able to distinguish clearly between the four segmentations.

The conclusion is that the four template sets facilitated segmentations of virtually equal quality. The highest absolute ranking on average was given to the segmentation facilitated by one of the surgeons (volume 3), while the author with the highest level of experience in ACSR segmentation showed no advantage and came second (volume 1). The segmentation facilitated by the radiologist had the lowest absolute ranking on average (volume 2).

Given that the rankings were inconsistent between observers, the average absolute grades given in table 7.11 do not serve as a useful measure of the overall quality of the four segmentations (it is assumed here that the four segmentations were of equal quality). To achieve a more realistic measure, a weighted average of the best absolute rankings by each participant was calculated with the weights based on the best relative rankings (more than one segmentation can have the highest grade in the absolute but not in the relative ranking). Four observers favoured volume 2, another four favoured volume 3, two favoured volume 1 and one favoured volume 4. If the weight for the eight observers favouring the volumes 2 and 4 are referred to as x, the two observers favouring volume 1 as y and the one observer favouring volume 4 as z then the following equations express the weights:

$$8x + 2y + z = 1 \tag{7.2}$$

$$y = \frac{x}{4} \tag{7.3}$$

$$z = \frac{x}{8} = \frac{y}{2} \tag{7.4}$$

Solving this set of equations gives the following weights:

$$x = \frac{8}{69} \qquad y = \frac{2}{69} \qquad z = \frac{1}{69} \tag{7.5}$$

The weighted average can now be calculated as the sum of the best absolute grades for each observer (determined by the segmentation rated best in the relative ranking) multiplied by their corresponding weights:

$$\mu_{hip\_abs} = \frac{8}{69}(5+6+6+5+6+5+5+5)+\frac{2}{69}(4+6)+\frac{1}{69}(6) = 5.3623 \approx 5.4 \tag{7.6}$$

Comparing this value to the mean summed absolute rankings from the previous experiments, places the quality of the hip bone segmentation between the cryo brain volume segmentation in section 7.3 (with a value of 4.8) and the natural colour image segmentation in section 7.2 (with a value of 5.9).

7.5. Qualitative evaluation of MRI volume segmentation with multiple variables.

The final human observer experiment was based on real and simulated MRI data ranked by specialist observers. For the simulated data the aim was to investigate the observers' preference for either multispectral or single-channel segmentations based on the observed quality of representation of the segmented tissue classes. For the real data the aim was to investigate if there was any significant difference in observed segmentation quality for segmentations based on EQ and N3 inhomogeneity corrected data, and how ACSR segmentation compared to a manually created gold standard ground truth.

7.5.1. Methods.

Six radiologists participated, four from St. Mary's Hospital and two from RSU Hammersmith. The MRI experiment was divided into two types of tasks. One type concerned the evaluation of single-channel vs. multispectral segmentation of the three BrainWeb segmentations reported in chapter 6, section 6.8. There were three tasks of

this type, one for each volume. The other type concerned the evaluation of segmentation of two real MRI volumes from the IBSR. Again there was one task for each volume. The experiment was carried out in one session of three blocks. Block A contained the first IBSR task. Block B contained all three BrainWeb tasks. Block C contained the second IBSR task. Block B was always second in every session, but the order of the individual tasks in the block was randomly selected. Blocks A and C were swapped randomly between the first and the third position for each session.

7.5.2 The IBSR tasks.

In the IBSR tasks participants first carried out a relative and then an absolute ranking similar to that described in section 7.4. Segmentations were based on two different template sets, two different types of inhomogeneity correction (EQ and N3) and one manual expert ground truth.

The two IBSR volumes used in the experiment were data sets 788_6_m (55 year old normal male) and 1320_2_max (5 year old schizophrenic male). They were acquired with a 1.5 Tesla General Electric Signa. The slice thickness was 3mm. These two volumes were selected as representative MRI brain scans, because they were acquired with a popular type of scanner with typical slice thickness. Furthermore these two volumes were more complete than several of the other IBSR volumes and contained no pathological anomalies, such as tumours. Finally a manually selected expert ground truth was available for both volumes. The age difference between the two subjects meant that the proportions of the three tissue types grey matter, white matter and CSF differed, providing an extra source of variation for the experiment.

A subvolume of 19 slices was selected from each of the full volumes. It was necessary to base evaluation on a smaller set of slices in order to keep the required time per session at a reasonable level. In both volumes the 19 slices were selected from the middle of the brain, both containing the ventricles and the thalamus, which are particularly challenging structures for intensity based segmentation.

One set of templates was selected manually by the author from the first and last slice of each of the two volumes, templating grey matter, white matter and CSF. Fig. 7.1 shows the manual templates from the first slice of the adult IBSR volume. Another set of templates from the first and last slice was created for each volume from the expert ground truth, similarly to the templates of the "simulated user" in the BrainWeb study described in chapter 6. However while the entire ground truth for the three classes was used in the selected slices in the BrainWeb volumes, all boundary points were discarded for the IBSR volumes. This was due to the possibility of incorrect boundary location in the manual ground truth. The two template sets were used to facilitate the segmentation of both volumes pre-processed with EQ and N3 (and using automatic template creation), thus yielding four segmentations per volume. The manual ground truth was included unmodified as a fifth segmentation. The ground truth segmentation was generated by trained investigators at Massachusetts General Hospital, assisted by a semi-automatic segmentation from an algorithm developed by Kennedy, Filipek and Caviness [181] and intensity histograms.

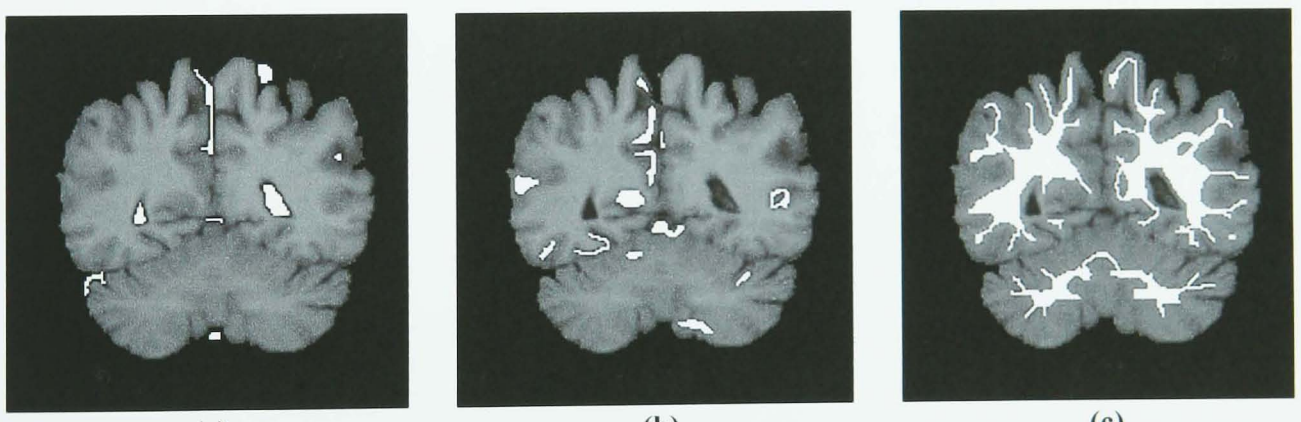


Fig. 7.1. Example of manually selected templates (shown in white) for the adult IBSR volume. (a) CSF. (b) Grey matter. (c) White matter.

The ranking task was performed in the same way as for the hip bone cryo section ranking described in section 7.4.2, the only difference being the arrangement of the image stacks on screen (see appendix D).

7.5.2.1. Results.

The summed relative rankings are shown in table 7.13 and the mean summed absolute rankings in table 7.14. The ICC for the summed rankings as well as individually for each of the two volumes are shown in table 7.15. Every ICC is high, but only one is in the very high range (>0.9) found in the previous experiments. This can be attributed to the small number of observers. In the tables "sim" refers to the simulated user (templates based on the ground truth) and "man" refers to the human user (manually selected templates).

Table 7.13. Summed relative rankings of the segmented MRI brain volumes for all subjects.

Volume	sim EQ	sim N3	ground truth	man EQ	man N3
IBSR adult	16	22	8	21	23
IBSR child	26	10	18	26	10
All	**42**	**32**	**26**	**47**	33

Table 7.14. Mean summed absolute rankings of the segmented MRI brain volumes for all subjects.

Volume	sim EQ	sim N3	ground truth	man EQ	man N3
IBSR adult	3.7	4.3	2.2	4.2	4.7
IBSR child	4.8	2.7	3.8	5.2	2.7
All	**4.3**	**3.5**	**3.0**	**4.7**	**3.7**

Table 7.15. ICC(3,11) for each MRI brain volume over all segmentations based on relative and absolute ranking.

Volume	ICC(3,6) relative ranking	ICC(3,6) absolute ranking
IBSR adult	0.733	0.823
IBSR child	0.919	0.897
All	0.775	0.835

Table 7.16 shows a number of U tests performed on EQ vs. N3 and ground truth vs. the best ACSR segmentation. All tests for EQ vs. N3 in the child volume were significant at the $p = 0.05$ level for both relative and absolute ranking. Similarly the difference between the ground truth and the best ACSR segmentation was significant in every test.

Table 7.16. Results of two-tailed Mann-Whitney U tests for relative and absolute rankings of MRI segmentations. Significance at p = 0.05.

Classes	Relative rankings	Absolute rankings
Sim EQ vs. sim N3 adult	p = 0.3209 not significant	p = 0.4082 not significant
Man EQ vs. man N3 adult	p = 0.6734 not significant	p = 0.4608 not significant
Ground truth vs. man EQ adult	p = 0.0094 **significant**	p = 0.0108 **significant**
Sim EQ vs. sim N3 child	p = 0.0039 **significant**	p = 0.0085 **significant**
Man EQ vs. man N3 child	p = 0.0103 **significant**	p = 0.0089 **significant**
Ground truth vs. man EQ child	p = 0.0228 **significant**	p = 0.0096 **significant**
Sim EQ vs. sim N3 all	p = 0.1445 not significant	p = 0.3263 not significant
Man EQ vs. man N3 all	p = 0.0571 not significant	p = 0.0099 **significant**
Ground truth vs. man EQ all	p = 0.0106 **significant**	p = 0.0046 **significant**

7.5.2.2. Discussion.

It is evident that the EQ corrected source data resulted in better segmentation than the N3 corrected data for the child volume, where large amounts of white matter in the cerebellum were misclassified as grey matter based on the N3 corrected data (see appendix D, fig. D.5). These artefacts were regarded as serious by the observers. Segmentations based on the manual template selection were higher ranked than those based on the templates derived from the ground truth. The manual ground truth was rated as the least good segmentation overall with an average absolute ranking of 3.0 compared to 4.7 for the ACSR segmentation with EQ pre-processing and initially manually selected templates. Even the segmentations where templates were based on the ground truth were rated better than the ground truth itself. This is essentially similar to partial ACSR where misrepresentations in an initial segmentation are corrected by the PGA in a subsequent segmentation, and demonstrates ACSR as an automatic optimisation tool.

7.5.3. The BrainWeb tasks.

In the BrainWeb tasks, participants carried out a categorisation, selecting one of two segmentations as their preference or stating that the two could not be distinguished. The sole purpose of these tasks was to investigate if observers would favour either the single-channel or multispectral segmentations, based on minimising the level of observed misrepresentations of the segmented tissue classes. Each task involved the categorisation of segmentations of one of the three BrainWeb volumes evaluated through ground truth comparison in chapter 6, section 6.8.

In each task the source data was displayed on screen as an image stack in the middle with the two segmentations below on either side. The position of each type of segmentation in each session and for each task was swapped randomly. Observers could select a fly-through or browse interactively as in the previously described experiments. For each volume they were asked to mark their segmentation of choice, or alternatively mark none of the two, if they felt that they could not be distinguished qualitatively.

7.5.3.1. Results.

Table 7.17 shows the frequencies of each categorisation for each volume. In 13 out of 18 cases the multispectral segmentation was preferred. All participants preferred the multispectral segmentation for the volume with 7% noise and 20% inhomogeneity. There was only one instance of an observer not being able to distinguish between the two segmentations for a single volume. Because of the small number of observers, the data was pooled for all observers and all volumes. The one observation with no difference between the two segmentations was discarded, giving 17 observations with of frequency of 13 for multispectral and 4 for single-channel. A chi-square test was performed on this data, assuming no difference between the two groups, i.e. an expected frequency of 8.5 for each group. This gave $x^2 = 4.76$. Since the critical value for the hypothesis of no difference at the 0.05 level for 1 degree of freedom is 3.84, the hypothesis could be rejected and consequently the multispectral segmentation was

significantly better rated compared to single-channel at the p = 0.05 level. However if Yates' correction for continuity [182] is applied, the value of x^2 drops to 3.76, meaning that the hypothesis of no difference could only be rejected at the p = 0.052 level.

Table 7.17. Frequencies of categories selected for each of the three BrainWeb volumes and the total for each category.

Category	3N 20RFI	3N 40RFI	7N 20RFI	All volumes
Single-channel	2	2	0	4
Multispectral	3	4	6	13
No difference	1	0	0	1

7.5.3.2. Discussion.

From the comments made by observers, it could be ascertained that their preference was mainly due to the multispectral segmentation producing connected CSF regions, while in the single-channel segmentation these regions appeared as broken islands. However because the data was simulated, it must be accepted that the better representation of CSF was in fact at the expense of losing some grey matter representation (appearing as a thinning of the boundaries), revealed by the ground truth evaluation. Are the observers then wrong? Clearly this is a matter of the application at hand. For a reconstruction of the ventricular system they would probably be right. For exact tissue quantification they would probably be wrong. The choice of the type of single-channel or multispectral ACSR segmentation discussed here might be used selectively depending on the required application.

7.6. Summary.

The results presented in this chapter have provided a qualitative evaluation of the data analysed in chapter 5 and 6 using ground truth comparison. In addition, results on new data have been presented.

Robustness to multiple initialisations was shown in the cryo hip bone experiment, where observers previously shown to rank highly consistently on a similar data set produced a negative correlation between rankings and no significant difference in observed segmentation quality over all volumes. The four segmentations were based on four different template sets and thus it could be concluded that different individuals were capable of facilitating ACSR segmentations of equal quality.

For the individual natural colour images, PBNN segmentation was preferred to LVQ and vice versa for the cryo brain sequence. This was consistent with the ground truth evaluation, showing that for this particular variation of the segmentation parameters, the observed quality of segment class representations followed the level of misrepresented points, when compared to a manual ground truth. However when varying the colour model there was no significant difference between the levels for OPC and RGB based segmentations, while human observers significantly preferred the OPC based segmentation for the natural colour images. It was also shown that multispectral MRI segmentations were preferred over single-channel segmentations because of the overall representation of the segmented tissue classes and their visual definition. Ground truth evaluation showed that single-channel segmentations were more accurate on a point by point basis.

A ground truth evaluation of the IBSR segmentations would have been highly ambiguous, given that the expert ground truth is possibly not as detailed as the semi-automatic ACSR segmentations. This was indicated by observers ranking the manual ground truth significantly lower than the best ACSR segmentation. The mean absolute rankings for the natural colour images and brain cryo series, as well as the weighted average for the cryo hip bone volume segmentation (where in all cases no qualitative comparison with a ground truth was performed) were 5.9, 4.8 and 5.4 respectively on a scale from 1 to 7. This must be considered an encouraging result, given the complexity of the images involved.

These findings indicate that human observers share a holistic view of image quality grounded in knowledge about expected image compositions (specialist knowledge about the expected anatomy in the case of the cryo section and MRI images). An

automatic evaluation achieving this would essentially require a system capable of passing an image interpretation equivalent of the Turing test. This does not however contradict the necessity for ground truth evaluation, but rather the two types of evaluation complement each other for different applications.

In conclusion, chapter 6 and 7 have provided both a quantitative and a qualitative evaluation of ACSR segmentation, showing promising results for potential applications in both tissue quantification (supported particularly by ground truth evaluation) and visualisation, including 3D reconstruction (supported particularly by visual ranking).

Chapter 8

Conclusions and future work

8.1. Conclusions from the presented work.

The goal of the work described in this thesis was to develop a new framework for robust semi-automatic segmentation of medical imaging data of multiple modalities. The aims were to devise a system in which the user would be able to specify the desired segment classes through an intuitive visual initialisation, followed by a fully automatic segmentation process, requiring no further interaction. For the purpose of initialisation, the imaging modality should be transparent to the user, requiring the same pattern of interaction regardless of the data to segment.

8.1.1. Background.

Chapter 1 gave an historical introduction to the areas of medical imaging, computer graphics visualisation and medical image processing. Some of the drawbacks of currently available methods, both in terms of performance and ease of use, were briefly discussed. This was followed by an outline of the research objectives.

The literature review in chapter 2 gave a technical introduction and background to image segmentation. The chapter concluded with a summary in the context of medical image analysis.

8.1.2. Contribution to knowledge.

The novel part of the work described in this thesis, constituting the contribution to knowledge, was described in chapters 3 to 7.

Specific problems related to medical image segmentation were identified in chapter 3 and a set of conceptual and technical requirements for a semi-automatic segmentation framework were specified. The work of Gerritsen [133] and Olabarriaga and Smeulders [134] on automatic and interactive segmentation systems was discussed. The seven conceptual requirements for the proposed framework addressed the problems of:

- Establishing a balance between automation and specification of the goal of a segmentation from the point of the user
- Enabling a focussing of proposed segmentation algorithms for the sake of efficiency
- Achieving accuracy and robustness by utilising information in all dimensions of a data set, while not imposing unwanted constraints and avoiding distortion of detail

The four technical requirements were specified as guidelines for implementing the conceptual requirements. Based on the literature review in chapter 2, it was concluded that the preliminary work should be based on a vector quantization neural network approach, and that the specific problem of segmentation near edges and correct boundary location was to be addressed in depth.

In chapter 4 a feature vector encoding for SOM and LVQ classifiers was developed. This encoding using the OPC colour model produced a 54-dimensional feature vector referred to as the PixelDefine encoding. It was demonstrated that this encoding was capable of producing segmentations of colour cryo sections based on both a supervised and an unsupervised approach. However the problems of the fixed size and shape of the sampling window prevented the method from producing good results on high and low frequency information simultaneously. It was shown with examples from the literature that this is a common problem, producing segmentation artefacts particularly at segment boundaries.

An algorithm for the encoding of plastic sampling windows for LVQ classification was developed. The Path Growing Algorithm (referred to as the PGA) is a graph based method, which considers all possible paths up to a specified length, originating

from the seed point (the equivalent of the centre pixel in the rigid window representation), as components in topologically different sampling windows for individual segment classes at every point in the source data. These paths exist in a hierarchy based on their match with a user specified class template and the total spread of point descriptors within the path. The sampling window for each class is built from these paths and may be used to find the classification of the seed point directly or through the use of a higher level classification system. Taking the latter approach with LVQ, it was demonstrated that the technique compared favourably to three other segmentation algorithms on previously published 2D natural colour images and cryo sections from the Visible Human Project. The initialisation by the user in the form of template selection and the combination of addressing the problem of segmentation near edges with a novel feature encoding and classification scheme was collectively named Adaptable Class-Specific Representation (referred to as ACSR). This was the first instantiation of the segmentation framework, which was the goal of the project. The introduction of the ACSR framework and the PGA with the results mentioned above, were published in the proceedings of the 15th IEEE International Conference on Pattern Recognition (ICPR 2000) [157].

ACSR was extended to 3D volume segmentation and used to segment two cryo section volumes from the Visible Human Project. An automatic focussing of the PGA was introduced to reduce the processing overhead. It was shown that the segmentation pipeline proposed in [157] could favourably be reversed by creating a fast initial LVQ segmentation using a standard fixed size sampling window, dilating the boundaries and finally applying the PGA only at boundary points, basing the final classification of these points directly on the winning class representation produced by the PGA. The final segmentation would consist of the masked initial LVQ segmentation with the added PGA segmentation. The first results on volume segmentation and the automatic focussing referred to as partial ACSR, were published in the proceedings of the Fifth IEEE Workshop on Applications of Computer Vision (WACV 2000) [161]. By the end of chapter 4 the ACSR framework had been defined for 2D and 3D colour image segmentation with preliminary evaluation.

More preliminary empirical evaluation was given in chapter 5. Issues relating to quantitative ground truth evaluation and qualitative visual ranking by human observers were discussed. It was concluded that the former is mostly applicable to comparative studies based on simulated or artificially composed data, while the latter could be a more realistic option for the evaluation of segmentation quality on real data. A pilot study, in which five individuals composed images and selected templates from photographs of real textures, showed that multiple template selections facilitated equally good segmentations. This study was published as part of a technical report [166]. Another study looked at the segmentation of six natural colour images. Segmentation accuracy was compared between LVQ, a Point Based Nearest Neighbour (PBNN) classifier, full and partial ACSR using the PGA with the RGB and OPC colour models. A similar comparison was performed for the segmentation of a series of five brain cryo sections. For the natural colour images, the PGA with the OPC colour model showed an accuracy between 94.83% and 99.58% based on comparison with a manually generated ground truth. In the brain slices the accuracy ranged between 91.72% and 93.75%. No significant difference between the results based on OPC and those based on RGB could be found, although results for OPC were slightly better overall. The fast PBNN classifier performed better than LVQ for the discrete 2D images while LVQ performed better for the image sequence, as expected. Partial ACSR achieving the same level of accuracy as full ACSR performed up to 20 times faster.

In chapter 6 the ACSR framework was extended to greyscale MRI segmentation. Three different path growing algorithms using the path median and average intensity difference as additional descriptors were tested in a comparative study based on segmentations of simulated BrainWeb volumes (with varying levels of noise and inhomogeneity) into the three classes CSF, grey matter and white matter. The algorithm using a single path as the class representation per point, and employing a seed point shifting in the direction of growth for the calculation of average intensity difference, generated the lowest error rates. This study was published as a short paper in Medical Image Computing and Computer-Assisted Intervention 2001, Lecture Notes in Computer Science, vol. 2208 [174], and in more detail in a technical report [175].

The problems of noise and inhomogeneity artefacts were targeted. An automatic template creation based on the initial LVQ segmentation was introduced. It was shown that this way of including more local information in templates increases the robustness to noise and ensures a consistent error rate throughout the volume. Two different algorithms were tested for inhomogeneity correction of the source data as a pre-processing step. The well established model based N3 algorithm [44,45] and the more recent volume intensity equalisation algorithm EQ [43] were employed. It was shown on the simulated BrainWeb volumes that a combination of the two optimisations always performed better then either one or none. The results of ACSR segmentation with inhomogeneity correction and automatic template creation, compared favourably to previously published results on the same volumes by Pham and Prince [120,130] using MRF segmentation with a standard EM algorithm and the AGEM algorithm [130] modelling inhomogeneity. N3 consistently facilitated better segmentation than EQ.

A multispectral segmentation using T2 weighed images for the CSF class and T1 images for the grey matter and white matter classes was performed and showed slightly higher error rates than for the single-channel segmentations. This was in spite of the CSF class appearing to be better visually represented in the multispectral segmentation.

The results using automatic template creation and inhomogeneity correction were published in the proceedings of SPIE Medical Imaging 2002: Image Processing [177]. The quantitative results in chapter 6 on simulated data suggested a high level of accuracy and robustness to noise and inhomogeneity artefacts. However no testing was performed on real MRI data due to the problems associated with ground truth evaluation and the subjective nature of a "gold standard" ground truth for real data.

Chapter 7 presented the final empirical evaluation of the ACSR framework for colour cryo section and MRI data. The evaluation was based on human observer experiments, using both naïve and expert observers to perform visual ranking.

The natural colour images and the brain cryo section sequence from chapter 5 were rated by computing science students and surgeons respectively. Results on the natural colour images showed a statistically significant difference between the OPC and RGB based results, favouring OPC, with an average rating of 5.9 on a scale from 1 to 7 for the PGA-OPC segmentation. The average rating for the PGA-OPC segmentation of the cryo section sequence was 4.8. ICC for both studies was above 0.9.

Four observers (two surgeons, one radiologist and the author) selected templates for the hip bone cryo section volume first presented in chapter 4. No significant difference between ratings of the four segmentations was found. This demonstrated robustness to multiple initialisations, by showing that the four template sets had facilitated segmentations of close to equal quality. The author was the only observer with previous experience in using ACSR segmentation, but his templates did not facilitate a significantly better segmentation. The average absolute rating based on a weighted average of the best grade by each observer was 5.4.

Finally six radiologists evaluated segmentations of simulated and real MRI data. In one part of the experiment, observers were presented with the single-channel and multispectral BrainWeb segmentations described in chapter 6. The results showed that the observed quality of the multispectral segmentations was significantly better than single-channel segmentations, although due to the small number of observers this could only be conclusively stated for the volume with the highest level of noise. This was contrary to the results on ground truth evaluation. It was argued that the multispectral segmentation might be an advantage for applications in visualisation, while the single-channel segmentation would be an advantage in applications for tissue quantification.

In the second part of the experiment the observers ranked segmentations of two real MRI volumes (one child and one adult) from the Internet Brain Segmentation Repository [42]. One set of templates had been manually selected by the author and another had been automatically generated from the manual ground truth in the two end slices. Each of these template sets were used to segment source data pre-processed with EQ and N3 and using automatic template creation, resulting in four

segmentations. A fifth segmentation was the unmodified expert ground truth. EQ based segmentations were significantly better rated than N3 for the child volume, where the N3 pre-processed data caused substantial amounts of white matter to be misclassified as grey matter. Segmentations from the manually selected templates were higher rated than those based on templates generated from the ground truth, and significantly better rated than the unmodified manual ground truth in all tests. This was an extremely positive result. The average absolute ranking for the segmentations using the manually selected templates with EQ inhomogeneity correction was 4.7.

It was concluded at the end of chapter 7 that while ground truth evaluation was an invaluable tool for comparative studies based on simulated data, the human observer experiments had revealed subtle observed differences in the overall visual representation (OPC vs. RGB and single-channel vs. multispectral), which had been missed by the ground truth evaluation, and enabled an objective testing on real data.

The results of the human observer experiments and how they related to ground truth evaluation were published in the proceedings of the International Conference on Diagnostic Imaging and Analysis 2002 [183].

8.2. Future work.

The automatic template creation, which has clearly improved segmentation accuracy for MRI data (compared to using manual templates directly), is likely to improve results for colour cryo section segmentation too. However to test this would require the availability of standard cryo section data sets with expert ground truth. Because there are no such data sets available currently, the second option would be to have a group of experienced observers create such a ground truth. This is a highly time consuming task, but would certainly be possible as part of future work.

It was mentioned in the introduction that the data from the Korean Visible Human [34] is to become available to researchers shortly. This data will require detailed and highly accurate segmentation similar to what the Visible Human Project data has been

subjected to, in order to create applications such as the VoxelMan atlases [19]. It is likely that the creation of such data sets will become more widespread for patient specific data in the future. This opens up new challenges for colour cryo section segmentation in the area of virtual pathology.

ACSR segmentation has been tested on simulated and real MRI data for the quantification of grey matter, white matter and CSF, showing very promising results. However the framework has yet to be tested on any specific clinical applications. An immediate possibility requiring little further development would be an application as a diagnostic tool for Multiple Sclerosis, Alzheimer's disease or Schizophrenia. They are all conditions, in which the proportions of the grey matter, white matter and CSF change in an abnormal way [184-188]. To establish how well ACSR would cope with this type of application, would require testing on a large number of data sets from both healthy and diseased subjects over a range of different ages. Preliminary testing on simulated BrainWeb data would be possible using the simulated Multiple Sclerosis volumes.

Detection of lesions in MRI volumes, such as tumours, are a possibility. However if the tumour tissue would need to be explicitly templated, then it would to some extent defy the purpose of an automatic detection process. It would be possible to automate the template selection for such lesions, based on experiments with specific contrast agents over a large number of data sets, and/or possibly combining the intensity based PGA segmentation with a simple form of shape matching.

With regards to the application of ACSR segmentation to other modalities than those described in this thesis, any modality, in which a mapping can be found between combinations of spatially connected grey levels and segment classes, is a possibility. This means, for example, that ACSR should be immediately applicable to CT segmentation (in fact some preliminary results suggest that it works well), but would require the development of new algorithms for ultrasound segmentation.

The results on natural colour image segmentation, which were presented due to the lack of standard cryo section test images, suggest that ACSR has applications outside

the medical area. ACSR might be integrated with paint packages, in order to extract objects from scenes, and assist artists in manipulating and compositing photo realistic images. Using a reduced set of paths for low processing overhead, ACSR might also be used in real time applications, such as tracking of objects or background extraction for augmented reality systems.

The current software implementation of ACSR segmentation is split over a number of different C programs for different tasks and different types of data. In order to make ACSR segmentation available for other people to use or do further research on, it would be beneficial to integrate the current implementation of ACSR in a set of C/C++ libraries, and possibly add a windows based front end for ease of use. The modular nature of ACSR, in which it is possible to "plug in" different colour models and different types of pre and post-processing, makes it suitable for an object oriented implementation.

Medical image segmentation will continue to face new challenges over the next decades. Successful solutions will depend on the tailoring of available techniques to specific applications. The development and testing of the ACSR segmentation framework has only just reached the level, where one might begin to use it on clinical problems. I encourage readers of this thesis to pursue the ideas presented herein for their own applications.

References

[1] R. Satava, "Medical Virtual Reality : The Current Status of the Future", *Medicine Meets Virtual Reality 1996*, IOS Press, Amsterdam, pp. 100-106 (quote from p. 100), 1996

[2] E. H. Shortliffe, L. E. Perreault, G. Wiederhold, L. M. Fagan (editors), *Medical Informatics*, Springer Verlag, New York, 2000

[3] W. C. Röntgen, *Über eine neue Art von Strahlung*, Physikalisch-Medizinische Gesellschaft, Universität Würzburg (The Physical-Medical Society, University of Würzburg), December 28, 1895. English translation appears in [4]

[4] W. C. Röntgen, "On a New Kind of Ray" (translation by Arthur Stanton), *Nature*, vol. 53, pp. 274-276, 1896

[5] G. M. Baxter, P. L. P. Allan, P. Morley (editors), *Clinical Diagnostic Ultrasound*, Blackwell Science, Oxford, 1999

[6] S. C. Bushong, *Magnetic Resonance Imaging Physical and Biological Principles*, CV Mosby Company, St. Louis, 1988

[7] A. C. Kak, M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, 1988

[8] G. T. Herman, H. K. Liu, "Three-Dimensional Display of Human Organs from Computed Tomograms", *Computer Graphics and Image Processing*, vol. 9, no. 1, pp. 1-21, 1979

[9] R. W. Prager, A. H. Gee, L. Berman, "STRADX: Real-Time Acquisition and Visualisation of Freehand 3D Ultrasound", technical report, no. CUED/F-INFENG/TR 319, Department of Engineering, University of Cambridge, 1998

[10] A. D. Linney, J. Deng, "Three-Dimensional Morphometry in Ultrasound", *Proceedings of the Institution of Mechanical Engineers*, vol. 213, part H, pp. 235-245, 1999

[11] W. Schroeder, K. Martien, B. Lorensen, *The Visualization Toolkit - An Object-Oriented Approach to 3D Graphics*, Prentice Hall, New York, 1998

[12] W. E. Lorensen, H. E. Cline, "Marching Cubes: A High Resolution 3D Surface Reconstruction Algorithm", *ACM Computer Graphics*, vol. 21, no. 4, pp. 163-169, 1987

[13] G. Farin, *Curves and Surfaces for Computer-Aided Geometric Design - A Practical Guide*, Academic Press, San Diego, 1997

[14] E. Catmull, J. Clark, "Recursively Generated B-Spline Surfaces on Arbitrary Topological Surfaces", *Computer-Aided Design*, vol. 10, no. 6, pp. 350-355, 1978

[15] A. Kalvin, C. B. Cutting, B. Haddad, M. E. Noz, "Constructing Topologically Connected Surfaces for the Comprehensive Analysis of 3D Medical Structures", *SPIE Proceedings on Image Processing*, vol. 1445, pp. 247-258, 1991

[16] W. J. Schroeder, J. A. Zarge, W. E. Lorensen, "Decimation of Triangle Meshes", *ACM Computer Graphics*, vol. 26, no. 2, pp. 65-70, 1992

[17] U. Tiede, K. H. Höhne, M. Bomans, A. Pommert, M. Riemer, G. Wiebecke, "Investigation of Medical 3D-Rendering Algorithms", *IEEE Computer Graphics Applications*, vol. 10, no. 2, pp. 41-53, 1990

[18] K. H. Höhne, U. Tiede, M. Riemer, M. Bomans, M. Heller, G. Witte, "Static and Dynamic Three-Dimensional Display of Tissue Structures from Volume Scans", *Radiology*, vol. 165, no. 2, p. 420, 1987

[19] U. Tiede, T. Schiemann, K. H. Höhne, "Visualizing the Visible Human", *IEEE Computer Graphics Applications*, vol. 16, no. 1, pp. 7-9, 1996

[20] A. Vesalius, *De Humani Corporis Fabrica*, Johann Oporin, Basel, 1543

[21] TeraRecon Incorporated, 2955 Campus Drive, Suite 325, San Mateo, CA 94403, USA, *URL*, http://www.terarecon.com/3d_products.html

[22] J. Waldsmith, *Stereo Views: An Illustrated History and Price Guide*, Wallace-Homestead Book Company, Radnor, PA, 1991

[23] Berezin Stereo Photography Products, *URL*, http://www.berezin.com/3d/viewmast.htm

[24] M. Billinghurst, "Do You See What I See?", *Virtual Reality Special Report*, vol. 2, no.1, pp. 21-26, 1995

[25] C. Cruz-Neira, D. Sandin, T. DeFanti, R. Kenyon, J. Hart, "The CAVE: Audio Visual Experience Automatic Virtual Environment", *Communications of the ACM*, vol. 35, no. 6, pp. 64-72, 1992

[26] V. Spitzer, M. J. Ackerman, A. L. Scherzinger, D. Whitlock, "The Visible Human Male: A Technical Report", *Journal of the American Medical Informatics Association*, vol. 3, no. 2, pp. 118-130, 1996

[27] M. J. Ackerman, "Accessing the Visible Human Project", W. Y. Arms, P. B. Hirtle, B. Wilson (editors), Cornell University, *D-Lib Magazine* (on-line digital library), October 1995, *URL*, http://www.dlib.org/dlib/october95/10ackerman.html

[28] W. L. Heinrichs, A. Pothen, R. Mather, P. Constantinou, M. Lewis, R. A. Chase, P. Dev, "3D Female Pelvic Organ Models: Comparison of the Human Visible Female with a Reproductive Age Pelvis", *Proceedings of the Visible Human Conference*, CD-ROM, 1996

[29] W. Nip, C. Logan, *Whole Frog*, technical report, no. LBL-35331, University of California, Lawrence Berkeley Laboratory, Berkely, CA, 1991

[30] D. Robertson, W. Johnston, W. Nip, "Virtual Frog Dissection: Interactive 3D Graphics via the Web", *Proceedings of the Second World Wide Web Conference*, (electronic proceedings) *URL*, http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/BioChem/robertson/robertson.html, 1994

[31] Middlesex University Vision and Image Processing Group – Volume rendering, *URL*, http://www.cs.mdx.ac.uk/staffpages/casper1/volren/

[32] K. Kayser, J. Szymas, R. S. Weinstein, *Telepathology*, Springer Verlag, New York, 1999

[33] B. D. Athey, *Development and Demonstration of a Networked Telepathology 3-D Imaging, Databasing, and Communication System*, report no. DARPA/DSO BAA 94-14, U. S. Army Medical Research and Materiel Command, Fort Detrick, Maryland, 1995

[34] M. S. Chung, J. Y. Kim, W. S. Hwang, J. S. Park, "Visible Korean Human: Another Trial for Making Serially Sectioned Images", *SPIE Medical Imaging 2002: Visualization, Image-Guided Procedures, and Display*, SPIE Proceedings, vol. 4681, pp. 171-183, 2002

[35] D. Crevier, *AI - The Tumultuous History of the Search for Artificial Intelligence*, BasicBooks, New York, p. 88, 1993

[36] T. F. Meaney, M. A. Weinstein, E. Buonocore, W. Pavlicek, G. P. Borkowski, J. H. Gallagher, B. Sufka, W. J. MacIntyre, "Digital Subtraction Angiography of the Human Cardiovascular System", *American Journal of Roentgenology,* vol. 135, no. 6, pp. 1153-1160, 1980

[37] Mayo Clinic, *URL*, http://www.mayo.edu/bir/

[38] MatLab Image Processing Toolbox from MathWorks, *URL*, http://www.mathworks.com/products/tech_computing/visimage.shtml

[39] VXL homepage at University of Oxford, *URL*, http://www.robots.ox.ac.uk/~vxl/

[40] OpenCV homepage at Intel Corporation, *URL*, http://www.intel.com/research/mrl/research/opencv/

[41] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, A. C. Evans, "Design and Construction of a Realistic Digital Brain Phantom", *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463-468, 1998

[42] D. N. Kennedy, A. J. Worth, V. S. Caviness Jr., "MRI-Based Internet Brain Segmentation Repository", *Proceedings of the International Society for Magnetic Resonance in Medicine*, vol. 3, p. 1657, 1996

[43] M. Cohen, R. D. DuBois, M. M. Zeineh, "Rapid and Effective Correction of RF Inhomogeneity for High Field Magnetic Resonance Imaging", *Human Brain Mapping*, vol. 10, no. 4, pp. 204-211, 2000

[44] J. G. Sled, A. P. Zijdenbos, A. C. Evans, "A Non-Parametric Method for Automatic Correction of Intensity Non-Uniformity in MRI Data", *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87-97, 1998

[45] J. G. Sled, A. P. Zijdenbos, A. C. Evans, "A Comparison of Retrospective Intensity Non-Uniformity Correction Methods for MRI", *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, vol. 1230, pp. 459-464, 1997

[46] Z. Li, *Pre-Attentive Segmentation in the Primary Visual Cortex*, technical report, MIT AI Memo no. 1640, MIT Artificial Intelligence Laboratory, 1998

[47] J. Seymour, T. L. Kriebel, "Virtual Human: Live Volume Rendering of the Segmented and Classified Visible Human Male in a CD-ROM Product for PCs", *Proceedings of the 1998 Visible Human Project Conference*, CD-ROM, 1998

[48] T. M. Strat, M. A. Fischler, "Context-Based Vision: Recognizing Objects Using Both 2D and 3D Imagery", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 1050-1065, 1991

[49] B. Julesz, "Method of Coding Television Signals Based on Edge Detection", *Bell System Technical Journal*, no. 38, pp. 1001-1020, 1959

[50] D. H. Hubel, T. N. Wiesel, "Receptive Fields and Functional Architecture of Monkey Striate Cortex", *Journal of Physiology*, vol. 195, no. 1, pp. 215-243, 1968

[51] D. Marr, *Vision - A Computational Investigation Into the Human Representation and Processing of Visual Information*, W. H. Freeman and Company, San Francisco, 1982

[52] L. G. Roberts, "Machine Perception of Three-Dimensional Solids", *Optical and Electro-Optical Information Processing*, MIT Press, Cambridge, MA, pp. 159-197, 1965

[53] I. E. Sobel, *Camera Models and Machine Perception*, PhD thesis, Electrical Engineering Department, Stanford University, 1970

[54] J. Prewitt, "Object Enhancement and Extraction", *Picture Processing and Psychopictorics*, Academic Press, New York, pp. 75-149, 1970

[55] J. F. Canny, *Finding Edges and Lines in Images*, MSc thesis, no. TR-720, MIT AI Lab, 1983

[56] J. F. Canny, "A Computational Approach to Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, 1986

[57] D. Ballard, C. Brown, *Computer Vision*, Prentice Hall, London 1982

[58] D. Marr, E. Hildreth, "The Theory of Edge Detection", *Proceedings of the Royal Society of London*, series B, vol. 207, pp. 187-217, 1980

[59] R. Fisher, S. Perkins, *The Hypermedia Image Processing Reference*, John Wiley and Sons, Chichester, West Sussex, 1996

[60] M. Petrou, P. Bosdogianni, *Image Processing - The Fundamentals*, John Wiley and Sons, Chichester, West Sussex, 1999

[61] K. Castleman, *Digital Image Processing*, Prentice Hall, London, 1996

[62] J. H. Elder, S. W. Zucker, "Local Scale Control for Edge Detection and Blur Estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 7, pp. 699-716, 1998

[63] R. M. Haralick, "Textural Features for Image Classification", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610-621, 1973

[64] M. M. Galloway, "Texture Analysis using Gray Level Run Lengths", *Computer Graphics and Image Processing*, vol. 4, pp. 172-179, 1975

[65] M. Levine, S. Shaheen, "A Modular Computer Vision System for Image Segmentations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, no. 5, pp. 540-557, 1981

[66] R. P. Schoenmakers, G. G. Wilkinson, T. E. Schouten, "Results of a Hybrid Segmentation Method", *Image and Signal Processing for Remote Sensing,* SPIE Proceedings, vol. 2315, pp. 90-101, 1994

[67] T. Pavlidis, Y. T. Liow, "Integrating Region Growing and Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 3, pp. 225-233, 1990

[68] K. Haris, S. N. Efstratiadis, N. Maglaveras, A. K. Katsaggelos, "Hybrid Image Segmentation Using Watersheds and Fast Region Merging", *IEEE Transactions on Image Processing*, vol. 7, no. 12, pp. 1684-1699, 1998

[69] S. L. Horowitz, T. Pavlidis, "Picture Segmentation by a Tree Traversal Algorithm", *Journal of the ACM*, vol. 23, no. 2, pp. 368-388, 1976

[70] Watershed Segmentation using the Insight Segmentation and Registration Toolkit, *URL*, http://www.cs.utah.edu/~cates/watershed/

[71] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform", *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 25, no. 3, pp. 235-238, 1977

[72] A. Grossmann, J. Morlet, "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape", *SIAM Journal of Mathematical Analysis*, vol.15, no. 4, pp. 723-735, 1984

[73] Impuls GmbH, Carl-Benz-Strasse 13, 82205 Gilching, Germany, *URL*, http://www.impuls-imaging.com/

[74] Robi Polikar's on-line wavelet tutorial, *URL*, http://engineering.rowan.edu/~polikar/WAVELETS/WTtutorial.html

[75] D. Gabor, "Theory of Communication", *Journal of the Institution of Electrical Engineers*, vol. 93, no. 3, pp. 429-459, 1946

[76] A. C. Bovik, M. Clark, W. S. Geisler, "Multichannel Texture Analysis using Localized Spatial Filters", *IEEE Transactions on Pattern Analysis and Machine Intelligence*", vol. 12, no. 1, pp. 55-73, 1990

[77] N. W. Campbell, B. T. Thomas, "Automatic Selection of Gabor Filters for Pixel Classification", *Proceedings of the Sixth International Conference on Image Processing and its Applications*, pp. 761-765, 1997

[78] J. Radon, "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Manningfaltigkeiten", *Berichte der königlich Sächsischen Akademie der Wissenschaften* (Proceedings of the Royal Saxon Academy of Sciences), vol. 69, pp. 262-267, 1917. English translation appears in [80]

[79] P. V. C. Hough, "Method and Means for Recognizing Complex Patterns", US Patent, no. 3,069,654, December 18, 1962

[80] S. Deans, *The Radon Transform and Some of Its Applications*, John Wiley and Sons, New York, 1983

[81] A. C. Copeland, G. Ravichandran, M. M. Trivedi, "Localized Radon Transform-Based Detection of Linear Features in Noisy Images", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 664-667, 1994

[82] R. O. Duda, P. E. Hart, "Use of the Hough Transform to Detect Lines and Curves in Pictures", *Communications of the ACM*, vol. 18, no. 1, pp. 11-15, 1972

[83] P. Toft, *The Radon Transform - Theory and Implementation*, PhD thesis, Department of Mathematical Modelling, Section for Digital Signal Processing, Technical University of Denmark, 1996

[84] B. B. Mandelbrot, *Fractal Geometry of Nature*, W. H. Freeman and Company, Oxford, 1982

[85] R. F. Voss, "Local Connected Fractal Dimension Analysis of Early Chinese Landscape Paintings and X-Ray Mammograms", *Fractal Image Encoding and Analysis* (Y. Fisher, editor), Springer Verlag in coorporation with NATO, pp. 279-297, 1995

[86] C. J. G. Evertsz, C. Zahlten, H. Peitgen, "Distribution of Local-connected Fractal Dimension and the Degree of Liver Fattiness from Ultrasound", *Fractals in Biology and Medicine* (T. F. Nonnenmacher, G. A. Losa, E. R. Weisel, editors), Birkhauser Verlag, Basel, 1994

[87] B. Kass, A. Witkin, D. Terzopoulos, "Snakes: Active Contour Models", *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321-331, 1987

[88] L. D. Cohen, "On Active Contour Models and Balloons", *Computer Vision Graphics and Image Processing: Image Understanding*, vol. 53, no. 2, pp. 211-218, 1991

[89] T. McInerney, D. Terzopolous, "Deformable Models in Medical Image Analysis: A Survey", *Medical Image Analysis*, vol. 1, no. 2, pp. 91-108, 1996

[90] A. Yezzi, S. Kichenassamy, A. Kumar, P. J. Olver, A. Tannenbaum, "A Geometric Snake Model for Segmentation of Medical Imagery", *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 199-209, 1997

[91] W. A. Barrett, E. N. Mortensen, "Interactive Live-Wire Boundary Extraction", *Medical Image Analysis*, vol. 1, no. 4, pp. 331-341, 1997

[92] J. Montagnat, H. Delingette, N. Scapel, N. Ayache, "Representation, Shape, Topology and Evolution of Deformable Surfaces. Applications to 3D Medical Image Segmentation", technical report, no. RR-3954, INRIA, 2000

[93] T. F. Cootes, A. Hill, C. Taylor, J. Haslam, "The Use of Active Shape Models For Locating Structures in Medical Images", *Image and Vision Computing*, vol. 12, no. 6, pp. 355-366, 1994

[94] T. F. Cootes, G. J. Edwards, C. J. Taylor, "Active Appearance Models", *Proceedings of the 5th European conference on Computer Vision*, vol. 2, pp. 484-498, 1998

[95] M. D. Fairchild, *Color Appearance Models*, Addison Wesley Longman, Reading, MA, 1998

[96] A. R. Smith, "Color Gamut Transform Pairs", *ACM Computer Graphics*, vol. 12, no. 3, pp. 12-19, 1978

[97] The Graphics Standards Planning Committee, "Status Report of the Graphics Standards Planning Committee", *ACM Computer Graphics*, vol. 13, no. 3, 1979

[98] E. Hering, *Zur Lehre vom Lichtsinn*, Gerold und Söhne, Vienna, 1878. English translation appears in [99]

[99] L. M. Hurvich, D. Jameson, *Outlines of a Theory of the Light Sense*, Harvard University Press, Cambridge, MA, 1964

[100] N.W. Campbell, B.T. Thomas, T. Troscianko, "Segmentation of Natural Images using Self-Organising Feature Maps", *Proceedings of the British Machine Vision Conference*, pp. 223-232, 1996

[101] K. Yamaba, Y. Miyake, "Colour Character Recognition Method Based on Human Perception", *Optical Engineering*, vol. 32, no. 1, pp. 33-40, 1993

[102] J. E. Dowling, *The Retina - An Approachable Part of the Brain*, Harvard University Press, USA, colour plate 8, 1987

[103] T. Schiemann, U. Tiede, K. H. Höhne, "Segmentation of the Visible Human for High Quality Volume Based Visualization", *Medical Image Analysis*, vol. 1, no. 4, pp. 263-271, 1997

[104] J. E. Stewart, J. H. Johnson, W. C. Broaddus, "Segmentation and Reconstruction Strategies for the Visible Man", *Proceedings of the 1996 Visible Human Project Conference*, CD-ROM, 1996

[105] P. Beylot, P. Gingins, P. Kalra, N. M. Thalmann, W. Maurel, D. Thalmann, J. Fasel, "3D Interactive Topological Modeling using Visible Human Dataset", *Computer Graphics Forum*, vol. 15, no. 3, pp. 33-34, 1996

[106] S. O. Senger, "User-Directed Segmentation of the Visible Human Data Sets in an Immersive Environment", *Proceedings of the 1998 Visible Human Project Conference*, CD-ROM, 1998

[107] T. O. Binford, T. S. Levitt, W. B. Mann, "Bayesian Inference in Model-Based Machine Vision", *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pp. 73-95, 1989

[108] A. Kehagias, "Bayesian Classification of Hidden Markov Models", *Mathematical and Computer Modelling* , vol. 23, no. 5, pp. 25-43, 1996

[109] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer Verlag, Tokyo, 2001

[110] N. Metropolis, S. Ulam, "The Monte Carlo Method", Journal of the American Statistical Association, vol. 44, no. 247, pp. 335-341, 1949

[111] N. M. Radford, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, technical report, no. CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993

[112] P. J. Green, "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination", *Biometrica*, vol. 82, no. 4, pp. 711-732, 1995

[113] A. P. Dempster, M. Laird, D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Proceedings of the Royal Statistical Society*, Series B, Vol. 39, pp. 1-38, 1977

[114] J. L. McClelland, D. E. Rumelhart, *Explorations in Parallel Distributed Processing - A Handbook of Models, Programs, and Exercises*, MIT Press, Massachusetts, 1988

[115] P. Picton, *Neural Networks*, Palgrave, Hamphshire, UK, p. 37, 2000

[116] S. Kirkpatrick, C. D. Gellat, M. D. Vecchi, "Optimisation by Simulated Annealing", *Science*, vol. 220, pp. 671-680, 1983

[117] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Berlin, 1997

[118] S. Dickson, *Investigation of the use of Neural Networks for Computerised Medical Image Analysis*, PhD thesis, Department of Computer Science, University of Bristol, 1998

[119] E. Kerre, M. Nachtegael (editors), *Fuzzy Techniques in Image Processing - Techniques and Applications*, Physica Verlag, Heidelberg, 2000

[120] D. L. Pham, J. L. Prince, "An Adaptive Fuzzy Segmentation Algorithm for Three-Dimensional Magnetic Resonance Images", *Information Processing in Medical Imaging 1999*, Lecture Notes in Computer Science, vol. 1613, pp. 140-153, 1999

[121] Peter Maybeck, "Stochastic Models, Estimation, and Control", vol. 1, pp. 1-16, Academic Press, New York, 1979

[122] A. Ortiz, M. Simo, G. Oliver, "Image Sequence Analysis for Real-Time Underwater Cable Tracking", *Proceedings of the Fifth IEEE Workshop on Applications of Computer Vision*, pp. 230-236, 2000

[123] I. T. Jolliffe, *Principal Component Analysis*, Springer Verlag, New York, 1986

[124] B. Moghaddam, A. Pentland, "Probabilistic Visual Learning for Object Detection", *Proceedings of the Fifth International Conference on Computer Vision*, pp. 786-793, 1995

[125] Trevor Darrell, *Active Face Tracking and Pose Estimation in an Interactive Room*, technical report, no. 356, MIT Media Laboratory Perceptual Computing Group, 1996

[126] M. Phelps, J. Mazziotta, H. Schelbert (editors), *Positron Emission Tomography of the Brain*, Springer Verlag, Berlin, 1983

[127] R. J. Jaszczak, "Tomographic Radiopharmaceutical Imaging", *Proceedings of the IEEE*, vol. 76, no. 9, pp. 1079-1094, 1988

[128] R. A. Brooks, *Model-Based Computer Vision*, Bowker Publishing Company, Essex, 1984

[129] S. Srivastava, K. van Leemput, F. Maes, D. Vandermeulen, P. Suetens, "Validation of Nonlinear Spatial Filtering to Improve Tissue Segmentation of MR Brain Images", *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, Lecture Notes in Computer Science, vol. 2208, pp. 507-515, 2001

[130] D. L. Pham, J. L. Prince, "A Generalized EM Algorithm for Robust Segmentation of Magnetic Resonance Images", *Proceedings of the 33rd Annual Conference on Information Sciences and Systems*, pp. 558-563, 1999

[131] K. van Leemput, F. Maes, D. Vandermeulen, P. Suetens, "Automated Model-Based Bias Field Correction of MR Images of the Brain", *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 885-896, 1999

[132] B. Likar, *Registration and Restoration of Medical Images*, PhD thesis, University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, p. 4, 2000

[133] F. A. Gerritsen, C. W. M. van Veelen, W. P. Th. M. Mali, A. J. M. Bart, H. L. T. de Bliek, J. Buurman, A. H. W. van Eeuwijk, M. J. Hartkamp, S. Lobregt, L. M. P. Ramos, L. J. Polman, P. C. van Rijen, C. P. Visser, "Some Requirements for and Experience with Covira Algorithms for Registration and Segmentation", *Medical Imaging*, IOS Press, Amsterdam, pp. 4-28, 1995

[134] S. D. Olabarriaga, A. W. N. Smeulders, "Setting the Mind for Intelligent Interactive Segmentation: Overview, Requirements, and Framework", *Proceedings of Information Processing and Medical Imaging*, pp. 417-422, 1997

[135] S. Dellepiane, "The Active Role of 2-D and 3-D Images: Semi-Automatic Segmentation", *Contemporary Perspectives in Three-Dimensional Biomedical Imaging*, IOS Press, Amsterdam, pp. 165-190, 1997

[136] A. L. Ratan, W. E. L. Grimson, "Training Templates for Scene Classification using a Few Examples", *Proceedings of the 1997 Workshop on Content-Based Access of Image and Video Libraries*, pp. 90-97, 1997

[137] K. W. Bowyer, P. Jonathon Phillips, "Overview of Work in Empirical Evaluation of Computer Vision Algorithms", *Empirical Evaluation Techniques in Computer Vision*, IEEE CS Press, Los Alamitos, CA, pp. 1-11, 1998

[138] K. W. Bowyer, "Validation of Medical Image Analysis Techniques", *Medical Image Procesing and Analysis 2000*, Handbook of Medical Imaging, vol. 2, pp. 567-607, SPIE Press, 2000

[139] R. A. Brooks, "Intelligence Without Reason", *The Artificial Life Route to Artificial Intelligence - Building Embodied, Situated Agents* (L. Steels, R. Brooks, editors), Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, pp. 23-81, 1995

[140] J. H. Holland, *Emergence - from Chaos to Order*, Oxford University Press, Oxford, 2000

[141] T. Kohonen, "Dynamically Expanding Context, with Application to the Correction of Symbol Strings in the Recognition of Continuous Speech", *Proceedings of the 8th International Conference on Pattern Recognition*, pp. 1148-1151, 1986

[142] T. Kohonen, "Self-Organization of Very Large Document Collections - State of the Art", *Proceedings of the 8th International Conference on Artificial Neural Networks*, vol. 1, pp. 65-74, 1998

[143] J. Iivarinen, M. Peura, J. Sarela, A. Visa, "Comparison of Combined Shape Descriptors for Irregular Objects", *Proceedings of the Eighth British Machine Vision Conference*, vol. 2, pp. 430-439, 1997

[144] S. Lawrence, C. L. Giles, A. C. Tsoi, A. D. Back, "Face Recognition: A Hybrid Neural Network Approach", technical report, no. UMIACS-TR-96-16 and CS-TR-3608, Institute for Advanced Computer Studies, University of Maryland, 1996

[145] J. Laaksonen, M. Koskela, E. Oja, "PicSOM - A Framework for Content-Based Image Database Retrieval using Self-Organizing Maps", *Proceedings of the 11th Scandinavian Conference on Image Analysis*, pp. 151-156, 1999

[146] K. L Ferguson, N. M. Allison, "Rate-Constrained Self-Organising Neural Maps and Efficient Psychovisual Methods for Low Bit Rate Video Coding", *Proceedings of the IEEE Signal Processing Society Workshop (Neural Networks for Signal Processing IX)*, pp. 390-399, 1999

[147] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*, Dover, New York, 1966

[148] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, *SOM_PAK: The Self-Organizing Map Program Package*, technical report, no. A31, Helsinki University of Technology, 1996

[149] J. W. Sammon, "A Non-Linear Mapping for Data Structure Analysis", *IEEE Transactions on Computers*, vol. 18, no. 5, pp. 401-409, 1969

[150] M. Blume, D. R. Ballard, "Image Annotation Based on Learning Vector Quantization and Localized Haar Wavelet Transform Features", *Applications and Science of Artificial Neural Networks III*, SPIE Proceedings, vol. 3077, pp. 181-190, 1997

[151] J. Alirezaie, M. E. Jernigan, C. Nahmias, "Neural Network based Segmentation of Magnetic Resonance Images of the Brain", *IEEE Transactions on Nuclear Science*, vol. 44, no. 2, pp. 194-198, 1997

[152] Y. Xiong, S. A. Shafer, "Variable Window Gabor Filters and Their Use in Focus and Correspondence", technical report, no. CMU-RI-TR-94-06, The Robotics Institute, Carnegie Mellon University, 1994

[153] I. T. Young, J. J. Gerbrands, L. J. van Vliet, "Fundamentals of Image Processing", PH Publications, Delft, The Netherlands, 1995

[154] M. Feng, X. Shaowei, "A Multiscale Approach to Automatic Medical Image Segmentation Using Self-Organizing Map", *Journal of Computer Science and Technology*, vol. 13, no. 5, pp. 402-409, 1998

[155] M. N. S. Swamy, K. Thulasiraman, *Graphs, Networks and Algorithms*, John Wiley and Sons, New York, p. 10, 1981

[156] D. C. Hoaglin, F. Mosteller, J. W. Turkey, *Understanding Robust and Exploratory Data Analysis*, John Wiley and Sons, New York, pp. 218-223, 1983

[157] C. F. Nielsen, P. J. Passmore, "A Solution to the Problem of Segmentation Near Edges Using Adaptable Class-Specific Representation", *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 1, pp. 436-440, 2000

[158] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, K. Torkkola, *LVQ_PAK: The Learning Vector Quantization Program Package*, technical report, no. A30, Helsinki University of Technology, Helsinki, 1996

[159] P. S. Williams, M. D. Alder, "Segmentation of Natural Images for CBIR", *Proceedings of the 14th International Conference on Pattern Recognition*, vol. 1, pp. 468-470, 1998

[160] M. Heath, S. Sarkar, T. Sanocki, K. W. Bowyer, "A Robust Visual Method for Assessing the Relative Performance of Edge Detection Algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, pp. 1338-1359, 1997

[161] C. F. Nielsen, P. J. Passmore, "Achieving Accurate Colour Image Segmentation in 2D and 3D with LVQ Classifiers and Partial Adaptable Class-Specific Representation", *Proceedings of the Fifth IEEE Workshop on Applications of Computer Vision*, pp. 72-78, 2000

[162] C. Fiorio, J. Gustedt, "Volume Segmentation of 3-Dimensional Images", technical report, no. 515/1996, Technische Universität Berlin, 1996

[163] M. J. Chantler, "Why Illuminant Direction is Fundamental to Texture Analysis", *IEE Proceedings of Vision, Image and Signal Processing,* vol. 142, no. 4, pp. 199-206, 1995

[164] L. Cinque, C. Guarra, S. Levialdi, "Reply: On the Paper by R. M. Haralick", *CVGIP: Image Understanding*, vol. 60, no. 2, pp. 250-252, 1994

[165] MIT Media Lab, *URL*,
http://www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html

[166] C. F. Nielsen, P. J. Passmore, "Pilot Study: Evaluating Re-Usability of Templates for Image Segmentation Using Adaptable Class-Specific Representation", *Proceedings of CAMDX*, technical report, no. CS-00-02, School of Computing Science, Middlesex University, pp. 67-77, 2000

[167] Z. Kato, "Bayesian Color Image Segmentation Using Reversible Jump Markov Chain Monte Carlo", technical report, no. 01/99-R055, European Research Consortium for Informatics and Mathematics (ERCIM), 1999

[168] Z. Kato, J. Zerubia, M. Berthod, "Bayesian Image Classification using Markov Random Fields", *Maximum Entropy and Bayesian Methods*, A. Mohammad-Djafari, G. Demoment (editors), Kluwer Academic, Dordrecht, pp. 375-382, 1996

[169] D. W. Shattuck, R. M. Leahy, "BrainSuite: An Automated Cortical Surface Identification Tool", *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2000*, Lecture Notes in Computer Science, vol. 1935, pp. 50-61, 2000

[170] K. van Leemput, F. Maes, D. Vandermeulen, P. Suetens, "Automated Model-Based Tissue Classification of MR Images of the Brain", *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 897-908, 1999

[171] Y. Zhang, M. Brady, S. Smith, "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm", *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45-57, 2001

[172] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, R. M. Leahy, "Magnetic Resonance Image Tissue Classification Using a Partial Volume Model", *NeuroImage*, vol. 13, no. 5, pp. 856-876, 2001

[173] S. Sandor, R. Leahy, "Surface-Based Labeling of Cortical Anatomy Using a Deformable Database", *IEEE Transactions on Medical Imaging*, vol. 16, no. 1, pp. 41-54, 1997

[174] C. F. Nielsen, P. J. Passmore, "Towards a Robust Path Growing Algorithm for Semi-Automatic MRI Segmentation", *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, Lecture Notes in Computer Science, vol. 2208, pp. 1382-1383, 2001

[175] C. F. Nielsen, P. J. Passmore, *Towards a Robust Path Growing Algorithm for Semi-Automatic MRI Segmentation*, technical report, no. CS-01-01, School of Computing Science, Middlesex University, London, 2001

[176] V. Kuperman, *Magnetic Resonance Imaging – Physical Principles and Applications*, Academic Press, San Diego, pp. 87-90, 2000

[177] C. F. Nielsen, P. J. Passmore, "Robust Semi-Automatic Segmentation of Single and Multi-Channel MRI Volumes through Adaptable Class-Specific Representation", *SPIE Medical Imaging 2002: Image Processing*, SPIE Proceedings, vol. 4684, part 3, pp. 1629-1639, 2002

[178] A. Lundervold, N. Duta, T. Taxt, A. K. Jain, "Model-guided Segmentation of Corpus Callosum in MR Images", *Proceedings of Computer Vision and Pattern Recognition*, pp. 231-237, 1999

[179] F. Hausdorff, *Set Theory*, Chelsea Publishing, New York, 1957

[180] S. Ishihara, *Tests for Colour-Blindness*, Kanehara Shuppan, Tokyo, 1963

[181] D. N. Kennedy, P. A. Filipek, V. S. Caviness, "Anatomic Segmentation and Volumetric Calculations in Nuclear Magnetic Resonance Imaging", *IEEE Transactions on Medical Imaging*, vol. 8, no. 1, pp. 1-7, 1989

[182] M. R. Spiegel, *Statistics*, McGraw Hill, New York, p. 263, 1999

[183] C. F. Nielsen, P. J. Passmore, P. Ziprin, A. Darzi, "Evaluation of Medical Image Segmentation: A Greater Role for Human Observer Experiments?", *Proceedings of the International Conference on Diagnostic Imaging and Analysis 2002*, pp. 442-449, 2002

[184] E. Wang, S. D. Snyder (editors), *Handbook of the Aging Brain*, Academic Press, London, 1998

[185] C. R. G. Guttmann, F. A. Jolesz, R. Kikinis, R. J. Killiany, M. B. Moss, T. Sandor, M. S. Albert, "White Matter Changes with Normal Aging", *Neurology*, vol. 50, no. 4, pp. 972-978, 1998

[186] J. L. Tanabe, D. Amend, N. Schuff, V. DiSclafani, F. Ezekiel, D. Norman, G. Fein, M. W. Weiner, "Tissue Segmentation of the Brain in Alzheimer's Disease", *American Journal of Neuroradiology*, vol. 18, January, pp. 115-123, 1997

[187] J. R. Mitchell, C. Jones, S. J. Karlik, K. Kennedy, D. H. Lee, B. Rutt, A. Fenster, "MR Multispectral Analysis of Multiple Sclerosis Lesions", *Journal of Magnetic Resonance Imaging (JMRI)*, vol. 7, no. 3, pp. 499-511, 1997

[188] C. G. Wible, M. E. Shenton, H. Hokama, R. Kikinis, F. A. Jolesz, D. Metcalf, R. W. McCarley, "Prefrontal Cortex and Schizophrenia", *Archives of General Psychiatry*, vol. 52, March, pp. 279-288, 1995

[189] F. J. Seinstra, D. Koelma, "Transparent Parallel Image Processing by way of a Familiar Sequential API", *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 4, pp. 824-827, 2000

Glossary

3DUS: 3D Ultrasound.

ACSR: Adaptable Class-Specific Representation. A framework for semi-automatic segmentation of medical imaging data of multiple modalities.

AFCM: Adaptive Fuzzy C-Means Algorithm. Fuzzy based automatic segmentation algorithm for brain MRI data developed by Pham and Prince [120].

AGEM: Adaptive Generalized Expectation Maximization algorithm. Automatic segmentation algorithm for brain MRI data developed by Pham and Prince [130]. Uses MRF models and a generalized EM algorithm to segment MRI data while correcting for inhomogeneity artefacts.

AID: Average Intensity Difference.

auto: In ACSR refers to the use of automatic template creation.

BrainWeb: An on-line database of simulated brain MRI scans with ground truth. Data simulated with varying levels of noise and inhomogeneity and available as T1, T2 and PD. Accessible through the Internet from the Brain Imaging Centre, Montreal Neurological Institute at McGill University. Homepage: http://www.bic.mni.mcgill.ca/brainweb.

BSE: Brain Surface Extractor. An algorithm for extraction of the cortical surface in MRI brain scans developed by Sandor and Leahy [172,173]. Uses anisotropic diffusion filtering followed by LoG boundary finding and morphological operations to create a brain mask.

CBIR: Content Based Image Retrieval.

CIELab: Colour appearance model specified by the CIE (Commission Internationale de l'Eclairage, the International Commission on Illumination). Based on psychometric colour matching experiments and considered perceptually uniform, i.e. colours which are close in parameter space are also perceptually close for human observers (as opposed to e.g. the RGB model).

CRT: Cathode Ray Tube. The CRT display (such as most current and older televisions) is still the most typical image display device (see also LCD).

CSF: Cerebrospinal Fluid.

CT: Computed Tomography. Also known as Computed Axial Tomography (CAT).

DFT: Discrete Fourier Transform. Discrete version of the Fourier transform suitable for computation (see FT).

DSA: Digital Subtraction Angiography. Angiography in which imaging of blood vessels is achieved by digitally subtracting the image before from the image after infusion of a contrast agent.

EM: Expectation Maximization. An iterative method first described by Dempster, Laird and Rubin [113]. Often used with MRF models for estimating model parameters from incomplete data.

EQ: Intensity equalisation algorithm for inhomogeneity correction of MRI data, developed at the Brain Mapping Division, UCLA. Source code available from: http://porkpie.loni.ucla.edu/BMD_ HTML/SharedCode/EQ/index.html.

f: In partial ACSR, the dilation factor expressed as the number of points added on each side of a given point in the original image (such as f1 for dilation of 1 point on each side).

FT: Fourier Transform. In digital imaging, transforms an image from spatial to frequency domain. The standard Fourier transform is continuous, but a discrete version (DFT) is used for digital images.

full: In ACSR refers to full ACSR, meaning that ACSR segmentation is applied at every point in an image or volume.

HLS: Hue, Lightness, Saturation. Colour model similar to HSV.

HMM: Hidden Markov Model. Statistical model describing the relation between input and output signals through chains of stable "hidden" states (Markov chains) and the probabilities of transitions between these states, which can be estimated from representative data (see also MRF).

HSV: Hue, Saturation, Value. Colour model which separates colour and intensity.

IBSR: Internet Brain Segmentation Repository. An on-line database of brain MRI scans from healthy and diseased individuals, acquired with a variety of scanners and acquisition modes (T1, T2 and PD). Most datasets available with ground truth. Accessible through the Internet for licensed users from the Center for Morphometric Analysis, Massachusetts General Hospital. Homepage: http://neuro-www.mgh.harvard.edu/cma/ibsr.

ICC: Intra-Class Correlation coefficient.

IIS: Intelligent Interactive Segmentation system.

KLT: Karhunen Loeve Transform. Also known as Principal Component Analysis (see PCA).

LCD: Liquid Crystal Display. Alternative to the older CRT display.

LHS: Short for CIELHS. Perceptually uniform colour appearance model specified by the CIE. Similar to CIELab (see CIELab).

LoG: Laplacian of the Gaussian. Edge detection operator developed by Marr and Hildreth [58]. Edges detected as zero crossings in the second derivative of an image following smoothing.

LUT: Look-Up Table.

LUV: Short for CIELUV. Perceptually uniform colour appearance model specified by the CIE. Similar to CIELab (see CIELab).

LVQ: Learning Vector Quantization. A supervised neural network similar to the SOM.

LVQ1: Learning Vector Quantization algorithm. Algorithm for training of LVQ neural network. Specifies the change made to the closest codebook vector depending on the input vector. The subtlety of change is determined by a global learning rate parameter $\alpha(t)$ which decreases as a function of time t (see appendix B).

M: In the PGA denotes the path length (number of vertices in a path excluding the end vertex which is the seed point).

MCMC: Markov Chain Monte Carlo. MRF based approach which uses Monte Carlo methods for parameter estimation.

median: In ACSR refers to the use of a median filter for post-processing of segmentation images.

MRI: Magnetic Resonance Imaging. Also known as Nuclear Magnetic Resonance Imaging (NMRI).

MRF: Markov Random Field. A 2D or 3D version of the Markov chain (see HMM). Used in image processing to model texture as an instantiation of a random field of intensities.

n: In the PGA denotes the number of dimensions in which paths are grown.

N: Noise. In simulated MRI volumes denotes the percentage level of added noise (such as 3N meaning 3% noise).

N3: The Non-parametric intensity Non-uniformity Normalization algorithm. A model based algorithm for inhomogeneity correction of MRI data, developed at the Brain Imaging Centre, Montreal Neurological Institute at McGill University. Source code available from: http://www.bic.mni.mcgill.ca/software/N3.

NURBS: Non-Uniform Rational B-Splines. A primitive suitable for visualising curved surfaces.

OLVQ1: Optimized Learning Vector Quantization algorithm. Algorithm for training of LVQ neural network. Similar to the original LVQ1 algorithm, but uses a local learning rate parameter $\alpha_i(t)$ (which decreases as a function of time t) for each node i rather than a global $\alpha(t)$ for faster convergence (see appendix B).

OPC: Opponent Process Colours. A colour model based on an approximation to Hering's opponent process theory of human colour vision.

partial: In ACSR refers to partial ACSR, meaning that ACSR segmentation is applied only at boundary points found in an initial segmentation step using a non-ACSR method.

PBNN: Point Based Nearest Neighbour classifier. A template based nearest neighbour classifier which matches single points with class templates based on colour descriptors.

PCA: Principal Component Analysis. Method for reducing the dimensionality of vectorial input data into its eigenvectors.

PD: Proton Density. MRI acquisition protocol reducing the effect of T1 T2. PD-weighed images show intensities proportional to the density of mobile protons equivalent to water content (e.g. urine, CSF).

PET: Positron Emission Tomography.

PGA: Path Growing Algorithm. Segmentation algorithm which implements ACSR. Uses topologically different spatial representations of segment classes at every point, which compete for the final classification. Representations are built from paths, which are acyclic chains of points originating from the point to classify (the seed point).

PGA-PD: Path Growing Algorithm with Path Descriptors (path median and AID for the MRI modality).

PGA-SPD: Path Growing Algorithm with Path Descriptors and Single path representation.

PGA-SPDS: Path Growing Algorithm with Path Descriptors and Single path representation with seed point Shifting for AID calculation.

PGA-DPDS: Path Growing Algorithm with Path Descriptors and Double path representation with seed point Shifting for AID calculation.

RF: Radio Frequency.

RFI: Radio Frequency Inhomogeneity. In simulated MRI volumes denotes the percentage level of added intensity inhomogeneity (such as 20RFI meaning 20% inhomogeneity).

RGB: Red, Green, Blue. Colour model developed for display devices.

RJMCMC: Reversible Jump Markov Chain Monte Carlo. Extension of MCMC first described by Green [112]. Jumps between parameter subspaces of different sizes, allowing for an automatic estimation of the number of classes.

ROI: Region of Interest.

SAD: Sum of Absolute Distances.

SGI: Silicon Graphics Incorporated.

SNR: Signal to Noise Ratio.

SOM: Self-Organizing Map. Unsupervised self-organising neural network. Also known as the Self- Organizing Feature Map (SOFM).

SPECT: Single-Photon Emission Computed Tomography.

SSD: Sum of Squared Distances.

STFT: Short Term Fourier Transform. Modification of the Fourier transform using windows of finite size.

T1: MRI acquisition protocol, referring to a specific relaxation time for the protons after the RF signal has been removed. T1-weighed images show good contrast between grey and white matter. Watery substances (such as CSF) appear dark.

T2: MRI acquisition protocol, referring to a specific relaxation time for the protons after the RF signal has been removed (shorter than T1). T2-weighed images do not have as good contrast between grey/white matter as T1, but watery substances (e.g. CSF, diseased tissues such as cysts) appear very bright with good contrast.

TE: Time to Echo. In MRI the time in milliseconds between the application of an RF pulse and the peak of the echo signal.

TR: Time to Repetition. In MRI the time in milliseconds between each RF pulse sequence applied to the same slice.

VHP: Visible Human Project. A collection of multimodal datasets of cryo section, MRI and CT data, available from the United States National Library of Medicine for licensed users.

VTK: The Visualization Toolkit. A C++ library developed by Kitware Incorporated for graphics visualisation.

Appendix A

Publications

A.1. Overview of publications.

This appendix contains the seven papers published to date on ACSR segmentation:

A.2. C. F. Nielsen, P. J. Passmore, "A Solution to the Problem of Segmentation Near Edges Using Adaptable Class-Specific Representation", presented at the 15th IEEE International Conference on Pattern Recognition (ICPR), Barcelona, 2000 [157].

A.3. C. F. Nielsen, P. J. Passmore, "Pilot Study: Evaluating Re-Usability of Templates for Image Segmentation Using Adaptable Class-Specific Representation", Middlesex University School of Computing Science technical report, 2000 [166].

A.4. C. F. Nielsen, P. J. Passmore, "Achieving Accurate Colour Image Segmentation in 2D and 3D with LVQ Classifiers and Partial Adaptable Class-Specific Representation", presented at the Fifth IEEE Workshop on Applications of Computer Vision (WACV), Palm Springs, 2000 [161].

A.5. C. F. Nielsen, P. J. Passmore, "Towards a Robust Path Growing Algorithm for Semi-Automatic MRI Segmentation", presented at Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2001, Utrecht, 2001 [174].

A.6. C. F. Nielsen, P. J. Passmore, "Towards a Robust Path Growing Algorithm for Semi-Automatic MRI Segmentation" (extended version of the MICCAI paper), Middlesex University School of Computing Science technical report, 2001 [175].

A.7. C. F. Nielsen, P. J. Passmore, "Robust Semi-Automatic Segmentation of Single and Multi-Channel MRI Volumes through Adaptable Class-Specific Representation", presented at SPIE Medical Imaging 2002, San Diego, 2002 [177].

A.8. C. F. Nielsen, P. J. Passmore, P. Ziprin and A. Darzi, "Evaluation of Medical Image Segmentation: A Greater Role for Human Observer Experiments?", presented at the International Conference on Diagnostic Imaging and Analysis (ICDIA) 2002, Shanghai, 2002 [183].

Appendix B

Vector quantization neural networks and fuzzy logic

B.1. SOM and LVQ.

If m_i denotes a codebook vector and m_c denotes the nearest codebook vector to the input vector x at the time t, then the learning function for the basic LVQ1 algorithm can be defined as (from [117]):

If x and m_c belong to the same class:

$$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)] \quad \textbf{(B.1)}$$

If x and m_c belong to different classes:

$$m_c(t+1) = m_c(t) - \alpha(t)[x(t) - m_c(t)] \quad \textbf{(B.2)}$$

For $i \neq c$:

$$m_i(t+1) = m_i(t) \quad \textbf{(B.3)}$$

$\alpha(t)$ is the learning rate parameter, which determines the subtlety at which adjustments are made to the codebook vectors. $0 < \alpha(t) < 1$ and usually decreases monotonically with time. In the Optimized Learning Vector Quantization algorithm (OLVQ1) an individual $\alpha_i(t)$ is assigned to each m_i, resulting in faster convergence:

If x and m_c belong to the same class:

$$m_c(t+1) = m_c(t) + \alpha_c(t)[x(t) - m_c(t)] \quad \textbf{(B.4)}$$

If x and m_c belong to different classes:

$$m_c(t+1) = m_c(t) - \alpha_c(t)[x(t) - m_c(t)] \quad \textbf{(B.5)}$$

For $i \neq c$:

$$m_i(t+1) = m_i(t) \quad \textbf{(B.6)}$$

$\alpha_i(t)$ changes as a function of the previous value:

$$\alpha_c(t) = \frac{\alpha_c(t-1)}{1+s(t)\alpha_c(t-1)} \tag{B.7}$$

$s(t)$ = 1 if x and m_c belong to the same class.

$s(t)$ = -1 if x and m_c belong to different classes.

Thus in order to avoid the situation where $\alpha_i(t)>1$, the initial value of all $\alpha_i(0)$ should be less than 0.5.

When the network has been trained through a number of training cycles, novel feature vectors may be compared to the codebook vectors. The closest match (minimum quantisation error) is the winning codebook and its label is the classification of the input vector. The ultimate purpose of LVQ is to precisely define the class borders in an arbitrary input set, given that some principal features will emerge from each of the classes in the training data during the learning phase.

The SOM is unsupervised, which means that no a priori knowledge is given to the network in the form of labelled input samples. A SOM can however be calibrated with labelled data sets after learning is completed. The basic SOM learning rule is similar to the LVQ1 algorithm except that the class of the winning node is unknown and a modification is made not only to the winning node but to its neighbourhood. The effect decreases with distance to the winning node and the topology of the neighbourhood depends on the chosen lattice type (usually rectangular or hexagonal). This way clusters are formed from randomly initialised nodes. The basic learning rule is:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{B.8}$$

$h_{ci}(t)$ is the neighbourhood function or kernel. The subscript c denotes the node with the closest match to an input $x(t)$ (similarly to LVQ above).

A number of different neighbourhood kernels can be used, typically a Gaussian:

$$h_{ci} = \alpha(t) * \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)} \right) \qquad \textbf{(B.9)}$$

$\alpha(t)$ is the learning rate parameter. r gives the location of a node on the lattice. σ defines the width of the kernel.

A lattice type, number of nodes, the dimensions of the lattice and the type of neighbourhood kernel must be given before learning. These parameters, particularly the dimensions, may significantly affect the outcome of the learning. Therefore another way of approximating to the probability density function should be used first to give an idea about the ideal values. Kohonen [117,148] recommends Sammon's mapping [149] for this purpose. The SOM offers fully automatic clustering of arbitrary input data, which is attractive in many applications. The best way to select the optimal parameters for a particular application (using a particular set of samples) for a SOM is however not well defined. This is not desirable if consistency is important and the nature of the clustering task is constantly changing. Rather than using the SOM directly, it might favourably be used to test the efficiency of a particular feature vector encoding before applying it to a LVQ network. Because codebook vectors are interchangeable between SOM and LVQ, the initialisation of codebook vectors prior to learning may even be carried out using a SOM before learning is started using LVQ. Both SOM and LVQ has been used extensively in many types of pattern recognition, including speech recognition, character recognition and feature detection in images (see [117] for an overview).

B.2. Sammon's mapping.

Sammon's mapping [149] is an automatic clustering algorithm which produces results comparable to the SOM [117]. It is a non-linear projection from a high to a lower dimensional space which preserves local relations in the data. Sammon's mapping tries to map coordinates in the high dimensional space to the low dimensional space in

such a way that the error function E is minimised. Let d_{ij} denote the distance between two points i and j in the high dimensional input space. d'_{ij} denotes the distance between the projected points in the low dimensional space. E is then given by [149]:

$$E = \frac{1}{\sum_i \sum_{j>i} d_{ij}} \sum_i \sum_{j>i} \frac{\left(d_{ij} - d'_{ij}\right)^2}{d_{ij}} \qquad \textbf{(B.10)}$$

As the error function expresses, local relations between every single point and all other points in the input space are preserved as well as possible in the output space. Since the optimisation depends on finding the best topology in the low dimensional space, the topology changes for every iteration. The major dimensions of the final topology can be used to determine the appropriate dimensions of the lattice for a SOM classifier.

B.3. Fuzzy logic.

In a fuzzy set the possibility of an element belonging to a set decreases linearly on both sides of the lower and upper limits in a crisp set (i.e. a set in classical set theory). For example if a particular intensity level in a sample lies between 100 and 200 in a crisp set corresponding to texture class A, then an intensity of 95 or 205 means that the sample definitely does not belong to class A. An intensity of 95 or 205 may mean that the sample definitely belongs to class B or C. If these limits are well established, then crisp sets is the preferred choice. However if there is a possibility that the class borders may be slightly plastic, and that the actual classification also depends on the membership of other sets for other descriptors in the same sample, then fuzzy theory may provide a better answer. In the fuzzy set an intensity between 100 and 200 would yield the possibility value 1. An intensity of 95 or 205 would *not* yield a possibility of 0, but e.g. of 0.5 (depending on the plasticity). A set of rules (fuzzy rules) specifying the relationships between variables in the domain (expressed in Boolean terms) must be found. Given actual values of some variables it is then possible to determine the possibility of an unknown variable lying within a certain range. For example, consider

a photograph of a homogeneous yellow object on a green background. In the boundary region the red channel will vary between zero and a maximum value. Assume that sets may be labelled as *zero*, *low* or *high*. A rule may specify that if the red channel is *high* AND the gradient magnitude is *zero* at a specific point THEN the point belongs to the *yellow object* class. Another rule could specify that if the red channel is *low* AND the gradient magnitude is *low* THEN the point belongs to the *boundary* class. In fig. B.1 an actual value is given to the amount of red (high but not maximum) and gradient magnitude (low but not zero) of a point. Intersecting the actual gradient magnitude with the graph for the gradient magnitude set *zero*, we get value *n*. Intersecting the actual level of the red channel for the red set *high* we get value *m*. Since the rule specified an AND operation, the possibility for the class set *yellow object* is min(*n*,*m*). This can be repeated for the gradient magnitude set *low* and the red set *low* according to the rules. Given a number of rules, a number of different possibilities for the unknown variable results for the same well known variables. These can be combined to find the overall possibility of the unknown variable belonging to a particular set. This process is known as defuzzification and may be based on heuristics or properties of the graph, such as the centre of gravity.

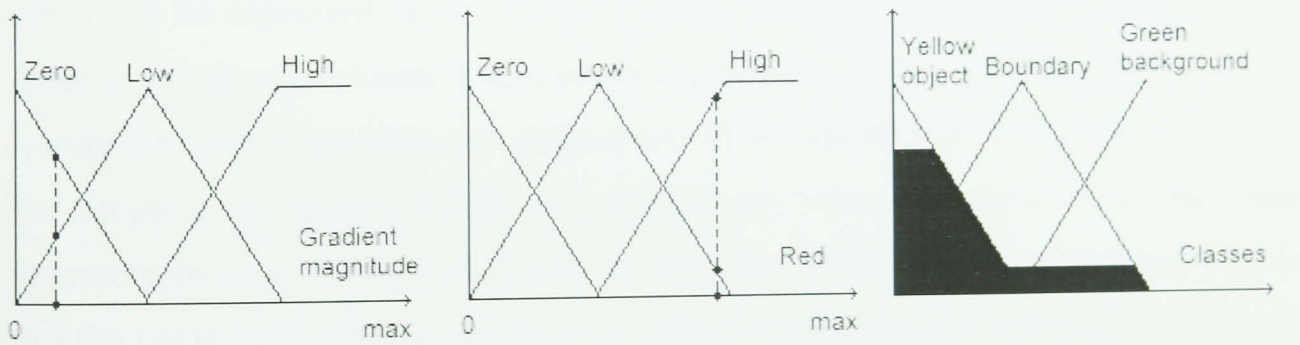


Fig. B.1. Example of fuzzy classes for image classification. (a) The sets for gradient magnitude with an actual value intersecting sets *zero* and *low*. (b) The sets for the red channel with an actual value intersecting sets *low* and *high*. (c) The resulting graph showing the possibilities of classes based on the graphs in (a) and (b).

Appendix C

Computational overhead and implementation of the PGA

C.1. Complexity and computational overhead of the PGA.

The computational overhead of using the PGA can be described by the number of representations that are created per point. This number depends on the path length M and the dimensionality n of the coordinate system, in which the paths exist. Ideally the number of paths should be expressed as a function of these two parameters. This has however proven to be a hard problem. It is trivial that as long as $M<4$ it is impossible that any chain of pixels will be cyclic when a $2n$-connected expansion is used. In that case the only condition, which has to be satisfied, is that the growth does not create the same edge twice (the word edge is used here in a graph theory context). The previous vertex cannot become the next vertex. In other words backtracking must be avoided. This simple constraint is easily incorporated into a formula for the total number of paths by giving $2n-1$ different directions of growth from each point, subsequent to the seed point, rather than $2n$. The total number of paths for $M<4$ and any n can be expressed as $2n*(2n-1)^{M-1}$. However, when $M>=4$ it is possible that a chain of pixels may become cyclic using the $2n$-connected expansion. If $M=4$ then it is only the cases when both end-vertices are the seed point that need to be eliminated. For longer path lengths though, it could be any repeated vertex where the chain intersects an edge, which has already been grown. What makes the problem hard is that the value of M alone at any given n does not correspond to only one topology. It corresponds to many, some of which are cyclic and some of which are not, when $M>=4$. There are particular regular topologies such as straight lines and zigzags, which are known to always be acyclic. Unfortunately these only account for some of the chains, which cannot be considered as paths. So far no complete solution has been found. The number of paths (counting only those connected subgraphs which are *acyclic*) can however be found by a computer program through multiple recursion. Each direction from a point is a separate function, which can call itself as well as all other directions except the opposite one (e.g. "up" cannot call "down"). This prevents

backtracking. The current number of vertices included in the path is passed with each new function call and a map is updated to keep track of vertices already included. This map is checked to verify the validity of a direction before its function is called. When $M+1$ vertices is reached, no new function call is made. Control is passed back to the previous function, which must now call another direction. This way the full search space is exhausted and all possible paths can be found. While a generic solution to the problem is still desirable, the number of paths per point can be found when M and n are known. As we are only ever likely to consider the 2D and 3D case, we can settle for these two values only for n. The value of M has varied between 3 and 5 in experiments, although all results presented in this thesis were achieved with $M=5$. Excluding cyclic cases is important to maintain consistency in the neighbourhood representation and also helps reduce the number of calculations needed per point. To further reduce computational overhead a subset of paths can be removed. Those are the paths which, although not cyclic, are still identical in terms of their vertices, but where the order in which they are grown differs (fig. C.1). Because we are simply considering the individual vertices of paths, the direction of growth is not important. We can thus eliminate paths with identical sets of vertices. Table C.1 shows the number of possible paths for values of M up to 8 in 2D and 3D and the number of paths with unique vertex sets.

An additional factor, which affects processing time, is the template matching, which occurs for each point for each texture class. This matching only needs to be carried out once and stored (it should be noted that this is only true for colour images - see section C.2). It can then be used as a look-up table for the calculation of path values. In the first implementation of the PGA an exhaustive linear search was employed. This has since been replaced by a faster algorithm, which does not have to traverse the entire search space to find the best match. The closer a novel feature set is to a template, the faster the algorithm finds the closest match. It follows that the processing time required per point for matching varies from point to point and also depends on the template sets. Examples of segmentations and their required processing time are shown in chapter 5, section 5.3.

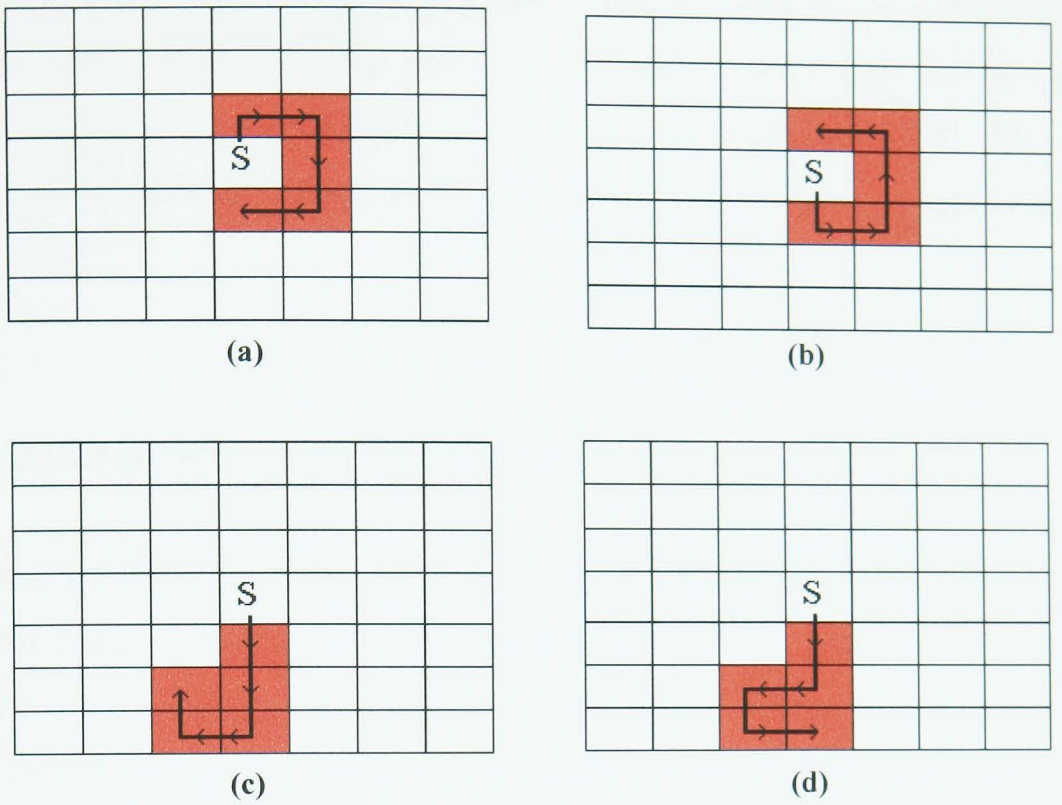


Fig. C.1. Different directions of growth from the seed point producing the same vertex sets in the PGA: (a) and (b), (c) and (d). Path length is set to 5.

Table C.1. Number of paths and their unique vertex sets for n dimensions with path length M in the PGA.

n	M	Paths	Unique vertex sets
2	3	36	32
2	4	100	92
2	5	284	248
2	6	780	696
2	7	2172	1872
2	8	5916	5169
3	3	150	138
3	4	726	678
3	5	3534	3162
3	6	16926	15266
3	7	81390	71498
3	8	387966	341421

Although the PGA has a high computational overhead, an important feature is that it is also highly parallelisable, in fact down to single pixel level. The PGA would be well suited for implementation on multi-processor architectures. The jump from sequential to parallel code might be easily achieved using a system such as the one

described in [189]. The PGA would possibly also be suitable for implementation in dedicated hardware. At a more basic level, a segmentation task can be split over several single-processor machines. This is particularly useful for volume segmentation (see chapter 4, section 4.7), where each machine can segment a separate sub-volume.

C.2. Calculation versus look-up of nearest neighbour match values.

The fastest way of obtaining nearest neighbour match values is through the use of pre-calculated Look-Up Tables (LUTs). Based on a template set the closest matches for all possible combinations of descriptors within the given ranges of the model used (colour model or greyscale range) can be placed in a look-up table. It is possible to do this e.g. for greyscale images using only 1 point descriptor. However if path descriptors are used (requiring the matching of combinations of 3 values) or in the case of colour images, where 3 point descriptors are generally used, the memory requirements and pre-processing time needed makes a purely LUT based solution practically impossible (assuming non-quantised information).

For colour images, where only point descriptors are used, the nearest match for *each image point* can be pre-calculated and the results placed in a LUT, where each entry point corresponds to a point in the image or volume. This LUT can then be used to calculate the path values for any combination of points. As mentioned above though, the actual matching itself cannot be LUT based.

When path descriptors are required for greyscale images, the combinations of points from which the descriptors are derived again make a LUT approach (even on an image point basis) practically impossible. Although the nearest neighbour match of the point descriptors can be LUT based, contrary to when colour images are used, the path descriptors must be calculated not just for every point, but for every path at every point. This means that the use of path descriptors is substantially more processor intensive than the use of point descriptors only. It is however at this time considered

to be a necessity for certain image modalities, providing a powerful means of adaptive filtering.

C.3. A fast algorithm for nearest neighbour match.

A naive nearest neighbour algorithm for template matching can be implemented as a linear search through the entire data set of templates. This is time consuming and often unnecessary. A faster algorithm for nearest neighbour match using n descriptors is employed. This algorithm seeks to reduce the search space. The algorithm is explained below in the context of its software implementation.

Pre-processing stage:

1. For template t each set of n descriptors are stored in an array A.
2. n arrays of pointers are created. Array $a_d \in [a_0,...,a_{n-1}]$ contains pointers to all entries in A in sorted order according to the value of descriptor d using an optimal bubble sort.
3. $LUT_d \in [LUT_0,...,LUT_{n-1}]$ is created. Each table provides three entries for each look-up value, where the look-up value may be any value within the range of descriptor d to be matched. The entries are:

    - Distance $dist_d$ to closest match with descriptor d
    - Position pos_d of closest match in array a_d
    - Number of times f_d the closest match appears in A (stored consecutively in array a_d) more than once

    If $dist_d$ is the same for two different values in A then pos_d is set to the lowest of the two matching values and f_d will include appearances of both values (i.e. one consecutive range of positions in array a_d covering both values).

Matching of novel sets of descriptors:

1. A novel set of descriptors $nov = (novd_0, ..., novd_{n-1})$ is given.

2. $novd_d$ is used as look-up value for LUT_d to obtain $dist_d$, pos_d and f_d for all d.

3. All entries of descriptors $entd_d$ in A pointed to by entries in a_d from pos_d to $pos_d + f_d$ are traversed for all d. At every entry of every traversal a match value is calculated as:

$$match_e = \sum_{x \neq d} |novd_x - entd_x| \qquad \textbf{(C.1)}$$

The minimum match value is found as $minmatch_d = \min(match)$ for all d. Notice that this value expresses the closest match with all *other* descriptors when descriptor d is equal to its closest match value(s) in A.

4. A search space for each a is defined by a lower and an upper search limit. The total best match for all descriptors other than d is calculated for all d as:

$$bestoth_d = \sum_{x \neq d} dist_x \qquad \textbf{(C.2)}$$

The difference between this value $bestoth_d$ and the best match for all descriptors other than d, $minmatch_d$, when descriptor d is equal to its closest match, represents the theoretical maximum decrease of the total closest distance when a search is carried out in either direction from pos_d. This value is expressed as $maxdist_d = minmatch_d - bestoth_d$. In other words it expresses the maximum increase and decrease of the value pointed to by pos_d , which may define the lower and upper limits for the search space in which the total closest distance can be found. At the distance $maxdist_d$ from pos_d the decrease in distance to other descriptors from finding $bestoth_d$ is cancelled out by the increase in distance to descriptor d. Therefore the closest distance cannot exist outside these limits.

5. Let $matchvalue_d$ be the actual value of descriptor d in the entry in a_d pointed to by pos_d. The value $templow_d$ is expressed as $templow_d = matchvalue_d - maxdist_d$. Let $minval$ be the minimum possible value of descriptor d. If $templow_d < minval$ then $templow_d = minval$. $templow_d$ is used as a look-up value for LUT_d to find pos'_d . The lower limit $limlow_d$ for search in a_d is

defined as $limlow_d = pos'_d$. The value $temphigh_d$ is expressed as $temphigh_d = matchvalue_d + maxdist_d$. Let $maxval$ be the maximum possible value of descriptor d. If $temphigh_d > maxval$ then $temphigh_d = maxval$. The value $temphigh_d$ is used as a look-up value for LUT_d to find pos''_d and f''_d. The upper limit $limhigh_d$ for search in a_d is defined as $limhigh_d = pos''_d$ and $limf_d = f''_d$.

6. The smallest search space is found as:

$$sspace = \min_{d=0\ldots n-1} \left[limhigh_d + limf_d - limlow_d \right] \qquad \textbf{(C.3)}$$

Let the value w be equal to the value of d yielding *sspace*.

7. The nearest neighbour match is carried out using the smallest search space:

$$totalmatch = \min_{x=limlow_w \ldots limhigh_w + limf_w} \left[\sum_{y=0}^{n-1} \left| A[a_w[x]][y] - novd_y \right| \right] \qquad \textbf{(C.4)}$$

The notation $A[a_w[x]][y]$ is interpreted as descriptor y in the entry of A at position $a_w[x]$.

Appendix D

Human observer experiments

D.1. Written material for human observer experiments.

The following pages show the written material which was used for the human observer experiments described in chapter 7.

In the experiment using natural colour images, participants were required to read and sign the document shown on pages 236-237. The relative order of segmentations for each image (specified by a participant placing the printed segmentations on a table from left to right in order of preference), was recorded manually by the investigator using the form on page 238. This was achieved by noting down the numbers at the back of each printed segmentation, identifying the algorithm (hidden from the participant). The form on pages 239-240 was used by participants for the absolute ranking. The desired grade for each relative position was marked with an X. By comparing the positions of the absolute grades on this form with the positions of each segmentation recorded by the investigator, grades and algorithms could be matched.

The document on page 241 was given to and signed by all participants in the human observer experiments using cryo section and MRI data. An additional sheet shown on page 242, with information about the two IBSR volumes, was given to radiologists ranking the MRI segmentations. Results in these experiments were recorded automatically (see section D.2).

The companion CD contains all source images/volumes and segmentations from the human observer experiments, and includes a similar software interface for viewing the medical images to what was used in the computer based tasks.

Consent form given to participants in the human observer experiment based on natural colour images.

Middlesex University Department of Computer Science
Research Consent Form

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

Research Project Title
A Robust Framework for Multi-Modal Medical Image Segmentation through Adaptable Class-Specific Representation.

Researcher
Casper F. Nielsen

Experiment Purpose

The purpose of this experiment is to use human observers to evaluate the segmentation quality of 6 standard machine vision colour images of natural scenes. Several segmentations of each image based on different segmentation algorithms are presented to the observers (hereafter referred to as "participants"). The results obtained from this experiment will help clarify issues related to the segmentation of medical colour cryo section data using non-medical standard test images where medical equivalents do not currently exist. Note that no medical images will be used in this experiment.

Participant Recruitment and Selection

Participants are undergraduate and/or postgraduate students at the School of Computing Science, Middlesex University, London, U.K. No particular experience is required, but participants must have normal colour vision.

Procedure

The experiment is expected to take between 30 and 40 minutes. Each participant will undergo a short paper and screen-based version of Ishihara's test for colour blindness. This test must be passed in order for participants to proceed to the main tasks. Each of the 6 images constitute a separate task in which participants will be presented with a colour image displayed on a computer monitor. Subsequently the participant will be given 6 printed segmentation images on paper, which must be arranged in order of preference (based on the perceived quality of segmentation). Following this relative ranking the participant is required to perform an absolute ranking by grading each segmentation image on an ordinal scale using a form supplied by the investigator on a sheet of paper. The 6 tasks will be completed in random order for each participant. Note that the test for colour blindness should not be considered as a precise diagnosis. Only a trained ophthalmologist can diagnose colour blindness.

Data Collection

No specifics of the colour test will be recorded. Participants who proceed to the main tasks will have successfully completed the colour test and no data will be recorded for participants who have not passed the test. The relative ranking of each image will be recorded by the investigator and the absolute rankings recorded by the participants will be collected and used. No part of the conversation between participants and the investigator will be recorded.

Confidentiality

The written records of collected data will not contain any information to identify the participants from whom the data originates. No participant will be identified in any publication containing results from this experiment.

Likelihood of Discomfort

Participants are not expected to feel any discomfort as a result of this experiment. Although a computer monitor will be used, participants will not be required to watch the monitor continuously throughout the experiment, so no discomfort should be caused by the monitor. Note however that computer monitors have been known to provoke seizures in individuals suffering from epilepsy and any history of epilepsy should be reported to the investigator prior to the experiment. Glasses or contact lenses may be worn during the experiment provided that they give full correction of any deficiencies in a participant's vision.

Researcher

Casper F. Nielsen is a PhD student and part-time lecturer at the School of Computing Science, Middlesex University, London, U.K. His research is in the development of usable systems for multi-modal medical image segmentation.

Finding out about Results

Results may be obtained from the investigator by request in writing to the following e-mail address: c.nielsen@mdx.ac.uk.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive you legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to not answer specific items or questions in interviews or on questionnaires. You are free to withdraw from the study at any time without penalty. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact the researcher.

_____ _____

Participant Date

_____ _____

Investigator/Witness Date

A copy of this consent form has been given to you to keep for your records and reference.

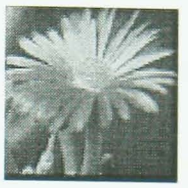
Form for recording the relative order of preference for segmentations of natural colour images. Used by the investigator.

ACSR Visual Ranking – Ordered Sequences

ID: _____

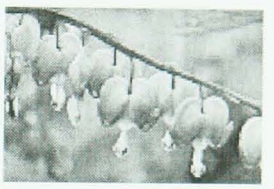
Poppy

1	2	3	4	5



Hearts

1	2	3	4	5

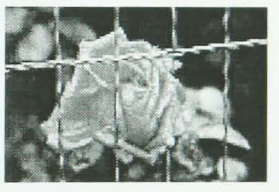


Seagull

1	2	3	4	5

Rose

1	2	3	4	5



Bird11

1	2	3	4	5

Bird12

1	2	3	4	5

Form for recording the absolute grades for all segmentations of each individual natural colour image. Used by participants.

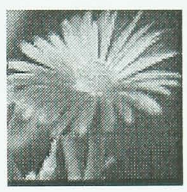
ACSR Visual Ranking – Absolute Grading

ID: __________

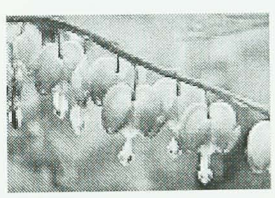
Guide to grades:

1 = Image shows no coherent representation of your perceived ideal segmentation

7 = Image matches your perceived ideal segmentation



1 (left)							2							3							4							5 (right)						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7



1 (left)							2							3							4							5 (right)						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7

1 (left)							2							3							4							5 (right)						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7

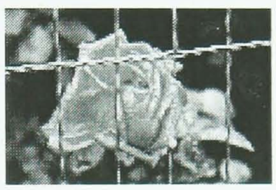
ACSR Visual Ranking – Absolute Grading

ID: __________

Guide to grades:

1 = Image shows no coherent representation of your perceived ideal segmentation

7 = Image matches your perceived ideal segmentation



1 (left)							2							3							4							5 (right)						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7

1 (left)							2							3							4							5 (right)						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7

1 (left)							2							3							4							5 (right)						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7

Consent form given to participants in the human observer experiments using cryo section and MRI data .

Research Consent Form

Research Project Title

A Robust Framework for Multi-Modal Medical Image Segmentation through Adaptable Class-Specific Representation.

Researcher

Casper F. Nielsen, Vision and Image Processing Group, Middlesex University.
E-mail: c.nielsen@mdx.ac.uk, tel.: 020 8411 6714

Experiment Purpose

The purpose of this experiment is to evaluate the ACSR segmentation framework through visual ranking. Participants are surgeons or radiologists. A small number of participants will be required to carry out the initialisation of ACSR segmentation for a cryo section or MRI volume combined with one ranking task. The remaining participants will take part in two ranking tasks involving cryo section data (surgeons) or three ranking tasks involving MRI data (radiologists). All tasks are computer based. Initialisation involves the use of a mouse while the ranking tasks are controlled using a standard keyboard. All results are recorded electronically throughout the experiment.

Likelihood of Discomfort

Participants are not expected to feel any discomfort as a result of this experiment. Note however that computer monitors have been known to provoke seizures in individuals suffering from epilepsy and any history of epilepsy should be reported to the investigator prior to the experiment. Glasses or contact lenses may be worn during the experiment provided that they give full correction of any deficiencies in a participant's vision.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive you legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions concerning matters related to this research, please contact the researcher.

_____ _____

Participant Date

_____ _____

Investigator/Witness Date

Information sheet about IBSR volumes given to participants in the human observer experiments using MRI data.

MRI Info Sheet

Acquisition details for MRI volumes:

The MRI scans were acquired with a 1.5 Tesla General Electric Signa. Contiguous 3.0 mm three-dimensional coronal T1-weighted spoiled gradient echo (SPGR) images of the brain were attained with the following parameters: TR = 40 msec, TE = 5 msec, flip angle = 40 degrees, field of view = 24cm, matrix = 256x256, and averages = 1

- **Subject: 55 year old male.**

- **Subject: 5 year old male.**

D.2. Computer based experiments.

The experiments described in chapter 7 using cryo section and MRI data were entirely computer based. Images were displayed on a CRT monitor using custom made software and the results of participants' interactions were recorded by the software and summarised in a file. The software was also responsible for the randomisation of individual blocks of tasks as well as viewports on screen assigned to individual volumes and the starting slice numbers. The software was written by the author in C++ using VTK libraries [11] for Microsoft Windows (compiled under Windows NT).

The physical setup used for all the computer based experiments is shown in fig. D.1. The software was run on a laptop connected to an external monitor and a mouse for the template selection task (cryo hip bone volume) or a keyboard for the ranking tasks. The investigator could control the software using the internal keyboard and pointing device on the laptop, which also duplicated the display of the external monitor on the internal LCD display.

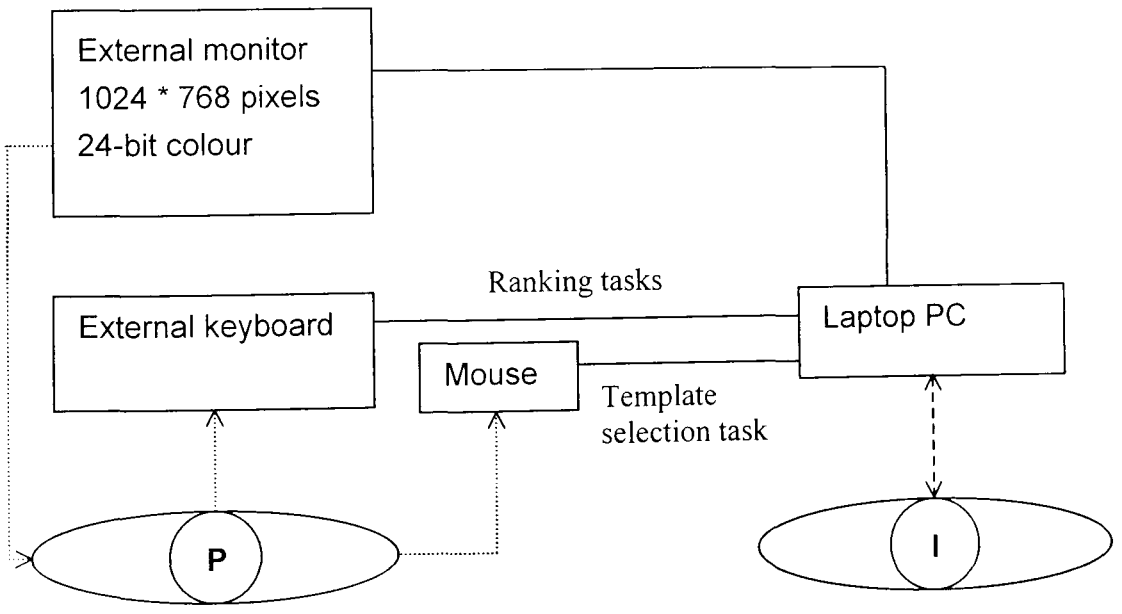
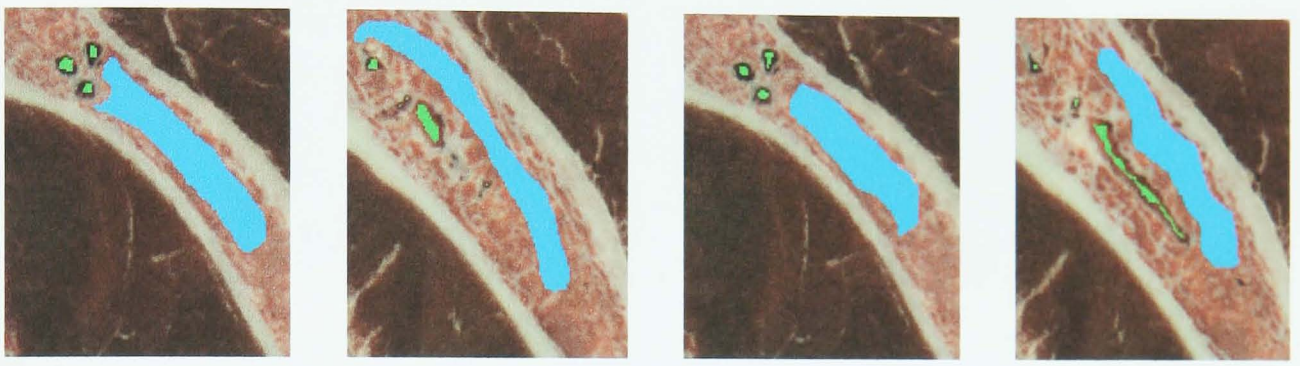


Fig. D.1. Schematic setup (viewed from above) used for computer based human observer experiments. Participants (P) interacted with the external keyboard and mouse and viewed tasks on the external monitor. The investigator (I) interacted directly with the laptop running the experiment software.

The template selection for the cryo hip bone segmentation was carried out in Paint Shop Pro v. 5.01 from JASC. Fig. D.2. shows examples of template selection from two participants. Fig. D.3 shows a screenshot from the ranking task using the segmented hip bone volumes. Each viewport showed a specific segmentation or the source data (centre) and participants could browse through each individual slice. Fig. D.4 shows a screenshot from the cryo brain series segmentation task, where each column corresponded to the same slice and each row corresponded to the same segmentation.



(a) (b) (c) (d)

Fig. D.2.Examples of template selection in the cryo hip bone experiment from two participants. (a-b) The two templated slices from one participant. (c-d) The two templated slices from another participant. Blood vessels and bone marrow selected in pseudocolours.

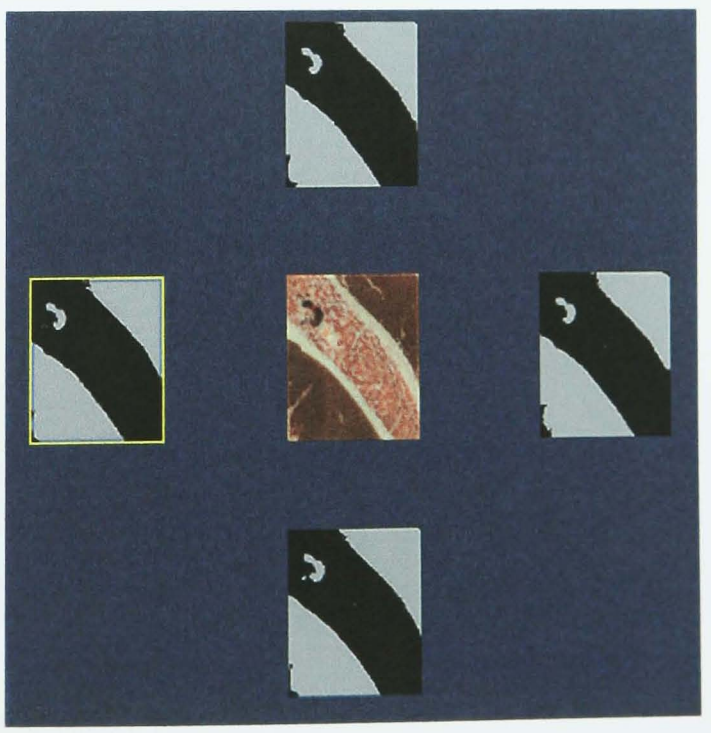


Fig. D.3. Screenshot from the cryo hip bone segmentation ranking task. Source volume in the centre, segmentation based on four different template sets surrounding. Leftmost segmentation volume currently selected (yellow selection box).

Fig. D.4. Screenshot from the cryo brain series segmentation task. Source images in the top row. Each row underneath corresponds to a separate segmentation. Two rows are selected for swapping (in the process of indicating the desired relative order).

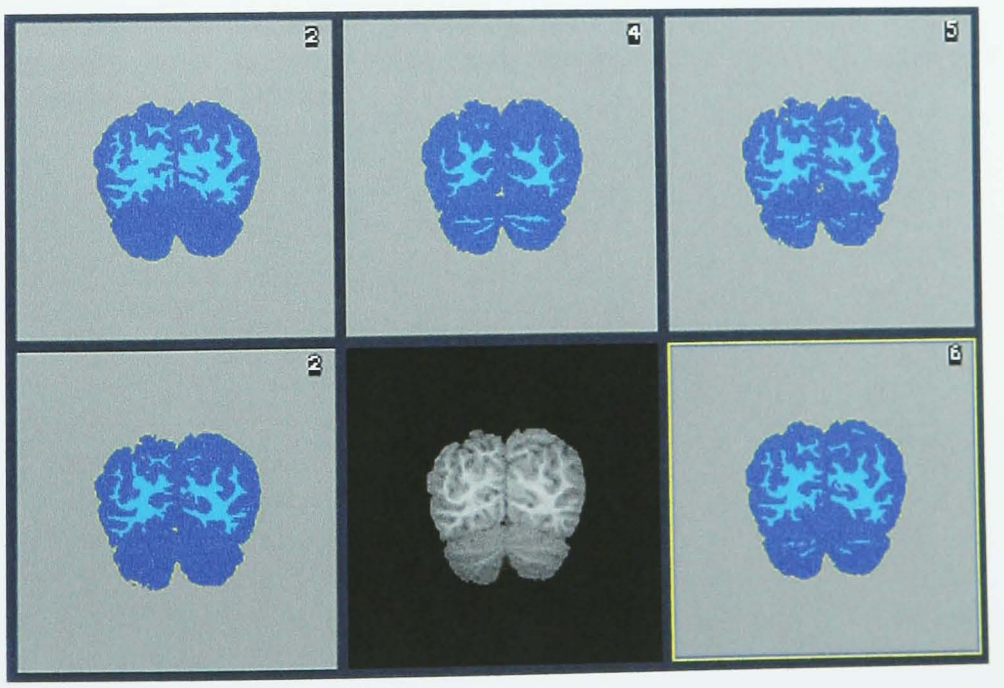


Fig. D.5. Screenshot from the IBSR MRI child volume ranking task. Source volume bottom centre, surrounded by the five segmentation volumes. Notice the missing white matter regions in the cerebellum in the two segmentations to the left, both based on N3 inhomogeneity correction. Absolute grades are shown in the top right corner of each viewport.

Fig. D.5 is a screenshot from one of the two IBSR MRI ranking tasks, showing a slice from the child volume and its five segmentations based on two template sets, EQ and N3 inhomogeneity correction and the gold standard manual ground truth. One of the three BrainWeb tasks is shown in fig. D.6. The multispectral segmentation is selected in this screenshot.

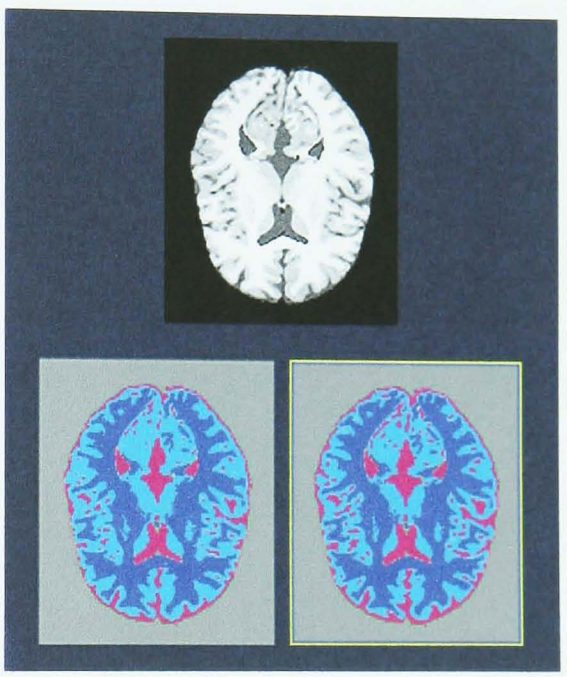


Fig. D.6. A screenshot from the BrainWeb ranking task, showing the volume with 3% noise and 40% inhomogeneity and its single (T1) and multi-channel (T1+T2) segmentation. The multispectral segmentation is selected.

Appendix E

The companion CD

E.1. Instructions for using the companion CD.

The CD included with this thesis contains source images and segmentations described in chapter 5, 6 and 7. It also contains two software interfaces for viewing the images. A Web browser based interface is used for viewing the natural colour images and cryo brain sections referred to in chapter 5 and 7. A Microsoft Windows based interface is used for viewing the cryo section and MRI volume data referred to in chapter 6 and 7.

The CD is in Joliet format, which is an extension of ISO9660 allowing for file names longer than 8 + 3 characters. It can be read on most IBM PC compatibles (running DOS, Windows or OS/2), Apple Macintosh/iMac and UNIX/Linux systems. However the long file names will only be accepted by Windows 95 or later, MacOS X (or older MacOS with the required file system extension) and later versions of Linux. On other platforms the file names will be truncated and the software interfaces will not work properly as a result. The Web browser based interface requires Netscape v. 4+ or Microsoft Internet Explorer v. 4+ with JavaScript enabled. Alternative browsers may work if they fully support frames, JavaScript and the document.images array. The interface for viewing volume data requires Microsoft Windows 95 or later. Both interfaces were designed for the following screen settings: 1024*768 spatial resolution, 24-bit colour resolution. It is highly recommended that these settings are used, particularly for the volume data interface.

Two ASCII text files are included on the root directory of the CD. The readme.txt file gives detailed information about additional compatibility issues and instructions about how to use the two software interfaces. The quickstart.txt file is an abbreviated version intended for the reader wishing to start using the software interfaces immediately. A hardcopy of readme.txt is included on the next pages.

Hardcopy of the file readme.txt on the companion CD.

ACSR Segmentation Viewers

Companion CD for the thesis "A Robust Framework for Medical Image Segmentation through Adaptable Class-Specific Representation".

This file gives detailed instructions about how to use the two software interfaces for viewing segmentations. Please read the file quickstart.txt to get started straight away.

Directories:

/html : contains a Web browser based interface for viewing natural colour image segmentation (chapter 5,7), natural colour image segmentation with 10% noise (chapter 5) and the brain cryo series segmentation (chapter 5,7)

/vtk : contains a VTK based application for viewing the cryo hip bone segmentations (chapter 7), BrainWeb single-channel and multispectral segmentations (chapter 6,7) and the IBSR segmentations (chapter 7)

The Web interface

Compatibility: Netscape v. 4+, Internet Explorer v. 4+
Note: Must have JavaScript enabled
Recommended screen setting: 1024*768, 24-bit colour

To launch the Web interface please open the file /html/menu.html in your browser. You will be presented with a menu from which you can select the set of segmentations you wish to view.

For the natural colour image segmentation and natural colour image segmentation with 10% noise, please select an image (top row of buttons) and a representation (left

column of buttons). The currently selected buttons will be displayed in green while other buttons remain red. The main area of the browser will display your selected combination and the average absolute ranking given to this particular combination in the human observer experiment described in chapter 7, section 7.2, will be displayed in the top left corner of the browser.

For the cryo brain series segmentation simply view the segmentations and scroll down if necessary. The image labels and the average absolute ranking given to the segmentations in the human observer experiment described in chapter 7, section 7.3 is displayed to the left and to the right of each row of images.

In all cases, to return to the menu screen, please click on the "Back to menu" button on the bottom left of your browser's display area.

The VTK interface

Compatibility: Windows 95/98/NT/2000 (not tested on ME/XP)
Minimum hardware spec.: 450 MHz CPU speed, 64MB memory (more recommended)
Required screen setting: 1024*768, 24-bit colour
Optional harddrive installation: 152MB free space required

The VTK interface is based on the Visualization Toolkit C++ libraries and requires the vtkdll.dll to run (included on the CD). Although the interface can be run from the CD, it is recommended that the full /vtk directory is copied to the harddrive and the interface run from there to ensure consistent performance.

Please open a DOS box and cd to the directory /vtk. Launch the interface by typing acsrview followed by the data sets you wish to view (see below).

THE DOS BOX SHOULD BE POSITIONED TO THE FAR RIGHT AND BOTTOM OF THE SCREEN _CLEAR_ OF THE TWO VTK WINDOWS AND _MUST_ BE THE CURRENTLY SELECTED WINDOW AT ALL TIMES TO ENABLE CONTROL OF THE APPLICATION.

The VTK windows will blank out if obscured by other windows. Should this happen, please move the other windows and press 'z' to refresh the VTK windows (this will only work after the volume data has initialised).

To launch the cryo hip bone data set viewer, type: acsrview cryo

To launch the BrainWeb date set viewer, type: acsrview brainweb

To launch the IBSR data set viewer, type: acsrview ibsr

The main VTK window displays the data sets. The small VTK window (top right) displays the label of the currently displayed data sets and a summary of available controls. When the small VTK window displays "Initialising..." please wait for volume data to load.

Controls are:

+/- : Browse forwards and backwards in the volumes

q : Automatically browse forwards and backwards in the volumes. Regardless of the currently displayed slice, auto-browse will start and end at the first slice in the volume and then reset the slice number to the one displayed before auto-browse

SPACE : In the IBSR and BrainWeb viewer use SPACE to skip to the next volumes

z : Refresh display of the two VTK windows (use if windows blank out)

ESC : Quit the viewer

The IBSR segmentation volumes are labelled with their segmentation algorithm. "GT" stands for Ground Truth. The IBSR and cryo hip bone segmentation volumes are labelled with the average absolute ranking given to them in the human observer experiments described in chapter 7, sections 7.5.2 and 7.4. The BrainWeb segmentation volumes are labelled with either "S" for single-channel or "M" for multispectral.

Please note that no such labelling was used in the experiments. The BrainWeb volumes will be displayed in random order similar to the way they were displayed during the ranking tasks. The IBSR volumes and cryo hip bone segmentations will be displayed in the same order every time the viewer is run unlike in the experiments.

Acknowledgements

The natural colour images are standard machine vision test images from Kodak, INRIA and the IEN (Istituto Elettrotecnico Nazionale) Computer Vision Research Group.

The RJMCMC segmentations are reproduced by permission from Zoltan Kato and were taken from the paper:

Z. Kato, "Bayesian Color Image Segmentation Using Reversible Jump Markov Chain Monte Carlo", technical report, 01/99-R055, European Research Consortium for Informatics and Mathematics (ERCIM), 1999

Images from the Visible Human Project appear courtesy of the United States National Library of Medicine.

The IBSR MR brain data sets 788_6_m and 1320_2_max and their manual segmentation were provided by the Center for Morphometric Analysis at Massachusetts General Hospital.

The file vtkdll.dll is copyright Kitware Incorporated and may not be included in any commercial applications without prior permission from Kitware.

Where other copyrights do not apply, the software interfaces on this CD are copyright Casper Nielsen 2002.