Retrospective techniques for segmentation of structural and functional MR brain images

Suresh Anand Sadananthan

2010

Suresh Anand Sadananthan. (2010). Retrospective techniques for segmentation of structural and functional MR brain images. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/42523>

<https://doi.org/10.32657/10356/42523>

RETROSPECTIVE TECHNIQUES FOR SEGMENTATION OF STRUCTURAL AND FUNCTIONAL MR BRAIN IMAGES

SURESH ANAND SADANANTHAN

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2010

Abstract

In this thesis, we focus on several segmentation problems arising in the area of structural and functional neuroimaging. Our solutions to these problems are based on a retrospective framework, wherein a complex segmentation procedure is divided into two simpler steps, initial segmentation and incorporation of prior information. While combining these steps might potentially lead to an optimal solution, we show that the simpler two-step approach can either be made equivalent to a combined procedure or achieve superior performance, due to simpler optimization.

We first consider the problem of detecting activated regions in functional Magnetic Resonance Imaging (fMRI). While activated regions are typically large, smaller spurious activations caused by noise are likely to appear in the segmentation. These can be removed by cluster size thresholding. Though cluster size thresholding can be regarded as a method to reduce false positives, it affects the smaller true activations. We show that in the context of Markov Random Field (MRF) based segmentation, simple removal of small regions after the segmentation is not optimal. We propose a retrospective correction approach that allows the regions to grow before they are eliminated based on the cluster size. This approach finds the best modification (removal or growth) and achieves superior performance to that of the standard MRF-based segmentation.

The second and the main problem considered in this thesis is skull stripping, i.e., the problem of separating the brain tissues (white matter, grey matter, cerebral spinal fluid) from the non-brain tissues (skull, scalp, eye sockets, neck tissues, etc.). Skull stripping performance suffers from the problem of narrow connections between brain and non-brain structures which usually results in preservation of significant amount of non-brain tissues. Many popular methods rely on iterative surface deformation to fit the brain boundary and tend to leave residual dura. We first approach the problem in a general framework of narrow connection removal. Here, we show that of all existing approaches, isoperimetric algorithm performs the best but can be sensitive to initialization. Instead, we propose a novel approach based on graph cuts that has beneficial features - global optimality, speed, and possibility of fully automated implementation for many applications. We also

show that incorporating intensity information into graph weight assignment can further improve performance. In the case of skull stripping, compared to the Hybrid Watershed Algorithm (HWA), our approach achieves an additional 10 – 30% of dura removal without incurring further brain tissue erosion. When used in conjunction with HWA, our approach substantially decreases and often fully avoids cortical surface overestimation in subsequent segmentation.

Lastly, we address the problem of skull stripping in multimodal images (combinations of T1-weighted and T2-weighted images). T1W and T2W images have different contrast properties and we show that the combined use of this information can help to further differentiate dura from brain structures, compared to using only T1W modality. We employ the graph cuts based skull stripping approach with suitably modified thresholding and post-processing procedures to suit multimodal images. The results obtained show significant decrease in the amount of dura in the resultant brain mask when using multimodal images.

Acknowledgements

First of all I would like to express my gratitude to my supervisor, Dr. Vitali Zagorodnov with whom I have been fortunate to work during the last four years. I am grateful to him for his excellent research directions, extended support and guidance without which my research would not have been interesting.

I would like to thank the School of Computer Engineering, Nanyang Technological University for providing me the financial support and the Biomedical Engineering Research Centre for providing computing facilities and a pleasant working atmosphere for research.

I also want to thank Dr. Michael W.L. Chee of Cognitive Neuroscience Laboratory for the research collaboration and providing valuable data for the experiments. Many thanks to Dr. Zheng Weili for her help and constructive suggestions during the project work.

I want to thank my family and friends for their continuous support and encouragement along this path.

Contents

Abstract	ii
Acknowledgements	iv
List of figures	ix
List of tables	xvi
1 Introduction	1
2 Background on MRI and image segmentation	6
2.1 Basics of MRI	7
2.1.1 MR image contrast	8
2.1.2 Functional magnetic resonance imaging	11
2.1.3 Multiecho imaging	12
2.1.4 Image quality issues	13
2.2 Background on segmentation	15
2.2.1 Thresholding	16
2.2.2 Region growing and region splitting-merging	19
2.2.3 Watershed segmentation	23
2.2.4 Active contour models	24
2.2.5 Graph theoretic approaches	28
2.2.6 Summary	33
2.3 Incorporation of prior information	34
2.3.1 Smoothing	34

2.3.2	Cluster size thresholding	35
2.3.3	Morphological processing	36
2.3.4	Watershed segmentation - preflooding	37
2.3.5	Graph cuts - isoperimetric ratio constraint	39
2.3.6	MRF based segmentation	40
2.4	Discussion	42
3	MRF segmentation with cluster size constraint	43
3.1	Review of existing techniques	44
3.1.1	Thresholding of pixel intensities	44
3.1.2	Controlling family-wise error and false discovery rate	45
3.1.3	Cluster based thresholding in fMRI	47
3.1.4	MRF-based detection of activations in fMRI	48
3.2	Motivation	49
3.3	Incorporating cluster size constraint	50
3.4	Problem statement	52
3.5	Retrospective cluster size thresholding	53
3.5.1	Finding a growth modification	53
3.5.2	Finding the smallest cost growth modification	55
3.6	Experimental results	56
3.7	Discussion	60
4	Removal of narrow connections: Critical analysis of existing approaches	61
4.1	Narrow connection removal: Problem definition	61
4.2	Review of existing methods	63
4.2.1	Morphological processing	63
4.2.2	Distance transform followed by watershed algorithm	64
4.2.3	Isoperimetric segmentation	65
4.3	Performance comparison of existing methods	67
4.3.1	Morphological processing	69
4.3.2	Distance transform followed by watershed transform	70
4.3.3	Isoperimetric segmentation	71

4.4	Summary	77
5	Distance transform based edge weight assignment for graph theoretic removal of narrow connections	79
5.1	Generalized ISO	80
5.1.1	Polynomial case	80
5.1.2	Polynomial weight assignment	81
5.2	Removing narrow connections using graph cuts	85
5.2.1	Performance evaluation	88
5.3	Discussion	90
5.4	Automated seed selection for GCUT	93
5.4.1	Seed selection	94
5.5	Discussion	97
6	Skull stripping	98
6.1	Review of existing skull stripping techniques	100
6.1.1	Region-based approaches	100
6.1.2	Boundary-based approaches	104
6.1.3	Hybrid approaches	109
6.2	Motivation	110
6.3	Proposed approach	111
6.3.1	Obtaining preliminary mask	112
6.3.2	Removal of narrow connections	113
6.3.3	Postprocessing	114
6.4	Data sets	115
6.4.1	Ground truth	116
6.4.2	Image quality	117
6.5	Evaluation metrics	118
6.6	Quantitative performance evaluation	120
6.6.1	Comparison with existing skull stripping approaches	120
6.6.2	Effect on FreeSurfer segmentation pipeline performance	124
6.6.3	Robustness and sensitivity to algorithm's parameters	126

6.7	Discussion	130
6.8	Conclusion	133
7	Skull stripping of multimodal MR images	135
7.1	Previous work	136
7.2	Multimodal brain segmentation	137
7.2.1	Brain tissue segmentation using multiecho MPRAGE	140
7.3	Multimodal dura segmentation	142
7.3.1	Skull stripping of multiecho MPRAGE	146
7.4	Main issues in multimodal skull stripping	149
7.5	Proposed approach	152
7.5.1	Obtaining preliminary mask	153
7.5.2	Removal of narrow connections	155
7.5.3	Post-processing	156
7.6	Data set	157
7.7	Performance evaluation	157
7.7.1	Quantitative performance evaluation	158
7.7.2	Qualitative performance evaluation	158
7.8	Discussion	159
8	Conclusions and future directions	161
A	Performance evaluation metrics	163
B	Lemmas	166
	Bibliography	178

List of Figures

2.1	(a) Structural and (b) functional MR images of a brain	7
2.2	MR imaging process	8
2.3	Relationship between TR and T1 contrast	9
2.4	Relationship between TE and T2 contrast	9
2.5	Courtesy of [1] (a) T1-weighted image and (b) T2-weighted image (c) PD-weighted image	10
2.6	Courtesy of www.fmrib.ox.ac.uk. Activated regions from fMRI experiment .	11
2.7	Multiecho MPRAGE (a) first echo (b) second echo (c) third echo (d) last echo	12
2.8	(a) Model image (b) model image corrupted with i.i.d. Gaussian noise, $SNR = 10dB$ (c) model image corrupted with i.i.d. Gaussian noise, $SNR = 4dB$ (d) model image corrupted by an intensity gradient in the diagonal direction	16
2.9	Type I and type II errors	18
2.10	Optimal hypothesis testing applied to (a) model image ($10dB$) (b) model image ($4dB$) (c) model image (intensity gradient)	18
2.11	Otsu's thresholding of (a) model image ($10dB$) (b) model image ($4dB$) (c) model image (intensity gradient)	19
2.12	Region growing using difference between the intensity of neighboring pixels as the criterion: First row - model image ($10dB$), second row - model image ($4dB$), third row - model image (intensity gradient)	21

2.13	Region growing using difference between pixel intensity and average region intensity as the criterion. (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)	21
2.14	Region splitting and merging: First row - model image (10dB), second row - model image (4dB), third row - model image (intensity gradient)	22
2.15	Watershed segmentation of (a) model image (10dB) (b) second row - model image (4dB)	23
2.16	Snakes: First row - close initialization on model image with intensity gradient and its corresponding output, second row - distant initialization and output on model image with intensity gradient, third row - initialization and output on model image corrupted with noise (smoothed)	25
2.17	Level sets on (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)	27
2.18	Graphical framework for a 5 x 5 image (a) region-based approach (b) boundary-based approach	29
2.19	Boundary based graph cuts (a) foreground and background seeds (b) model image (10dB) (c) model image (4dB) (d) model image (intensity gradient) .	32
2.20	Gaussian smoothing and Otsu's thresholding: First row - model image (10dB) processed with (a) 3 x 3 filter (b) 7 x 7 filter, second row - model image (4dB) processed with (c) 3 x 3 filter (d) 7 x 7 filter	35
2.21	Cluster size thresholding applied to (a) Figure 2.11(a) (b) Figure 2.11(b) . .	36
2.22	Morphological opening operation applied to the Gaussian smoothed image in Figure 2.20 (b) using disk structuring element	37
2.23	Illustration of preflooding - merging procedure	38
2.24	Watershed transform with preflooding applied to (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)	39
2.25	Isoperimetric segmentation applied to (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)	40
2.26	MRF based segmentation: First row - model image (10dB), second row - model image (4dB), third row - model image (intensity gradient)	41

3.1	Controlling type I error at the expense of type II error	45
3.2	(a) Model image (b) model image corrupted by noise (c) thresholding (d) MRF segmentation (e) cluster size thresholding of (c) (cluster size > 10) (f) cluster size thresholding of (d) (cluster size > 10)	51
3.3	(a) Partition π (has the smallest cost) (b) Partition π' (non-optimal)	52
3.4	A small region and boundary edge	54
3.5	Possible and impossible changes to the region's shape after modifying boundary edge weight (a) Original region and boundary edge e (b) Region Δ doesn't contain e (impossible) (c) Region Δ consists of two disconnected parts (impossible) (d) Correct modification	55
3.6	(a) Model image, (b) model image corrupted with noise, (c) MRF segmentation of (b), (d) MRF segmentation with retrospective cluster size thresholding (cluster size > 7) - grown region highlighted in yellow	57
3.7	Performance comparison of the segmentation techniques	58
3.8	(a) Statistical parametric map (SPM), (b) corresponding structural MR image slice, (c) thresholding of the SPM, (d) cluster size thresholding of (c) (cluster size > 10), (e) thresholding result (d) overlaid on (b), (f) MRF segmentation of the SPM, (g) cluster size thresholding of (f) (cluster size > 10), (h) MRF segmentation with retrospective cluster size thresholding (cluster size > 10) - one small region grown beyond cluster size highlighted in red, (i) result of retrospective correction approach (h) overlaid on (b) . .	59
4.1	Examples of narrow connections in medical images: row 1-abdominal fat segmentation, row 2-brain segmentation, row 3-liver segmentation	62
4.2	Morphological processing (a) sample image (b) erosion of (a) by disk structuring element, size 25 (c) Selection of the largest connected component (d) dilation of (c) by disk structuring element, size 25	64
4.3	Distance transform followed by watershed algorithm (a) sample image (b) distance transform of (a) (c) distance transform of (a) - inverted (d) plot of (c) (e) watershed transform of (c)	65
4.4	Cutting narrow connections with ISO	66

4.5	Model images (a) Arbitrarily shaped regions (b) Two circular regions (c) Circle-rectangle	68
4.6	MORPH using disk structuring element of different sizes (a) sample image (b) size 25 - little smoothening (c) size 50 - original shape damaged	69
4.7	DWAT - oversegmentation (a) original image (b) distance transform of (a) (c) distance transform of (a) - inverted (d) watershed transform of (c) . . .	71
4.8	Possible cuts for two-circles object	73
4.9	Possible cuts for circle-rectangle object	74
4.10	Effect of position of seed point on ISO	77
5.1	Selecting background seeds in GCUT manually provides good segmentation	86
5.2	Cutting narrow connections with GCUT	86
5.3	GCUT on the two-circles image (a) model image with $w < 2.5\sqrt{r_2}$ (b) GCUT on (a), (c) model image with $w > 2.5\sqrt{r_2}$ (d) GCUT fails to remove the narrow connection in (c)	91
5.4	GCUT on the circle-rectangle image (a) model image with $w < 2b$ (b) GCUT on (a), (c) model image with $w > 2b$ (d) GCUT on (c)	92
5.5	(a) Original MR abdominal image slice (b) SAT (red) and VAT (green) highlighted	93
5.6	(a) Original MR brain slice (b) image after skull stripping	94
5.7	Fat segmentation (a) original abdominal MR image slice (b) preliminary mask (c) foreground seed (d) segmented output - SAT (e) segmented output - VAT	95
5.8	Selection of foreground seed is performed by finding the brightest and the most uniform cube (b), followed by conservative region growing (c)	96
6.1	MR brain image (a) tissue labels (b) pial surface highlighted in red	99
6.2	Typical results of existing skull stripping techniques (a) original image, (b) ground truth mask, (c) BSE, (d) WAT, (e) BET, (f) HWA	101
6.3	Watershed transform - basic idea (a) conceptual representation of brain image (b) interpretation of intensity as height information (c) inverted intensity interpretation	103

6.4	Impact of change in the value of parameters on skull stripping. Left: Original MR brain slices (b) BSE with parameters (3,25,0.8,2) (c) BSE with parameters (3,25,0.75,2) (e) BET with default parameter value = 0.5 (f) BET with parameter value = 0.6 (h) WAT with default preflooding height = 0.145 (i) WAT with preflooding height = 0.15	105
6.5	Problem in boundary determination in BET: A single dark voxel (pointed by arrow) can alter the intensity threshold and therefore the shape of the surface of the deformable model	108
6.6	Intensity overlap in the boundary region of the brain	111
6.7	The pipeline of proposed skull stripping approach	112
6.8	Effect of different threshold values on the quality of initial mask. Too low threshold (second column) leads to insufficient separation between brain and non-brain structures, too high threshold (right column) results in brain loss	113
6.9	Original image (left column), graph cuts with weight assignment based on distance transform (middle), graph cuts with weight assignment based on distance transform and intensity (right). First row is an image from data set 3, second row - from data set 1	115
6.10	Example of ground truth masks, top row - data sets 1 and 2 (IBSR), cerebellum included, bottom row - data sets 3 and 4 (Siemens Allegra 3T), cerebellum excluded	116
6.11	Skull stripping result 1, FP=27%, FP_adj=8% (left), skull stripping result 2, FP=31%, FP_adj=6% (right)	119
6.12	Problematic FreeSurfer segmentation performance using HWA brain mask (left) is improved using GCUT_HWA (right), see pial surface (highlighted in red) overgrown on the left	127
6.13	ROC curves (false positive rate vs. true positive rate) for $k = 2.3$ (Table 6.8) for (a) data set 1 (b) data set 2 (c) data set 3 (d) data set 4	129
6.14	ROC curves (false positive rate vs. true positive rate) for $T = 0.36I_{WM}$ (Table 6.9) for (a) data set 1 (b) data set 2 (c) data set 3 (d) data set 4	130

6.15	Typical errors exhibited by HWA (middle) and GCUT (right). HWA is often confused by double boundary between scalp/dura/GM, resulting in inclusion of large chunks of skull/dura mater, rows 1-2. GCUT fails to cut connections where there is no noticeable intensity separation between dura and GM, row 3	132
7.1	(a) T1-weighted image and (b) T2-weighted image	136
7.2	Intensity distributions of T1W and T2W images in Figure 7.1 and their combination	138
7.3	Histograms of the (a) T1W intensity distribution (b) T2W intensity distribution	139
7.4	(a) Thresholding based only on T1W image intensity information (b) Thresholding based on T1W and T2W image intensity information	139
7.5	Multiecho MPRAGE (a) first echo (b) last echo	141
7.6	Distribution of intensities of first and last echo of MEMPRAGE	141
7.7	Intensity distribution of T1W and T2W images in Figure 7.1 (a) plot of WM, GM, and dura-CSF (b) plot of only dura-CSF class	143
7.8	Typical results of dura-CSF overlay on the original T1W image (a) Original image (b) dura overlay (c) CSF overlay	144
7.9	Classification of CSF voxels in the intensity distribution of Figure 7.7 (b)	145
7.10	Classification of dura voxels in the intensity distribution of Figure 7.7 (b)	146
7.11	(a) Segmentation based only on T1W image intensity information (b) segmentation based on T1W and T2W image intensity information	147
7.12	Intensity distributions of first and last echoes in Figure 7.5 (a) plot of WM, GM, and dura-CSF (b) plot of only dura-CSF voxels	148
7.13	Typical results of dura-CSF overlay on the first echo of MEMPRAGE (a) original image (b) dura overlay (c) CSF overlay	149
7.14	Thresholding based on ratio of first and last echo intensity information (a) only T1W threshold (b) van der Kouwe threshold (c) new ratio threshold	150

7.15	Typical results of dura elimination in multiecho and multimodal images (a) original image slices (b) dura layer classified from MEMPRAGE overlaid on original image (c) dura layer classified from multimodal skull stripping overlaid on original image	151
7.16	Pipeline of proposed multimodal skull stripping approach	153
7.17	Intensity distribution and decision boundaries used for thresholding	154
7.18	ROC curve for $T = 0.36I_{WM1}$ and different values of d_{WM}	155
7.19	Thresholding using multimodal (T1W and T2W) information (bottom row) results in additional dura elimination and narrower connections compared to thresholding using only monomodal (T1W) information (middle row) . .	156
7.20	Inaccuracy in the ground truth (a) original image slices (b) ground truth overlaid on the original image	159
7.21	Additional dura elimination in multimodal skull stripping (bottom row) compared to monomodal skull stripping (middle row)	160
B.1	Minimum isoperimetric ratio cut - illustration	167
B.2	Plot of the function in (B.1)	167
B.3	Minimum isoperimetric ratio cut (cut within larger region) - illustration . .	168
B.4	Plot of the function in (B.3)	170
B.5	Plot of the function in (B.4)	171
B.6	Minimum isoperimetric ratio cut (cut within smaller region) - illustration .	171
B.7	Plot of the function in (B.6) for $p \leq 2$	173
B.8	Plot of the function in (B.8)	175
B.9	Plot of the function in (B.10)	176

List of Tables

2.1	Classification of pixels	17
2.2	Performance comparison of segmentation techniques	33
3.1	Outcomes of hypothesis testing	45
4.1	Performance comparison: Maximum width of the connection that can be cut by different methods	78
5.1	Performance comparison: Maximum width of the connection that can be cut by different methods	85
5.2	Performance comparison: Maximum width of the connection that can be cut by different methods	90
6.1	Estimated CNR between GM and dura/CSF, and coefficient of variation within WM for tested data sets	117
6.2	Comparison of graph cuts skull stripping approach (GCUT) with existing skull stripping approaches, Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), and Hybrid Watershed Algorithm (HWA), using data set 1 (18 1.5mm scans, IBSR). GCUT_HWA stands for mask obtained by intersecting GCUT and HWA mask.	121
6.3	Comparison of GCUT with existing skull stripping approaches using data set 2 (20 normal subjects, IBSR)	122
6.4	Comparison of GCUT with existing skull stripping approaches using data set 3 (Siemens Allegra 3T scanner, good quality)	123

6.5	Comparison of GCUT with existing skull stripping approaches using data set 4 (Siemens Allegra 3T scanner, poor quality)	124
6.6	Effect of brain masks on subsequent estimation of pial surface position in 15 hemispheres with prior overestimation problem	125
6.7	Effect of brain masks on subsequent estimation of pial surface position in 15 hemispheres without prior overestimation problem	126
6.8	Sensitivity of GCUT performance to intensity threshold parameter for $k = 2.3$	128
6.9	Sensitivity of GCUT performance to parameter k controlling the influence of voxel intensity on cut positions for $T = 0.36I_{WM}$	128
7.1	Confusion matrices for unimodal-only T1W (left) and multimodal-T1W and T2W (right) brain segmentation	138
7.2	Multimodal brain segmentation misclassification rate	140
7.3	Multimodal skull stripping misclassification rate for CSF vs GM+WM separation on the intensity distribution in Figure 7.9	144
7.4	Multimodal skull stripping misclassification rate for the thresholding on the intensity distribution in Figure 7.11	145
7.5	Skull stripping of MEMPRAGE: Misclassification rate	148
7.6	Comparison of monomodal and multimodal skull stripping approaches . . .	157
A.1	All possible outcomes of a hypothesis test	163

Chapter 1

Introduction

Research in neuroimaging has seen significant progress in the recent years, thanks to the advancement of non-invasive and high resolution techniques for visualizing various anatomical structures of the brain. Neuroimaging can be classified into two main categories - structural imaging and functional imaging broadly referred to as Magnetic Resonance Imaging (MRI) and functional Magnetic Resonance Imaging (fMRI) respectively. Structural imaging provides volumetric views of the brain. It helps in visualization of the brain structure and diagnosis of brain tumors, strokes and certain disorders such as multiple sclerosis, Alzheimer's disease, etc. Functional imaging is used to measure the metabolic changes taking place in the brain as the person performs various cognitive tasks. It visualizes the task-relevant brain areas using the BOLD (Blood Oxygenation Level Dependent) response which is based on the fact that neural activity is generally accompanied by an increase in blood flow in the activated brain areas. The applications of fMRI include diagnosis of metabolic diseases, mapping brain activity and building brain computer interfaces. It is an important research tool in cognitive neuroscience and neuropsychology. MRI and fMRI are often seen as complementary to each other - MRI helps in visualization of the brain structure and fMRI helps in mapping the structure with specific mental function.

With MRI and fMRI playing a vital role in the analysis of the structure and functioning of the brain, there is substantial on-going work dealing with extraction of features of these images for analysis. This is typically accomplished by segmentation, an image processing operation that facilitates the delineation of anatomical structures such as the white matter

(WM), gray matter (GM), cerebral spinal fluid (CSF), etc. in the case of structural MR images and detection of activated regions, brain mapping etc. in the case of functional MR images. Segmentation is also an important tool for applications such as image-guided surgery, pathology detection, and volume rendering.

In this thesis, we focus on the problem of segmentation in structural and functional MR images. In particular, we consider two segmentation problems: detection of activated regions in fMRI and skull stripping in structural MRI. We propose novel retrospective segmentation techniques based on graph cuts to solve these problems. The word ‘retrospective’ has several meanings, but the one closest to ours is ‘retrospective correction’, as in ‘retrospective non-uniformity correction’, which generally means a refinement of the corrupted data. In the cases when the term is used, the correction can be done retrospectively, after corruption has occurred, or prospectively, to prevent the corruption in the first place. By choosing retrospective correction, we effectively decide not to prevent the corruption. In all our approaches, the constraints on the size or shape of the objects can theoretically be included in the main segmentation step, but that would greatly complicate the optimization procedure. Hence we choose to introduce the constraints retrospectively, and demonstrate that in some cases such approach can produce results equivalent to or better than single step segmentation approaches.

It is well known that some of the best and most beautiful engineering work is never implemented in practice due to simplified assumptions or lack of robustness to real-world imperfections. For example, many image segmentation algorithms developed for medical applications never make it to clinical practice. To avoid falling into the same trap, in this thesis we tried to give equal priority to theory development and clinical significance. In particular, Chapters 1-5 constitute the “engineering” part of the thesis and describe our theoretical work on segmentation of fMRI (Chapter 3) and removal of narrow connections (Chapters 4-5). Chapters 6-7 deal primarily with clinical applications and include several large scale performance evaluation studies conducted on real brain images. We believe it is these two last chapters that provide real significance to the theory developed in the earlier chapters. Our contributions in this thesis include

- **Detection of activations in fMRI.** A novel retrospective correction approach

that combines the benefits of MRF-based segmentation and cluster size thresholding for improved detection of activations in fMRI.

- **Removal of narrow connections.** Generalized isoperimetric algorithm and graph cuts based segmentation method with distance transform based edge weight assignment for the removal of narrow connections.
- **Skull stripping of T1W MR images.** Graph cuts based segmentation algorithm for skull stripping. The proposed approach reduces dura attachments by 10-30% compared to the hybrid watershed algorithm (HWA). The intersection of resultant brain mask with HWA mask solves cortical thickness overestimation problem. Our skull stripping algorithm is incorporated in FreeSurfer 5.0.0 under the name `mri_gcut`.
- **Skull stripping of multimodal images.** A novel algorithm that combines the benefits of T1W and T2W image information to improve skull stripping performance. The use of multimodal images results in reduced dura attachments compared to only T1W images.

Segmentation in structural and functional images is considered a difficult task because of the sheer volume of the data set and the variation in image modalities. The problem is made even more complicated due to the convoluted nature of the surface of the brain and substantial amount of noise and artifacts present in the images. Existing segmentation techniques such as intensity thresholding, region-based methods, and boundary-based methods do not always produce satisfactory results due to the presence of these complexities. The main way to improve the performance is by incorporation of prior information into the segmentation process.

Prior information in the segmentation problems considered can come in different forms, i.e., the minimum cluster size in fMRI and the smoothness of the brain mask in skull stripping. However, when this information is incorporated into the core of segmentation algorithm, it often leads to over-complicated optimization that is not guaranteed to converge to a global optimum. One possible solution is to incorporate prior information as a separate retrospective procedure that follows the main algorithm. This makes the prob-

lem simpler, but doing so no longer guarantees optimality. However, as we intend to show in this thesis, retrospective correction can provide optimal solution for many applications and even when not optimal, it possesses many useful properties that can be very beneficial for the overall algorithm performance. Before we introduce our problems, we give a background of MRI and some of the segmentation techniques that are employed in structural and functional image processing in Chapter 2.

It is known that the mental activities are linked with the brain cortical areas, also called the functional regions. These links can be established through functional Magnetic Resonance Imaging (fMRI) experiments that use statistical processing of MR image sequences to visualize the activated areas. Noise, head motion, imaging artifacts can all lead to false activations. Usually, improved detection while maintaining sufficiently low rate of false positives can be achieved by cluster size thresholding where small regions are removed in the post-processing step. Such small regions are likely to be caused by noise and are hence better removed. Cluster size thresholding is not optimal and Markov Random Field (MRF) based segmentation provides a good alternative because it decreases the occurrence of small spurious regions. However, MRF-based segmentation is local by nature and does not contain any notion of cluster size. Our contribution in Chapter 3 is a novel retrospective correction approach that combines the benefits of MRF-based segmentation and cluster size thresholding for improved detection of activations in fMRI.

Image segmentation deals with the problem of identifying and isolating the individual regions from the whole image. In some cases such as brain image segmentation, spurious connections can be formed between distinct regions, leading to their erroneous treatment as a single region. These connections are usually narrow, and the problem associated with their removal came to be known as “narrow connection removal”. It is often solved retrospectively, where the first pass of the segmentation algorithm identifies all regions without attempting to separate them, while the goal of the second pass is to remove the narrow connections. The problem is made more complicated when there is no clear demarcation between the regions. Chapters 4 and 5 address the theoretical issues in the removal of narrow connections between regions. In Chapter 4, we review several existing solutions, such as morphological processing, distance transform followed by watershed algorithm and isoperimetric ratio algorithm. We perform critical analysis of each of these

methods and demonstrate their performance theoretically.

In Chapter 5, we propose generalized isoperimetric algorithm and graph cuts based method for narrow connection removal to overcome the drawbacks of the existing methods. We show how the graph cuts segmentation technique with a suitably modified edge weight assignment can be used to remove narrow connections and separate the regions. Contrary to the common belief that graph cuts requires manual intervention for seed selection, the proposed graph cuts algorithm can be made fully automated.

In Chapter 6, we consider our main application - the problem of segmenting the brain (white matter (WM), gray matter (GM), cerebral spinal fluid (CSF)) from the non-brain tissues (skull, scalp, eye sockets, neck tissues, etc.). This technique, often termed as skull stripping, represents one of the essential steps in brain image analysis. Operations such as brain tissue classification, registration, volumetric analysis of the brain, and cortical surface reconstruction require the brain to be skull stripped before they are applied. Our main contribution here is a new algorithm based on graph theoretic segmentation for the skull stripping of MR brain images. We use intensity thresholding followed by removal of narrow connections using graph cuts and post-processing to obtain the brain mask. The proposed skull stripping approach reduces dura attachments by 10 – 30% compared to the current state of the art algorithm, hybrid watershed algorithm (HWA). Further, the resultant brain mask of the proposed approach intersected with the HWA mask eliminates the cortical thickness overestimation problem.

Skull stripping usually has been restricted to only T1W images because of the difficulty in acquiring T2W images, which are properly aligned with their T1W counterparts. But this has been made possible with the recent advances in MR image acquisition and registration thus creating an opportunity to exploit the T2W image information for improved skull stripping. T1W and T2W images have different image contrasts and in Chapter 7 we show how the combined use of this information can help remove additional dura that could not be distinguished from the brain using only T1W information. We employ graph cuts based skull stripping approach, similar to the one described in Chapter 6, with modified thresholding and post-processing procedures to suit multimodal images. The results obtained show significant decrease in the amount of dura in the resultant brain mask when using multimodal images.

Chapter 2

Background on MRI and image segmentation

Magnetic Resonance Imaging (MRI) is a technique that allows non-invasive high quality imaging of the human body. It uses strong magnetic fields and radio frequency waves to produce high quality two or three-dimensional images of different biological tissues. The ability of MRI to obtain images in multiple planes and to characterize and differentiate tissues using their physical and biochemical properties increases its utility for diagnostic purposes [2–8]. In this thesis, our focus on MRI is confined to neuroimaging, i.e., imaging of the brain. It can be broadly classified into structural imaging commonly referred to as MRI and functional magnetic resonance imaging (fMRI). Structural imaging gives volumetric views of the brain and functional imaging is used to measure the metabolic changes occurring in the active parts of the brain. Figure 2.1 provides examples of typical structural and functional MR images.

Analysis of MR images requires extraction of features from these images and this is accomplished by an image processing procedure called segmentation. Segmentation delineates anatomical structures such as the white matter (WM), gray matter (GM), cerebral spinal fluid (CSF), etc. from the structural MR brain images. Segmentation is also an important tool for image-guided surgery, pathology detection, and volume rendering. In order to design an automated algorithm for the segmentation of MR images, knowledge of different types of MRI sequences is essential. Furthermore, information on the source of

Figure 2.1: (a) Structural and (b) functional MR images of a brain

noise and artifacts and their characteristics enables making the algorithm less dependent on the quality of the image. In this chapter, we describe the basics of MRI, different types of acquisitions and typical properties of noise and artifacts in MR images. We also provide background on existing segmentation techniques for MR images.

2.1 Basics of MRI

MRI system is constructed of at least three basic subsystems: a main magnet to produce a strong, homogenous, static field denoted as the B_0 field; a subsystem for generation of a gradient magnetic field for signal localization; and a radio-frequency (RF) subsystem, for generation and transmission of a rotating magnetic field, denoted as the B_1 field, and measurement of Nuclear Magnetic Resonance (NMR) signals. Magnetic resonance imaging uses the signal from the nuclei of hydrogen atoms for image acquisition. When the hydrogen nuclei are exposed to an external magnetic field, the magnetic moments gradually align with the direction of the field, like a compass needle, and undergo precession at a characteristic frequency, called Larmor frequency which is proportional to the strength of the applied magnetic field. When the nuclei are in the stable spin system, electromagnetic wave (RF) is applied to the object and the energy is absorbed. This process, known as excitation, induces an alternating voltage in the receiver coil. When the field is turned off, the atomic nuclei that had absorbed electromagnetic energy in a specific frequency, release

that energy [7, 8]. Sensors read these emissions and use them to generate the MR image. MR signal then rapidly fades due to spin-lattice interaction and spin-spin interaction. This process is called relaxation. Finally, the system returns to the stable state present before excitation. The overall process is shown in Figure 2.2. Using MRI, it is possible to create images of both surface and subsurface structures, with a high degree of anatomical detail.

Figure 2.2: MR imaging process

2.1.1 MR image contrast

The difference in brightness that helps to distinguish different structures in the image is called the image contrast, and MRI is versatile in generating images of a wide range of different tissues. The acquisition sequences in MRI can be adjusted to maximize the image contrast to increase visibility of different tissue types. MR image contrast is based either on proton density or on the relaxation times of each of the tissues, which is the time taken by the excited nuclei to return to their stable state. Depending on the inherent properties of the tissues in dissipating the energy, there are different types of MR scans, such as T1-weighted (T1W), T2-weighted (T2W), T2*-weighted, and proton density (PD) weighted images. For example, WM appears bright and CSF appears dark in T1W images and dark WM and bright CSF characterize T2W images, see Figure 2.5.

T1 contrast

T1 relaxation time, which is also called the longitudinal relaxation time or spin-lattice relaxation time, determines the time taken by the excited spins to get back to equilibrium and be available for the next excitation. If the contrast of the image is determined by T1 relaxation time, it is called T1-weighted image or T1W image. The time between the successive excitations is called the repetition time (TR) and this controls the T1-weighting of the image. If a short TR is selected, tissues with short T1 return to equilibrium rapidly and regain most of the longitudinal magnetization. Thus, they emit stronger signals and

Figure 2.3: Relationship between TR and T1 contrast

appear bright on the image. Tissues with a long T1 (for eg. CSF) take longer time to relax and leave less magnetization available before the next excitation. Therefore, they emit only weaker signals resulting in dark appearance. If long TR is selected, all tissues have longer time to return to equilibrium and hence have low T1-weighting [6, 8–10]. Figure 2.3 shows the relationship between T1 contrast and TR.

Figure 2.4: Relationship between TE and T2 contrast

T2 contrast

T2 relaxation time, which is also called the transverse relaxation time or spin-spin relaxation time, is the time taken for the transverse magnetization to decay or the time till the phase coherence is lost. T2-weighting is characterized by the echo time (TE) which is the time between the initial RF pulse and the detection of the signal. If a short TE is used,

then the decay is low and therefore amounts to low T2-weighting. If a long TE is used, tissues with short T2 lose most of their signal and appear dark whereas tissues with long T2 produce stronger signals and therefore appear bright. Thus TE can be used to control T2-weighting [6,8–10]. Figure 2.4 shows the relationship between TE and T2 contrast.

Figure 2.5: Courtesy of [1] (a) T1-weighted image and (b) T2-weighted image (c) PD-weighted image

T2* contrast

T2*-weighted scans use a gradient echo sequence with long TE. They have the combined effect of spin-spin interaction (T2) and magnetic field inhomogeneities resulting in the decay of transverse magnetization. T2* decay is faster than T2 decay and their relationship is given by $\frac{1}{T2^*} = \frac{1}{T2} + \frac{1}{T2'}$, where T2' denotes the effect of field inhomogeneity. T2*-weighted images are sensitive to the amount of deoxygenated haemoglobin and therefore they are the basis for the BOLD contrast used in fMRI [6,8,9,11].

Proton density contrast

Proton density (PD) contrast is based on the net magnetization of each voxel which depends on the number of protons in that voxel. The proton density contrast can be maximized by minimizing T1 and T2 contrasts and therefore is characterized by long TR and short TE. Refer to Figure 2.5 for a sample PD-weighted image. The image has the highest signal in CSF and ventricles with less signal in the gray matter and white matter.

This difference in the intensity values can be used to improve labeling of the different tissues [8, 11].

2.1.2 Functional magnetic resonance imaging

Functional magnetic resonance imaging captures neuronal activity of the brain using T2*-weighted imaging that effectively records the blood oxygenation level-dependent (BOLD) response, while the subject performs behavioral or mental tasks. The BOLD response is based on the magnetic susceptibility of haemoglobin and this is identified by the T2*-weighted pulse sequences. The images are obtained using echo planar imaging (EPI) where multiple gradient echoes are obtained per excitation and all these echoes are used to generate the image. Since the acquisition time is very short, the image generated is generally of low resolution. The resolution and the signal to noise ratio can be improved by using multishot echo planar techniques, but these are more sensitive to motion due to increase in the acquisition time [6, 11].

Figure 2.6: Courtesy of www.fmrib.ox.ac.uk. Activated regions from fMRI experiment

The fMRI data show increased intensity in those parts of the brain that are activated by stimulation (Figure 2.6). Statistical analysis of these activations is performed to determine which voxels are activated by the stimulation. It relates the signal changes with the timing of the subject's behavior. The resulting output is called the statistical parametric map (SPM) and it indicates the likelihood that a particular brain region is involved in processing that particular task or behavior [9]. Further image processing tasks and analysis are performed on this statistical parametric map of the brain.

2.1.3 Multiecho imaging

The multiecho imaging decreases the acquisition time in single images and generates images with different tissue contrasts. There are two classes of multiecho techniques, namely the echo planar technique and the fast spin echo technique. Echo planar imaging involves acquisition of multiple gradient echoes per excitation pulse whereas fast spin echo involves acquisition of multiple spin echoes per excitation pulse.

Multiecho imaging has several advantages. They produce less distortions since the alternating gradient echoes are acquired with opposite readout directions [12]. They provide extra information that can be used for further processing such as the $T2^*$ map which can be used to improve dura elimination [13]. The $B0$ offset can be calculated when the phase information is available. Further, these properties of multiecho imaging can be used for sequence-independent segmentation [14].

Figure 2.7: Multiecho MPRAGE (a) first echo (b) second echo (c) third echo (d) last echo

One example of a multiecho sequence, which will be used for experiments in Chapter 7, is the multiecho magnetization prepared rapid gradient echo (MEMPRAGE) acquisition sequence developed with an intention to have more contrast between the tissues of interest and hence better separation between the classes. MEMPRAGE has higher bandwidth and lesser distortion and yet good signal to noise ratio (SNR) [13]. The $T2^*$ information in the multiple echoes of the MEMPRAGE helps to segment tissues such as dura which are very difficult to separate with only T1W image information. The MEMPRAGE data consist of four echoes and each progressive echo contains a larger weighting of $T2^*$ contrast (see Figure 2.7).

2.1.4 Image quality issues

The quality of MR images depends on the noise and artifacts present in the image. Such artifacts differ from each other depending on their source and can be patient related, image processing related, radio frequency related, artifacts due to external magnetic field, magnetic susceptibility, and gradient related artifacts.

Patient related artifacts consist of the motion artifacts and the magic angle artifacts. The motion artifacts are due to the voluntary and involuntary motions of the patient which result in ghosting effects in the image. Magic angle artifact occurs due to the change in the T2 contrast when fibrillary tissues (ligaments and nerves) are at 55° (magic angle) to the main magnetic field. However, the T1 contrast is unaffected in this case.

Image processing artifacts arise during the processing of the signal / data. The wraparound / aliasing, chemical shift, truncation, and partial volume artifacts belong to this category. Wraparound artifact is the representation (wrap around) of the image data outside the field of view (FOV) on the opposite side of the image because the FOV is usually smaller than the body part being imaged. Chemical shift misregistration artifacts are the result of erroneous mapping of protons with different chemical shifts. They occur when fat and water are misregistered relative to each other. The increase in the sampling bandwidth per pixel can reduce the artifact. Truncation artifacts, also called Gibbs artifacts, occur as a result of Fourier transformation of truncated signals and appear as alternating light and dark bands called Gibbs ringing. Partial volume artifacts occur when a single pixel represents two tissues and could be resolved by increasing the spatial resolution. Increasing spatial resolution of the image will also help to reduce truncation and partial volume effects.

Radio frequency related artifacts occur due to the changes in the RF pulse sequences. These include cross talk which is due to radio pulses targeted at one slice affecting adjacent slices, zipper artifacts caused by deficiency or extraneous radio signals detected by the MRI receiving coil and RF noise. Magnetic susceptibility artifacts occur at the interfaces of tissues with different magnetic susceptibilities such as interface between tissue and air and lead to signal loss and distortion. This artifact is more pronounced in gradient echo techniques; it can be reduced by using spin echo with very short TE. Gradient related

artifacts occur due to eddy currents which are small electric currents caused by rapidly switching gradients. They produce distortions which result in reduced signal intensity in the periphery [4, 6].

External magnetic field artifacts are caused by magnetic inhomogeneities which are due to improper shimming, hardware imperfections, and environmental factors, and are unavoidable. They are usually characterized by the presence of smoothly varying and multiplicative intensity variations within the tissues, i.e., intensity non-uniformity, where voxels of the same tissue appear in different intensity at different locations. This leads to severe problems in the accurate measurement of brain structures. Though the artifacts can be minimized by the use of appropriate shimming coils, retrospective correction measures [15–17] are usually used to obtain better results.

The quality of MR images can be numerically evaluated using measures such as signal to noise ratio (SNR), contrast to noise ratio (CNR), and coefficient of variation (CV). SNR is a measure of the amount of noise present in the image and is defined as the ratio of the mean intensity of the region of interest (usually WM or CSF) to the standard deviation of the noise in the background that does not include the head [18]. Images with high SNR are generally of good quality where different structures in the image can easily be distinguished.

CNR is the ratio of the difference in mean intensities of two regions of interest to the standard deviation of the background noise. It gives information on how tissue differentiation is affected by noise [3, 19]. The noise here usually refers to intensity variations within a tissue class that includes background imaging noise and variations caused by inhomogeneities and partial volume effects [20, 21]. It has been observed that the intensity variation within a tissue class is larger than the standard deviation of the noise in the background. For functional MR images, the noise is defined as the non-task variability over time [22]. Though SNR and CNR are the most commonly used methods to define image quality, they correlate very little with the diagnostic utility and physician preference [23].

Coefficient of variation (CV) represents the ratio of the standard deviation to the mean and is an indirect measure used to evaluate the performance of nonuniformity correction [24]. It is evaluated on the basis of remaining tissue intensity variability (CV of WM, GM) rather than the actual estimated bias field.

2.2 Background on segmentation

MR brain imaging helps to visualize the different structures and also the active regions of the brain. In order to extract information useful for analysis from these images, suitable image processing procedures, such as segmentation, are required. The segmentation is complicated by the complex shapes of the anatomical structures, lack of clear boundaries between anatomically distinct tissues, and distortions due to noise, artifacts, and image inhomogeneities. Improved segmentation performance can be achieved by using any prior knowledge about the image such as information about the image modality, pixel intensity distribution, size, and shape of the structures. In this section, we provide a background on some of the most widely used techniques for MR brain image segmentation.

The segmentation techniques are generally based on three distinct concepts: thresholding, pixel discontinuity, and pixel similarity. Thresholding-based segmentation approaches are usually referred to as pixel-based methods. Segmentation techniques which work on the basis of similarity among the pixels in the local neighborhood are called region-based methods. They group pixels in the image into small regions based on some uniformity criteria. Techniques which work on the basis of pixel discontinuity, i.e., edge information, are called edge or boundary-based techniques. They typically use an initial boundary location and then deform it to fit the actual region of interest [25]. To illustrate these three main classes of segmentation techniques, we will use a simple model image (Figure 2.8(a)) consisting of two regions or labels, background region (π_0) and foreground region (π_1) with intensities $\mu_0 = 0$ and $\mu_1 = 0.4$ respectively. The model image possesses characteristics typical of an MR image which is comprised of several anatomical regions, each displayed as an area of uniform intensity with no texture. The regions have complex region boundaries and non-uniformity in intensity even within the same region. They are corrupted by noise mainly due to intensity inhomogeneities and artifacts. Figures 2.8(b)-(c) show the model images corrupted by i.i.d. Gaussian noise resulting in the SNR of 10dB and 4dB. Figure 2.8(d) shows the model image corrupted by an intensity gradient in the diagonal direction. These properties make the model images a good representation of MR images.

Figure 2.8: (a) Model image (b) model image corrupted with i.i.d. Gaussian noise, $SNR = 10dB$ (c) model image corrupted with i.i.d. Gaussian noise, $SNR = 4dB$ (d) model image corrupted by an intensity gradient in the diagonal direction

2.2.1 Thresholding

The simplest and the most straightforward technique for segmentation to classify image pixels is thresholding of the pixel intensities. This provides an easy way to extract the object from the background in images where intensities of the pixels of the object/foreground differ significantly from the intensities of the pixels of the background. Let I represent the model image and $I(s)$ represent the intensity of pixel s . The operation of thresholding can be defined as

$$G(s) = \begin{cases} 0, & \text{if } I(s) < T \\ 1, & \text{if } I(s) \geq T \end{cases} \quad (2.1)$$

where G is the output image after thresholding and T is the intensity threshold. The segmentation is determined by this single parameter T . In complex segmentation problems, multiple thresholds can be specified to set a range of values as foreground and the rest as

background.

The threshold T as in (2.1) is called the global threshold if it is applied to the whole image. If it is applied only within a local neighborhood, it is called the local threshold. The main problem in thresholding is the determination of the correct value of the threshold.

Hypothesis testing based thresholding

Let p_0 and p_1 be the probability density functions of pixel intensities in the foreground and background respectively. For identically distributed zero mean Gaussian noise, $p_0 \sim N(\mu_0, \sigma^2)$ and $p_1 \sim N(\mu_1, \sigma^2)$.

$$\begin{aligned} I(s) &\sim p_0 & \text{if } s \in \pi_0 \\ I(s) &\sim p_1 & \text{if } s \in \pi_1 \end{aligned} \quad (2.2)$$

Segmentation by thresholding can be formulated as a hypothesis testing problem, which aims to decide between the following two hypotheses:

$$\begin{aligned} H_0 : & \quad s \in \pi_0 \\ H_1 : & \quad s \in \pi_1 \end{aligned} \quad (2.3)$$

We illustrate the realization in Figure 2.9. Hypothesis testing leads to appearance of four categories of pixels, depending on whether or not the pixel truly belongs to class π_1 and whether or not it is declared as belonging to π_1 , as shown in Table 2.1. Two errors arise in this hypothesis testing, the type I error (probability of false alarm P_{FP}) and the type II error (probability of miss detection P_{FN}). Type I error $p(H_1|H_0)$ occurs when we decide H_1 when H_0 is true, i.e., declaring a π_0 pixel to belong to class π_1 and type II error $p(H_0|H_1)$ occurs when we decide H_0 when H_1 is true, i.e., declaring a π_1 pixel as belonging to class π_0 .

Table 2.1: Classification of pixels

	Declared class π_1	Declared class π_0
Truly class π_1	P_{TP}	P_{FN}
Truly class π_0	P_{FP}	P_{TN}

Figure 2.9: Type I and type II errors

Also called the binary hypothesis testing, this method requires prior knowledge about the probabilities of occurrence of each hypotheses. If the prior probabilities are equal, the optimal decision is taken based on the likelihood ratio test given by

$$L = \frac{p_1(I(s))}{p_0(I(s))} \geq 1 \quad (2.4)$$

which is usually a monotonic function of pixel intensity. Therefore, the likelihood test is equivalent to comparing the pixel intensity to a threshold T . For i.i.d. Gaussian noise $T = \frac{\mu_0 + \mu_1}{2}$, and therefore optimal thresholding does not require the knowledge of the variance of the noise distribution. Figure 2.10 shows the result of optimal hypothesis testing, assuming known distribution means and equal priors.

Figure 2.10: Optimal hypothesis testing applied to (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)

Otsu's thresholding

Otsu's method [26] of threshold determination assumes that the histogram of the image is bimodal and that the illumination is uniform. The method operates directly on the gray level histogram and hence is fast. It minimizes the weighted sum of within-class variances of the foreground and background pixels to establish an optimum threshold. This method gives satisfactory results when sizes of the two classes are close to each other. It fails to provide a good threshold when one class is large and the other class is small. Figure 2.11 shows the result of applying Otsu's threshold to binarize the model images in Figure 2.8. The presence of noise produces many spurious regions in the output giving large type I error and intensity gradients in the image lead to poor performance.

Figure 2.11: Otsu's thresholding of (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)

Though there are many methods available for threshold selection, there is no universal procedure that guarantees to work on all images. Sezgin et al. [27] provide a detailed survey of image thresholding techniques. Since thresholding is based on pixel intensities, spurious regions are likely to appear in the presence of noise in the original image. Furthermore, thresholding does not guarantee connectivity between the regions.

2.2.2 Region growing and region splitting-merging

Region growing and region splitting and merging are the basic region-based segmentation techniques.

Region growing

Region growing groups pixels into coherent regions. Coherency is defined based on the homogeneity in the intensity of pixels within the region. Some probabilistic criteria defining the distribution of pixels can also be used to decide the growth of the region. The most popular and widely used criterion is uniformity in the intensity of the pixels. The uniformity criterion could be defined in terms of the absolute difference between the intensity of the neighboring pixels or the absolute difference between the pixel intensity and the average region intensity falling below a predefined threshold τ . The region growing method starts by choosing a seed pixel. The region is then grown from the seed pixel by adding the neighboring pixels that satisfy the uniformity criterion. For images with many regions, multiple seed selection can be employed. Different criteria for region growing can be used depending on application [28–30].

Shown in Figure 2.12 is the result of region growing technique applied to the model images in Figure 2.8 using the absolute difference between the neighboring pixels as the uniformity criterion. Small τ values lead to region underestimation while larger τ may lead to overestimation (inclusion of additional regions). High intensity uniformity among the pixels is required in order to achieve good results. In the presence of noise, the region growing procedure tends to leak into nearby regions of similar intensity distribution through weak boundaries as can be seen in the images in the third column of Figure 2.12. This uniformity criterion fails completely for the gradient corrupted image because of the less variation between the pixels. Figure 2.13 shows the result of region growing using the difference between pixel intensity and the average region intensity as the criterion. Though this criterion does not give satisfactory results for the noise corrupted image, it works well on the gradient image. The poor segmentation obtained earlier for the gradient image can be avoided using this criterion. Region growing depends heavily on the selection of the right seed point and the optimal value of the threshold which usually require manual intervention. We used the code obtained from Matlab central file exchange to implement region growing.

Figure 2.12: Region growing using difference between the intensity of neighboring pixels as the criterion: First row - model image (10dB), second row - model image (4dB), third row - model image (intensity gradient)

Figure 2.13: Region growing using difference between pixel intensity and average region intensity as the criterion. (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)

Figure 2.14: Region splitting and merging: First row - model image (10dB), second row - model image (4dB), third row - model image (intensity gradient)

Region splitting and merging

Region splitting and merging considers the entire image as the seed region. The pixels inside the seed region are checked for the similarity condition. If they are not found to be homogeneous, the region is split, typically into four sub regions. Then, each new region is considered as a seed region and the process of splitting is continued until all pixels in the region are homogeneous. The splitting step may result in adjacent regions with similar properties. The merging step is used to overcome this drawback by merging the sub regions which are homogeneous. Since the criterion for these approaches is based on the intensity

of the pixels, the quality of segmentation is highly sensitive to image brightness [31]. The advantage of this method over region growing is that no seed points are needed. Since the region is usually a square block of size in powers of 2, the result may have a checkerboard effect. Taking smaller values for the size of the smallest block allowed s_b helps to reduce this effect, but it results in more number of spurious regions, see Figure 2.14. An optimal value of s_b has to be chosen for good results. This method completely fails for the model image corrupted with intensity gradient due to the difficulty in finding a homogeneous region. We used the code obtained from Matlab central file exchange to implement region splitting and merging.

2.2.3 Watershed segmentation

Watershed algorithms for segmentation are a special case of region based methods. The use of the concept of watersheds for image segmentation started with the work of Beucher et al. [32]. A good review of different algorithms used for watershed segmentation is given in [33]. Watershed segmentation works on the basis of image intensities, which are interpreted as height information, i.e., the whole image is viewed in three dimensions, two spatial coordinates and the intensity values. Under this topographic interpretation of intensity values, bright pixels represent the hills and dark pixels correspond to the valleys. Two regions are said to be connected if there is no valley separating them. The image is then segmented into different regions, each of which corresponding to a connected basin.

Figure 2.15: Watershed segmentation of (a) model image (10dB) (b) second row - model image (4dB)

There are many algorithms available to compute watersheds in images, most of which

are based on mathematical morphology. One of the main drawbacks of the watershed algorithms is that they frequently result in oversegmentation due to noise and local irregularities of the gradient (see Figure 2.15). In the case of image corrupted with intensity gradient, watershed may fail to separate the basins and consider the entire image as a single basin. Some form of pre/post-processing is required to overcome this problem. One popular solution is the marker controlled watershed segmentation [34]. Hahn et al. [35] proposed a 3D watershed transform which includes a merging procedure called preflooding to prevent oversegmentation. More details on preflooding are provided in Section 2.3.4.

2.2.4 Active contour models

Boundary based techniques use edge information to locate the object boundaries. The active contour models have been an intensely researched area since they were originally proposed by Kass et al. [36]. Popularly called snakes, the active contours are computer generated curves that undergo deformation to fit the region of interest. The deformation is controlled by a set of internal forces that are defined within the curve, set of external forces that are computed from the image data and the optimization function. The internal forces typically comprise the elasticity forces and the bending forces that are designed to keep the curve smooth during deformation and prevent it from stretching and bending too much. These forces are in general designed using the prior shape information/smoothness. The external forces move the snake towards the desired boundary and are usually represented by edge information. The main advantage of these deformable models is their mechanism to allow expert knowledge in the form of prior information to be incorporated in the model-based image segmentation.

The active contour models can be further subdivided into the parametric active contours and geometric active contours. Parametric deformable models represent curves explicitly in their parametric forms during deformation [37, 38]. This parametric representation allows direct interaction with the model. However the main drawback with this method is that it cannot adapt to topological changes, such as splitting or merging during the deformation. Geometric deformable models, on the other hand, can handle topological changes naturally [39]. These models are based on the theory of curve evolution and the level set method [40, 41], which represents the curves and surfaces implicitly as a level set

of a higher-dimensional scalar function. Their parameterizations are computed only after complete deformation, because of which they have the capability to adapt to topological changes. Despite this fundamental difference, the underlying principles of both methods are very similar. A good survey of the deformable models for medical image analysis can be found in [42].

Figure 2.16: Snakes: First row - close initialization on model image with intensity gradient and its corresponding output, second row - distant initialization and output on model image with intensity gradient, third row - initialization and output on model image corrupted with noise (smoothed)

Parametric active contours - snakes

The parametric deformable models use an energy minimizing formulation of deformable contours that minimizes the weighted sum of the internal energy and external energy. The

snake is a parametric curve $\mathbf{x}(s) = (x(s), y(s))$, where $s \in [0, 1]$ and the energy function to be minimized is of the form

$$E(\mathbf{x}) = E_{int}(\mathbf{x}) + E_{ext}(I, \mathbf{x}) \quad (2.5)$$

where I represents the image. The internal energy is given by the form

$$E_{int}(\mathbf{x}) = \int_0^1 \left(\alpha |\mathbf{x}'(s)|^2 + \beta |\mathbf{x}''(s)|^2 \right) ds \quad (2.6)$$

where α and β are the coefficients associated with the elastic energy and the bending energy respectively. The external energy is usually a function of the image gradient and is given by

$$E_{ext}(\mathbf{x}) = \int_0^1 -|\nabla I(x, y)|^2 \quad (2.7)$$

The snake that minimizes $E(\mathbf{x})$ has to satisfy the Euler equation

$$\alpha \mathbf{x}''(s) - \beta \mathbf{x}''''(s) - \nabla E_{ext} = 0 \quad (2.8)$$

The optimal position of the contour is obtained when the external forces and the internal forces balance each other.

Figure 2.16 shows typical results of snakes obtained using the graphical demonstration software written by Dejan Tomazevic on the basis of [43]. Snakes suffer from two main problems, close initialization requirement and inability to progress towards boundary concavities. Further, the noisy model images need to be smoothed for the snakes to successfully segment the foreground. The close initialization requirement can be overcome by the use of multiresolution methods [44], pressure forces [45], and distance potentials [46]. The basic idea is to extend the capture range of the external force fields to guide the contour towards the desired boundary. The use of gradient vector flow (GVF) fields [43] captures the snake at long range and forces it into concave regions but does not solve the problem completely. Further extensions of GVFs to improve the results can be found in [47, 48].

Figure 2.17: Level sets on (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)

Geometric active contours - level sets

Geometric deformable models are represented implicitly as level sets of higher-dimensional scalar function and evolved in an Eulerian framework [38, 49]. In this framework, curves evolve using geometric measures, resulting in a contour evolution that is independent of the curve's parameterization. This avoids the need to repeatedly reparameterize the curve or to explicitly handle topological changes. Level sets are a class of geometric deformable models proposed originally by Osher and Sethian [40]. The level set model implicitly represents a curve as the level set of a signed distance function $\phi(x, y, t)$, which varies in space and time. Its value at (x, y) is equal to the signed minimum distance from the pixel to the curve. The sign is positive or negative if the pixel is inside or outside the zero level curve

$$\phi(x, y, t) = 0. \quad (2.9)$$

The curve evolution is controlled by a speed function that defines the speed with which the curve moves. Differentiating both sides of (2.9) and applying the chain rule gives

$$\nabla \phi(x, y, t) \mathbf{x}'(t) + \frac{\partial \phi(x, y, t)}{\partial t} = 0 \quad (2.10)$$

where $\mathbf{x}'(t)$ is the speed term representing the movement of the point $\mathbf{x} = (x, y)$ over time and ∇ is the gradient operator.

Level set models have significant advantages over other contour models. They are topologically flexible and can represent complicated shapes that either split to form multiple objects or merge with other objects to form another object, see Figure 2.17. For the

implementation of level sets, we used the software written by Chunming Li on the basis of [50]. Similar to snakes, the images need to be smoothed for level sets to produce better segmentation. From the output images, we can see that level sets produce satisfactory segmentation even in the presence of noise and intensity gradient. There is no need to re-parameterize the model as it undergoes changes in shape. One major limitation of the method is its strong dependence on the choice of the speed function. Choosing an incorrect function may result in wrong segmentation. [51–53] provide improved versions of level set segmentation using priors on shape and intensity. A fast implementation of level set method is provided in [54].

2.2.5 Graph theoretic approaches

Similar to active contour models, graph theoretic approaches represent the segmentation problem in an energy minimization framework. The difference between them is that discrete energy minimization function is defined in the case of the latter and a continuous function in the case of the former. Graph theoretic approaches represent the image in the form of a weighted undirected graph $G = (V, W, E)$, where each vertex $v_i \in V$ corresponds to a pixel in the image, edge $e_{ij} \in E$ connects neighboring vertices v_i and v_j and $w_{ij} \in W$ is the weight assigned to the edge e_{ij} . Weighted graphs are generalized form of unweighted graphs, where $w_{ij} = 1$ for all $e_{ij} \in E$.

The graph theoretic approaches can be classified into two categories: boundary-based and region-based, each requiring different graph structure. Region-based approaches [55] segment regions with similar properties and do not require selection of seeds. Here, every vertex in the graph is also connected to two additional vertices (source and sink). The weights of the edges are assigned based on the pixel intensity distribution. Figure 2.18(a) shows an example of a graph constructed for a 5×5 image.

Boundary-based graph theoretic approaches [56,57] try to find the boundaries between the regions. They require the selection of both the foreground and background seeds. The background seeds are connected to the source and the foreground seeds to the sink. Figure 2.18(b) shows an example of the boundary-based graph theoretic approach. The shaded region corresponds to the foreground and the rest of the pixels correspond to the background region. The weights of edges connecting background seeds with the source

Figure 2.18: Graphical framework for a 5 x 5 image (a) region-based approach (b) boundary-based approach

are denoted by W_0 and the weights of edges connecting foreground seeds with the sink are denoted by W_1 . These weights are assigned some larger value so that they are never cut. The weights of the edges between pixels are assigned usually based on the difference in intensities of the pixels they connect or the shape information.

A graph cut is a set of removed edges that partitions the vertices in the graph into two disjoint sets, with sink and source separated from each other. The cost of the cut is the sum of the weights of all the removed edges. The minimum cut on the graph produces the optimal segmentation and it can be determined by computing the maximum flow on the graph. The maximum flow on the graph is equal to the minimum cut and it can be found using the Ford-Fulkerson maximum flow algorithm in polynomial time [58]. In our experiments, we used the minimum cut algorithm implementation by Boykov and Kolmogorov [59].

Region-based graph theoretic approaches

Region-based graph theoretic approaches are designed to find regions with uniform properties. The distribution of the foreground as well as the background pixels is given as prior information into the segmentation process. This avoids the need for selecting the

foreground and background seeds explicitly. The weights of the edges connecting the vertices to the source/sink are assigned based on the foreground/background pixel intensity distributions. Let $W_0(s)$ represent the weight of the edge connecting vertex s with the source and $W_1(s)$ represent the weight of the edge connecting s with the sink. The sink and source represent foreground and background respectively. If a pixel is more likely to belong to foreground, the weight that connects it to sink should be made large, while the weight that connects it with the source should be made small. All other edge weights are usually set equal to some predefined constant β . The value of β is chosen empirically. One example of weight assignment motivated by MRF segmentation is as follows:

$$W_0(s) = -\log p_0(I(s))$$

$$W_1(s) = -\log p_1(I(s))$$

where p_0 and p_1 are as in (2.13). Figure 2.18(a) shows an example of a region-based graph framework. The segmentation is then obtained by finding the minimum cut that separates the source and the sink on this graph. A good example for this approach is the Markov random field (MRF) segmentation solved using graph cuts (refer to Section 2.3.6).

Boundary-based graph theoretic approaches

Boundary-based graph theoretic approaches are employed to determine the boundary between regions, similar to the approach taken by active contours. They work by cutting along the boundaries of the region. This approach requires the selection of foreground seeds and background seeds. These seeds act as hard constraints in the segmentation process, i.e., pixels selected as foreground seeds always appear as foreground and pixels selected as background seeds always appear as background in the output. The background seeds and foreground seeds are connected to the source and sink respectively. The weights of the edges connecting these seeds with the source/sink are assigned some large values to guarantee that these edges are not severed when finding the minimum cut. The following

formula can be used as a guideline for the minimum value of the sink/source edge weights.

$$W_0(s) = W_1(s) \geq \sum_{k \in N_s} w_{sk}$$

where N_s is the neighborhood of s consisting of all its adjacent vertices. The weights of edges connecting the vertices are chosen in such a way that the weights of the edges connecting likely boundary pixels are small while the rest of the weights are large. The edge weight assignment function can be chosen depending on the application. Finding the minimum cut that separates the two seeds yields a globally optimal segmentation under the chosen edge weight criterion. Figure 2.18(b) shows an example of the graph framework for the boundary-based approach.

Given below is an example of a function [56, 60, 61] to determine the weights of the edges connecting the vertices of the graph.

$$w_{ij} = \exp\left(-(I(i) - I(j))^2\right)$$

where w_{ij} is the weight of the edge e_{ij} connecting the vertices v_i and v_j and $I(i)$ is the intensity value at vertex v_i . This edge weight assignment gives smaller weights to the edges connecting less similar pixels, which are likely to belong to the boundary between regions. The results of boundary based graph cuts are shown in Figure 2.19. It produces satisfactory results even in the presence of noise and is robust to intensity gradients in the image. However, its performance depends on the selection of the seeds. Usually seeds have to be sufficiently large to make sure that minimum cut does not select a region that is too small. This could happen because graph cuts penalizes longer cuts over shorter ones [62].

In order to overcome this bias towards smaller regions, Shi and Malik [61] proposed the normalized cut criterion, which is defined as the ratio of the cost of the removed edges to the total cost of the edges inside the region. This ratio is large for smaller regions and hence the cut will prefer larger regions. Eigen vector approximation is used for solving the NP-hard problem of minimization of the normalized cut criterion. Wang and Siskind [63] addressed this by the ratio cut, that minimizes the mean value of a cut by dividing the total weight by the number of removed edges, thus favoring regions with the highest contrast

Figure 2.19: Boundary based graph cuts (a) foreground and background seeds (b) model image (10dB) (c) model image (4dB) (d) model image (intensity gradient)

boundary. However, this tends to produce cuts around spurious regions if the image is noisy. Hochbaum [64] proved that the NP-hard problem of a variant of normalized cut can be solved in polynomial time using a monotone integer programming formulation. Interactive graph cuts approach [56] used s-t minimum cut as an optimization method for user interactive selection of object and background regions. Xu et al. [65] assumed that the cut corresponding to the desired object boundary is a global minimum of the cuts inside some contour neighborhood specified by the user and aimed to find the closest contour that is a global minimum within its neighborhood, given an initial contour. Rother et al. [66] used an iterative version of optimization of graph cut and developed grab cut which has a simpler user interaction and almost same performance as graph cuts. Vicente et al. [67] used connectivity priors in graph cuts to improve the segmentation of elongated objects. Couprie et al. [68] proposed power watersheds which is a general seeded segmentation framework that extends graph cuts, random walker [69], shortest path segmentation [70] and watersheds using the earlier work [71]. Cousty et al. [72,73] developed watershed cuts which introduced the notion of watershed in edge-weighted graphs leading to improved

delineation in watershed based segmentation procedures.

2.2.6 Summary

Table 2.2 provides the performance comparison of the segmentation techniques based on the algorithm and implementation mentioned in the respective sections. However, it must be noted that this comparison is limited to the most basic versions of each algorithm and some drawbacks can be mitigated by appropriate modifications. Though it might be impossible to decide the best out of all the methods, this performance comparison might aid in selecting an appropriate method for a particular application. For example, graph theoretic approaches can be chosen over active contours or watershed algorithm for images that contain large amount of noise. Boundary-based approaches perform better than region-based approaches when the image contains smooth intensity gradients.

Table 2.2: Performance comparison of segmentation techniques

Property	Thresholding	Region growing	Watershed algorithm	Active contours	Graph theoretic methods
Algorithm complexity	simple	simple	average	complex	complex
Speed	very fast	fast	average	average	average
Sensitivity to noise	very high	high	very high	average	low
User interaction	sometimes	yes	no	yes	yes
Sensitivity to smooth image gradients	high	low (only for uniformity based on average region intensity), high for others	average	low	low (only for boundary-based), high for region-based
Single / multiple region segmentation	arbitrary	arbitrary but requires multiple user inputs for multiple regions	arbitrary	arbitrary for level sets, snakes require multiple user inputs for multiple regions	arbitrary (region-based), multiple user inputs (boundary-based)

One of the ways to improve the performance of all segmentation algorithms is to incorporate prior knowledge about the image into the segmentation procedure. This is discussed in the next section.

2.3 Incorporation of prior information

The segmentation performance can be improved by incorporating suitable prior information into the segmentation procedure. We consider all information that is known beforehand and is useful for segmentation as prior information and do not perceive it in a statistical Bayesian sense, as a prior probability distribution on intensity, boundary shape, etc. Some common types of such prior information are smoothness, minimum cluster size, shape simplicity - absence of narrow connections or compactness. While some of these can be expressed in the form of a probability distribution, e.g. smoothness in the MRF model, most are too complex or abstract and thus cannot be. For example, we do not know how to express shape compactness, defined as isoperimetric ratio, in the form of a probability distribution of the boundary shape. However, we refer to the algorithm that minimizes isoperimetric ratio as the algorithm that incorporates prior information (or prior knowledge) that the shape is expected to be as compact as possible. In this section, we discuss how these prior information can be added to the segmentation procedure.

2.3.1 Smoothing

In the presence of noise, thresholding and region-based methods may produce many spurious regions. One way to remove the impact of noise in the image is smoothing. Smoothing blurs the image, improves the SNR and reduces the probability of error. Thresholding of the smoothed image thus reduces the number of spurious regions. The simplest way to incorporate the smoothness constraint is to apply a smoothing filter, Gaussian filter for example. However, selecting the appropriate window size for the filter remains a concern. In Figure 2.20, we convolved the noisy corrupted model image with 3×3 and 7×7 Gaussian filters. Using larger window sizes leads to better removal of spurious regions, but tends to oversmooth the corners. The filter size is a trade-off between noise reduction and preservation of object boundaries.

Figure 2.20: Gaussian smoothing and Otsu's thresholding: First row - model image (10dB) processed with (a) 3×3 filter (b) 7×7 filter, second row - model image (4dB) processed with (c) 3×3 filter (d) 7×7 filter

2.3.2 Cluster size thresholding

In medical images, the objects of interest occur in groups/clusters rather than single pixels. Using statistical language, we can say that a pixel has more probability to belong to the region if its neighboring pixels belong to that region. Thus, it is more appropriate to process the pixels jointly rather than individually.

From this it follows that true pixels tend to occur in relatively large clusters, while spurious ones occur either individually or in very small clusters. Hence, the latter can be removed based on some cluster size. This approach is called cluster size thresholding. Shown in Figure 2.21 is the result of applying cluster size thresholding on the Otsu thresholded images in Figure 2.11. As is the case with normal intensity thresholding, selection of the right threshold for cluster size is a difficult task. Smaller cluster size threshold does not remove spurious regions properly and causes increase in the false alarm rate (type I error). Larger cluster threshold performs better in removing spurious regions, but may

Figure 2.21: Cluster size thresholding applied to (a) Figure 2.11(a) (b) Figure 2.11(b)

remove some of the true regions as well, increasing the miss detection rate (type II error). Use of prior information regarding the size of the regions in the image may be of significant advantage in the segmentation process.

2.3.3 Morphological processing

Morphological operations are used to improve the segmentation result by removing narrow connections or smoothing the boundary. Morphological image processing is based on the mathematical field of set theory and is used to identify objects within binary and grayscale images. Its main goal is to distinguish meaningful shape information from irrelevant one. This is achieved by designing a shape operator, called the structuring element and performing certain basic operations with the structuring element on the image [74]. By varying the size and shape of the structuring elements, we can extract useful information about the shape of different parts of the image. The most commonly used ones are the disc, square, and rhombus shaped elements.

Erosions and dilations are the two most elementary operations and more complex operations can be designed by combining these. Opening (erosion followed by dilation) and closing (dilation followed by erosion) are examples of secondary operations. Opening operation is used to disconnect the regions, which were erroneously connected by spurious pixels, while closing operation can be used to fill in the holes formed by pixels wrongly classified as background. Both operations have a smoothing effect on the image. Amount and type of smoothing is determined by the shape and size of the structuring element.

By suitably combining these basic operations, we can perform thinning, thickening, skeletonization, pruning, distance transform, and so on.

Figure 2.22: Morphological opening operation applied to the Gaussian smoothed image in Figure 2.20 (b) using disk structuring element

The application of morphological operations is effectively an incorporation of smoothing and shape constraint into the segmentation process. Due to intensity overlap between different regions of interest, traditional segmentation techniques often fail to isolate the region of interest. Narrow connections exist between the region of interest and other irrelevant regions in the image. This is one of the widely occurring instances in the segmentation of medical images, particularly with MR brain images and echocardiographic images. Such connections can be removed by morphological opening. Similarly, suitable combination of morphological operations provide an easy way to extract the largest connected component in the image, which is another requirement most commonly encountered in medical image segmentation. Figure 2.22 shows the result of morphological opening operation performed on the Gaussian smoothed image in Figure 2.20 (b). Due to the presence of more noise, smoothing blurs the boundary between the regions and merges them. Applying opening operation separates the regions.

2.3.4 Watershed segmentation - preflooding

Watershed segmentation is highly sensitive to noise. Noise can create spurious basins and the size of the basins or their depth can not be controlled by the traditional watershed segmentation. The presence of noise and variation in the intensity levels in the image usually result in oversegmentation. Hahn [35] proposed a 3D watershed transform which

includes a merging procedure called preflooding to prevent oversegmentation. In this procedure, the condition of connectivity between pixels is relaxed to allow for the variation in the intensity of pixels in the region. This variation in intensity is allowed up to a maximum difference which is termed as the preflooding height h_{pf} . This denotes the depth of the basin.

Figure 2.23: Illustration of preflooding - merging procedure

By preflooding, all neighboring basins whose depth relative to the current pixel intensity is less than or equal to h_{pf} will be merged with the same basin as the pixel itself, i.e., the deepest neighboring basin (see Figure 2.23). After the transform with an appropriate value of preflooding height, smaller regions are merged to form a single large basin. Thus preflooding provides a way to incorporate the smoothness constraint into the watershed segmentation process. The value of h_{pf} indicates the robustness of the segmentation to noise in the image, the broader the range, the more robust is the segmentation. Figure 2.24 shows the results of watershed transform with preflooding on the model images. Since the model images are very noisy, preflooding includes some adjacent regions as well. From the results, we can see that preflooding helps in segmenting the foreground regions partic-

ularly in the model image corrupted with the intensity gradient which completely failed without preflooding.

Figure 2.24: Watershed transform with preflooding applied to (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)

The value of the preflooding height determines the amount of smoothness in the image and selection of the appropriate preflooding height becomes an important task for quality segmentation. The number of basins monotonically decreases whereas their sizes increase with increasing preflooding height. Further it was shown empirically on MR brain images that a preflooding height $h_{pf} = 0.11I_{max} + 3.5n$ works well [35]. I_{max} here is the maximum intensity value in the image and n indicates the variance of the noise in the image.

2.3.5 Graph cuts - isoperimetric ratio constraint

Minimum cut on the graph tends to penalize longer boundaries and favor smaller regions. This problem of graph cuts showing bias towards smaller regions can be overcome by incorporating shape and size constraint into the graph theoretic framework.

Grady [75] introduced a graph partitioning algorithm based on the optimization of the isoperimetric constant. It is influenced by the properties of the classic isoperimetric problem to find a region with the minimum perimeter for a fixed area. The algorithm finds partitions that minimize the isoperimetric constant h defined as

$$h = \inf_S \frac{|\partial S|}{Vol_S} \quad (2.11)$$

where S is a region, $Vol_S \leq \frac{1}{2}Vol_{total}$ denotes the volume of S and $|\partial S|$ is the length of

the boundary or surface area of S . In other words, the algorithm finds a compact region, where the compactness is defined as the ratio of perimeter and volume. The method requires only single pixel foreground seed for initialization and [75] suggested using the node in the graph that has the maximum degree.

A modified version of the algorithm can be used to remove weak connections between regions, termed as mask segmentation [76]. We will return to this topic later, in Section 4.2.3. Applications of this algorithm to medical and natural images are shown in [76] and [77] respectively. Figure 2.25 shows the result of isoperimetric segmentation on the model image in Figure 2.20. The algorithm is sensitive to noise and fails on both the noisy model images. However, it produces excellent result on the intensity gradient corrupted image. Another limitation of the algorithm is its sensitivity to the position of seed point. Furthermore, the algorithm does not provide the option to select multiple seed points to process multiple regions. Finding the isoperimetric sets is a NP-hard problem and the problem is usually simplified by using a technique called relaxation, which reduces the optimization to a solution of a set of linear equations.

Figure 2.25: Isoperimetric segmentation applied to (a) model image (10dB) (b) model image (4dB) (c) model image (intensity gradient)

2.3.6 MRF based segmentation

Markov Random Field (MRF) is a statistical model which provides a systematic way to incorporate prior information regarding the spatial connectivity (smoothness) between the neighboring pixels into the segmentation model. Let S be a set of pixels in the image, L be the set of all labels and the neighborhood be defined by N . Let X be the set of

random variables defined on the set S where $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in L$ is the label for s_i . Then, X is said to be a MRF with respect to the neighborhood N if and only if it satisfies $P(x) > 0$ (positivity property) and $P(X_s = x_s | X_r = x_r, \forall r \neq s) = P(X_s = x_s | X_r = x_r, \forall r \in N_s)$ (Markovian property).

Figure 2.26: MRF based segmentation: First row - model image (10dB), second row - model image (4dB), third row - model image (intensity gradient)

The Hammersley-Clifford theorem establishes the equivalence between MRF and Gibbs random field (GRF) and states that the probability density for MRF has a Gibbs distribution

$$P(x) = \frac{1}{Z} \exp \left(- \sum_{c \in C} V_c(x_c) \right) \quad (2.12)$$

where c is a clique, C is a set of all cliques, $V_c(x_c)$ is the clique potential and Z is the normalization constant. $V_c(x_c)$ describes the prior probability of the elements of the clique c and is equal to $\beta L(\pi)$ (for Ising model), where $L(\pi)$ is the length of the boundary and β is a constant. The segmentation is formulated as a maximum a posteriori probability (MAP) estimation problem, which in the case of bi-level partition $\pi = (\pi_0, \pi_1)$ and Ising MRF model is equivalent to the minimization of the following energy cost function

$$J(\pi) = - \sum_{\pi_0} \log p_0(I) - \sum_{\pi_1} \log p_1(I) + \beta L(\pi) \quad (2.13)$$

Here p_0 and p_1 represent the probability density functions of the pixel intensities (I) in the background (π_0) and foreground (π_1) respectively. The Ising MRF model penalizes longer boundaries and favors small regions. It can be solved by determining a region-based graph cut with appropriate selection of weights (refer to Section 2.2.5). Figure 2.26 shows the results of MRF based segmentation of the different model images. Overall, MRF based segmentation has the least sensitivity to noise among all the methods discussed so far. However, the method fails for the intensity gradient corrupted image. The result also depends on the value of the smoothness constant β , with larger β leading to better noise suppression and smoother segmentation.

2.4 Discussion

Accurate segmentation of MR images is a complex task and might require prior information about the input images in the form of size, intensity, shape, etc. In this chapter, only the most basic versions of each of the segmentation algorithm are discussed and compared. The drawbacks of the algorithms can be alleviated to a certain extent by employing suitable modifications. Incorporation of prior information is one of the steps to alleviate the limitations of the algorithms. This can help in increasing the performance of the segmentation procedure. In the case of the problems addressed in this thesis, the existing techniques do not solve the problems as can be seen in the corresponding chapters. Therefore, new techniques need to be developed to solve the problems.

Chapter 3

MRF segmentation with cluster size constraint

Functional Magnetic Resonance Imaging (fMRI) is an experiment that allows measuring metabolic changes taking place in an active part of the brain. It uses statistical processing of MR image sequence to visualize the activated areas. Since the actual changes in intensity are subtle compared to image noise, typical fMRI experiment consists of multiple (tens or hundreds) repetitions of stimulation (task) and rest, to increase the statistical power of detection. The detection of activated areas is performed by the computation of a statistical parameter indicating the significance of activation for every pixel in the image. These parameters, obtained over all the voxels result in the statistical parametric map (SPM) of the brain. In other words, SPM is an image in which intensity values represent statistical significance (p-value) obtained under the null hypothesis of no activation.

Detection of brain activations helps in the identification of brain structures involved in various cognitive tasks, such as visual perception, problem solving, memory, etc. This detection can be considered as an instance of segmentation problem on the SPM of the brain. What makes this problem difficult compared to standard segmentation is the fact that the activation regions are usually small and the noise is large, despite the multiple task repetitions.

In this chapter, we concentrate on the problem of detecting small activated regions in fMRI. We review earlier techniques used for the detection of brain activations. We

then show that in the context of Markov random field (MRF) based segmentation, cluster size thresholding - simple removal of small regions after the segmentation is not optimal. We propose a novel method, the retrospective correction approach that allows the regions to grow before they are eliminated based on their size. The proposed approach achieves better performance compared to simple removal of small regions after MRF segmentation.

3.1 Review of existing techniques

The segmentation in fMRI aims to separate the image into activated and non-activated regions. The most commonly used approach is thresholding of individual pixel intensities. Since thresholding does not take into account the spatial correlation between the activated pixels, appearance of spurious regions is likely. Processing pixels jointly helps to suppress spurious regions and improve the segmentation performance. Modeling prior information in a statistical model provides another way to improve the performance of the segmentation. In this section, we review some of the techniques used for the detection of activated regions in fMRI.

3.1.1 Thresholding of pixel intensities

Thresholding of individual pixel SPM values is probably the most widely used solution to activation detection in fMRI [78]. The thresholding techniques to isolate active regions can be grouped into three main categories, type I error control thresholding, false discovery rate (FDR) control thresholding, and posterior probability thresholding [79]. Type I error control thresholding comes under the following hypothesis testing problem:

$$\begin{aligned} H_0 : & \text{ pixel is inactive (background), } I(s) \sim N(\mu_0, \sigma^2), \mu_0 = 0 \\ H_1 : & \text{ pixel is active (foreground), } I(s) \sim N(\mu_1, \sigma^2), \mu_1 > 0 \end{aligned} \quad (3.1)$$

It is a composite hypothesis testing problem where the second distribution is represented as a family of distributions parameterized by the parameter μ_1 . The result of hypothesis testing produces four categories of pixels depending on whether or not the pixel is truly active and whether or not it is declared active, as shown in Table 3.1. Here, n is the number of pixels, D_a denotes the number of pixels declared active and D_i denotes the

number of pixels declared inactive. Two errors arise in this hypothesis testing, the type I error (probability of false alarm or false positives, P_{FP}) and the type II error (probability of miss detection or false negatives, P_{FN}). Type I error ($p(H_1|H_0)$) occurs when we decide H_1 when H_0 is true, i.e., declaring an inactive pixel to be active and type II error ($p(H_0|H_1)$) occurs when we decide H_0 when H_1 is true, i.e., declaring an active pixel to be inactive.

Table 3.1: Outcomes of hypothesis testing

	Declared active	Declared inactive	Total
Truly active	P_{TP}	P_{FN}	T_a
Truly inactive	P_{FP}	P_{TN}	T_i
Total	D_a	D_i	n

Figure 3.1: Controlling type I error at the expense of type II error

Determining a direct value of threshold is difficult since the second distribution is unknown, which makes it impossible to evaluate type II error. However, regardless of μ_1 , the optimal decision rule is still a thresholding. The threshold is chosen in such a way that the type I error is less than or equal to α , a predefined significance level, usually 0.05.

3.1.2 Controlling family-wise error and false discovery rate

Let us consider an image consisting of a thousand pixels (or voxels). In this case, it is not sufficient to guarantee 0.05 FP rate, because this small chance will be multiplied by the

number of pixels. In statistics, this problem is usually referred to as multiple independent tests problem, where the number of tests is equal to the number of pixels. Family-wise error (FWE) is the probability of getting at least one false positive among all hypotheses when performing multiple tests. The Bonferroni procedure is the simplest way to control the probability of having any false positives among all the tests. The method divides the threshold significance level α by the number of hypotheses n for each test. The new threshold significance level $\alpha' = \alpha/n$ guarantees that FWE is never larger than α . The problem with this procedure is that it is too conservative. By controlling the probability of false positives, each test is set at a high standard thereby increasing the probability of type II error. Some of the other techniques for controlling FWE are based on random field theory and resampling. Reviews of these techniques are provided in [80] and [81]. A good review of FWE controlling techniques can be found in [82].

False discovery rate (FDR) is the proportion of false positives (P_{FP}) to the total number of positives ($P_{FP} + P_{TP}$).

$$FDR = \frac{P_{FP}}{P_{FP} + P_{TP}} \quad (3.2)$$

By controlling FDR at some predefined level q , the expected proportion of incorrectly rejected null hypotheses (type I errors) is reduced [83, 84].

$$E(FDR) \leq \frac{P_{FP} + P_{TN}}{n} q \leq q \quad (3.3)$$

It is a less conservative procedure for comparison, with greater power than FWE control, at a cost of increasing the likelihood of obtaining type I errors [85]. Controlling FDR does not limit the number of false positives, but if a certain activation map is produced, no more than q proportion of all activated pixels are false positives and the rest $(1 - q)$ proportion are true positives.

In posterior probability thresholding, the threshold is applied to the estimated map of posterior probabilities that each pixel is activated. Hartvig and Jensen [86] formulated a spatial mixture model that takes into account to some extent, the spatial structure of the activation pattern. The posterior probability map obtained is thresholded at a significant

level that depends on the loss function specified by the user. A good comparison of performance of the thresholding techniques is given in [79] and [87].

3.1.3 Cluster based thresholding in fMRI

Activated regions in fMRI tend to occur in clusters rather than individual pixels. Probabilistically put, a pixel is more likely to be considered active if its neighboring pixels are active. Individual thresholding techniques test each pixel independently of its neighbors without taking into consideration the spatial correlation between pixels. This results in a poor trade-off where there are either too many falsely activated voxels detected (too low threshold) or too many truly activated voxels missed (too high threshold). Spurious regions can be suppressed by processing pixels jointly, rather than individually, thereby increasing segmentation performance. The simplest way to overcome this is to smooth the images, often done implicitly in the pre-processing step. Software packages like the SPM package (developed in UCL, London), which is used for the analysis of neuroimaging data, provides an option for removing small activations with the size less than a certain (predefined) number of pixels. It is believed that such small activations are likely to be spurious and hence their removal would improve the detection performance. But this type of thresholding can remove the small truly activated regions as well.

Friston et al. [88] proposed a technique to determine a cluster-based threshold based on the Gaussian random field (GRF) theory. A cluster is declared activated based not only on its pixel intensities but also on its size. The technique considers a simple model of Gaussian signal with noise and smoothed by a point spread function. It then derives an exponential form for the cluster size as a function of smoothness, pixel intensity, and the dimensionality of the image. The intensity threshold is low for signals wider than the full width half maximum (FWHM) and high for signals sharper than the FWHM, i.e., smaller thresholds for larger clusters and larger thresholds for smaller clusters. Since the GRF theory assumes the image to be smooth, it is necessary to filter the image to ensure that it conforms to the GRF theory, which may suppress significant high-frequency spatial information.

Poline et al. [89] proposed a method that combines two tests for better detection of activation regions of all sizes. One of them is the peak height test that sets the high

intensity threshold and the other is the spatial extent test that requires two thresholds, a low intensity threshold and a size threshold. But this method is sensitive to signals with either high intensity or large spatial extent and works only when the intensity or the cluster size is significantly large. Bullmore et al. [90] proposed a permutation based method that combines intensity and cluster size. It computes cluster mass as the integration of the pixel intensities above the threshold for each cluster. The maximum cluster mass is then used as a test statistic in a permutation test in place of the maximum intensity or cluster size. The method by Hayasaka et al. [91] uses several combining functions so that the result gets the benefits of all of them. The combining functions used are the minimum p-value approach [89], cluster mass combining function [90] and the Fisher combining function [92]. Lazar et al. [92] provide a good survey of different functions for combining data in functional neuroimaging.

3.1.4 MRF-based detection of activations in fMRI

MRF segmentation aims to reduce the number of spurious regions in the output image, while maintaining low probability of miss detection. It takes into consideration the cluster spatial extent as well as the information about its pixel intensities. The knowledge about the second distribution is necessary in order to have a control over the type II error and MRF theory provides a systematic way to incorporate this information as well as region smoothness in the segmentation process [93, 94]. The spatial interactions between the neighboring pixels are modeled using MRF and are applied through a Bayesian framework. The segmentation is obtained by maximizing the a posteriori probability. Despite the computational complexity and difficulty in the proper selection of parameters, MRF model based segmentation is widely used for the segmentation of brain MR images since it is more robust to noise than the earlier models [94, 95]. In the case of binary segmentation (activated vs. non-activated regions) the partition can be efficiently estimated by finding a minimum cut on a specially defined graph [55]. Refer to Section 2.3.6 for details on MRF based segmentation.

Descombes et al. [96] considered MRFs in the context of signal and image restoration. They used a spatiotemporal MRF as a prior in a Bayesian framework to improve the restoration of fine structures and transitions around edges. The first two dimensions of

the model correspond to the spatial dimensions of an fMRI slice and the third one relates to time. The prior information is incorporated by specifying a Φ -model to preserve the transitions between activated and non-activated pixels in the spatial domain and between activated periods and baseline periods in the temporal domain. Φ -models are defined on pairwise interactions to preserve transitions. In addition to the spatiotemporal MRF for restoration, [97] uses another MRF for analyzing the hemodynamic function. Parameter maps are used to inhibit interactions between the activated areas and the background.

The method by Rajapakse et al. [98] used the likelihoods of the activation directly obtained from the SPMs to detect significant brain activation. The posterior probabilities are derived from the likelihoods given by intensities of the SPM and the prior probabilities given by clique potentials using the Bayes' theorem. Salli et al. [99] presented an iterative contextual analysis method based on MRFs. They used a contextual clustering algorithm to segment the SPM into active and non-active regions. The prior probability distribution for a pixel activation is specified conditionally on the activation pattern in the neighborhood of that pixel. The problem is formulated as a MAP estimation problem and iterated conditional modes (ICM) algorithm is used to determine the maximum. The algorithm was modified to perform hypothesis testing.

Cosman et al. [100] proposed a method for detection of activations in fMRI using a General Linear Model (GLM) with an Ising spatial prior. Ising model is one of the simplest MRF models where the log probability is proportional to the length of the boundary. Woolrich [101] developed a spatial mixture model with spatial regularization information provided by MRF. The discrete classification mixture models were approximated using the continuous weights mixture models to allow adaptive spatial regularization. Ou and Golland [102] used a Mean Field solver for MRF-based fMRI detection. Further, the method incorporates anatomical information to refine the activation priors.

3.2 Motivation

Detection of activation regions in fMRI undergoes difficulty when handling small regions. To illustrate this, we consider a binary grayscale model image with five activated regions (as shown in Figure 3.2(a)). The model image is corrupted by i.i.d. Gaussian noise

with zero mean and variance 0.1 (Figure 3.2(b)). We consider two different segmentation techniques (thresholding of pixel intensities and MRF segmentation) being applied on the noise corrupted model image. As can be seen from Figures 3.2(c) and 3.2(d), that due to noise both thresholding and MRF segmentation yield many spurious regions (the latter approach yields less), most of which can be suppressed by cluster size thresholding (see Figures 3.2(e) and 3.2(f)). Clearly, MRF segmentation produces better result than simple intensity thresholding, but not as good as thresholding combined with cluster size thresholding. However, with cluster size thresholding, the actual activations may also get affected, for example, when a single active region gets split into two or more regions, or simply becomes smaller than the threshold size (Figures 3.2(c) and 3.2(d)). Hence the motivation of our work is to find an optimal combination of cluster size thresholding and MRF segmentation.

3.3 Incorporating cluster size constraint

In this section, we explain cluster size thresholding in the context of MRF segmentation. The image is formulated as a graph G and MRF segmentation is solved by determining mincut on this graph. Refer to Section 2.2.5 and Section 2.3.6 for details on graph construction and MRF segmentation respectively. Consider a partition $\pi = (\pi_0, \pi_1)$ in G that minimizes the cost function

$$J(\pi) = - \sum_{\pi_0} \log p_0(I) - \sum_{\pi_1} \log p_1(I) + \beta L(\pi) \quad (3.4)$$

and another partition π' that is identical to π except for one removed (reclassified from foreground to background) small region π_s , whose size is below some predefined threshold. Clearly $J(\pi') > J(\pi)$, because π already achieves the smallest cost. The cost of π' shows an increase of $C(\pi_s) = J(\pi') - J(\pi)$, which is the cost of reclassification and also the cost of satisfying the cluster size constraint. In other words, satisfying cluster size constraint increases partition cost, but it is not clear whether there is any other way of satisfying the same size constraint at a smaller cost increase.

MRF segmentation with Ising model incorporates only pairwise pixel interactions and

Figure 3.2: (a) Model image (b) model image corrupted by noise (c) thresholding (d) MRF segmentation (e) cluster size thresholding of (c) (cluster size > 10) (f) cluster size thresholding of (d) (cluster size > 10)

thus cannot explicitly model the cluster size. A simple solution to the problem is to modify the energy minimization function of (3.4) as

$$J(\pi) = - \sum_{\pi_0} \log p_0(I) - \sum_{\pi_1} \log p_1(I) + \beta L(\pi) + h(size) \quad (3.5)$$

Figure 3.3: (a) Partition π (has the smallest cost) (b) Partition π' (non-optimal)

where

$$h(\text{size}) = \begin{cases} 0 & \text{if size of the smallest region in } \pi > T \\ \infty & \text{if size of the smallest region in } \pi \leq T \end{cases} \quad (3.6)$$

This formulation will penalize partitions where the smallest regions have size less than T . But, since the direct optimization of (3.5) is difficult, we propose a retrospective correction approach that is optimal under some assumptions that will be stated later.

We seek to minimize (3.5) by first minimizing the energy functional (3.4) and then modifying the solution so that (3.5) is minimized. It is straightforward to show that solution to (3.5) will be similar to that of (3.4). That is, if there is a region in (3.4) that has size above the cluster size threshold, exactly the same region would appear in the minimization of (3.5). If there was a small region (below cluster size threshold) in (3.4), then it would either disappear in (3.5) or alternatively, (3.5) will have a larger (than cluster size threshold) region that overlaps with this small region in (3.4). We finally show that this larger region can be found by examining the energy of all possible growths of the small region, which can be efficiently performed by modifying its boundary edges one-by-one.

3.4 Problem statement

We consider a binary grayscale model image with a background region (π_0) and k small disconnected foreground regions $\pi_{1i}, i = 1, \dots, k$. Let n_{1i} be the number of pixels in π_{1i} . Figure 3.2(a) shows an example of such an image for $k = 5$ and $n_{1i} = 30$, $n_{2i} = n_{3i} =$

$n_{4i} = n_{5i} = 15$. Let p_0 and p_1 be the probability density functions of pixel intensities in the foreground and background respectively. We assume that the model image is corrupted by i.i.d. Gaussian noise with zero mean and variance σ^2 (Figure 3.2(b)). Hence $p_0 \sim N(0, \sigma^2)$ and $p_1 \sim N(I_1, \sigma^2)$.

Let $\hat{\pi} = (\hat{\pi}_0, \hat{\pi}_{11}, \dots, \hat{\pi}_{1k})$ be the partition that minimizes (3.4). As illustrated in Section 3.2, the quality of $\hat{\pi}$ can be potentially improved by applying cluster size thresholding in the post-processing step. However, such retrospective correction does not necessarily have the smallest cost, as shown in Section 3.3. Our goal is to devise a method to minimize (3.4) subject to a constraint $\forall i, \hat{n}_{1i} > T$, where T is the cluster size threshold. This is also equivalent to minimizing (3.5).

3.5 Retrospective cluster size thresholding

Considering the two partitions as in Figure 3.3, we have shown that the application of cluster size thresholding (simple removal of small regions) to the result of MRF segmentation is not optimal, i.e., does not guarantee to minimize (3.5). In addition to simple removal of all small clusters, there could be other modifications of π , which have a smaller cost than that of simply removing the region. For example, we can try growing a region with size below the threshold until its size exceeds it, subject to constraint that the increase in the corresponding partition cost is smaller than that resulting from the region removal. If successful, the expanded region will satisfy the size constraint at a possibly smaller cost than simple region removal. Since there could be many ways to grow the region, in this section, we address the question how to find the smallest cost growth modification.

3.5.1 Finding a growth modification

Removing a small region π_s from the partition π increases its cost from $J(\pi)$ to $J(\pi) + C(\pi_s)$. Our goal is to find a way to increase the size of π_s above T such that the increase in the partition cost is minimized and is smaller than $C(\pi_s)$. To achieve this we propose the following approach. We modify the original graph G by increasing the weight of a boundary edge of π_s by $C(\pi_s) - \varepsilon$, where ε is an arbitrary small number (see Figure 3.4). We then run the min-cut algorithm on the altered graph G' to obtain π' . Lemma 1

summarizes several important observations about the properties of π' .

Figure 3.4: A small region and boundary edge

Lemma 1. *Let π' be the minimum cost partition on graph G' and Δ be the set of all pixels that have different classification in π' compared to π . Then*

1. $J'(\pi') < J(\pi) + C(\pi_s)$
2. Δ is a connected region
3. If $\Delta \neq \emptyset$, then it contains one of the pixels connected to π_s by the modified boundary edge.

Proof. If the cost $J'(\pi')$ of the new partition π' , is greater than $J(\pi) + C(\pi_s)$, then $J'(\pi') > J(\pi) + C(\pi_s) - \epsilon = J'(\pi)$. This contradicts the fact that π' is a minimum cost partition on G' . Therefore, $J'(\pi') < J(\pi) + C(\pi_s)$. Thus statement 1 is true. If Δ consists of two disconnected regions, then removing one of them (reclassifying one from background back to foreground) which is not adjacent to the modified edge, would decrease the cost of the partition, contradicting the optimality of π' . Thus statement 2 is true.

To prove statement 3, let us assume that the growth modification Δ does not include the pixel connected by the modified boundary edge (see Figure 3.5(b)). Then, the cost of the new partition is $J'(\pi') = J(\pi) + C(\pi_s) - \epsilon + J(\Delta)$, which again contradicts the optimality of π' . Thus statement 3 is true. ■

Shown in Figure 3.5 is the illustration of the results of Lemma 1. In summary, the proposed edge weight modification can lead to either growth or shrinking of π_s , or no

Figure 3.5: Possible and impossible changes to the region's shape after modifying boundary edge weight (a) Original region and boundary edge e (b) Region Δ doesn't contain e (impossible) (c) Region Δ consists of two disconnected parts (impossible) (d) Correct modification

change in the partition structure. This growth must always occur from the side of the modified edge. If π_s grows beyond T , then such modification achieves smaller cost than removing π_s . However, Lemma 1 does not guarantee that the found growth modification, if any, has the smallest possible cost. For example, there can be other modifications involving growth through another boundary edge that achieve even smaller cost.

3.5.2 Finding the smallest cost growth modification

To find the best growth modification (the growth modification that satisfies the cluster size constraint and has the smallest cost) we can perform a series of corrections, where in each correction we modify the weight of one boundary edge of π_s , and then choose the modification with the smallest cost. According to Lemma 2, this guarantees that

this approach will find the smallest cost modification, provided it exists and has the cost smaller than $C(\pi_s)$.

Lemma 2. *Let Δ^* be the smallest cost growth modification of region π_s , so that $C(\Delta^*) < C(\pi_s)$. Then Δ^* can be found by modifying each boundary edge of π_s , one by one, and choosing the modification with the smallest cost.*

Proof. Since Δ^* is a growth modification, it consists of foreground pixels connected to π_s . Let e be a boundary edge of π_s connecting a pixel in π_s with a pixel in Δ . Then it is straightforward to show that increasing the weight of this edge by $C(\pi_s) - \varepsilon$ will lead to a change in partition that coincides with Δ^* . ■

Let π'_s be the found grown region after trying each boundary edge. It is possible that the size of π'_s is still smaller than the threshold. In this case we can repeat the whole procedure on π'_s to determine whether it can be grown to a larger size. This will involve increasing the weight of each boundary edge of π'_s by $C(\pi'_s) - \varepsilon$ and choosing a growth modification with the smallest cost. Iterations can stop after the region size exceeds the threshold value or when no further changes happen to region shape.

3.6 Experimental results

We have conducted a performance comparison of various segmentation techniques on the model image (Figure 3.2 (b)) corrupted by i.i.d. Gaussian noise with $\sigma = 0.1$. Here we included smoothing with a Gaussian kernel followed by thresholding (with and without cluster size constraint), MRF segmentation (with and without cluster size constraint) and the proposed retrospective correction approach. In our experiments, we use the minimum cut algorithm implementation by Boykov and Kolmogorov [59].

Figure 3.6 shows the result of the proposed approach on the model image. The true activated region obtained by MRF segmentation got split into two (highlighted in yellow in Figure 3.6(c)) and would have been removed by cluster size thresholding because of its small size. The application of the retrospective correction approach grows the region beyond the cluster size (highlighted in yellow in Figure 3.6(d)) and preserves the true activation. One of the true activation was not able to be grown beyond the cluster size

Figure 3.6: (a) Model image, (b) model image corrupted with noise, (c) MRF segmentation of (b), (d) MRF segmentation with retrospective cluster size thresholding (cluster size > 7) - grown region highlighted in yellow

and is therefore eliminated. In the retrospective correction approach, since the original partition contains several small regions, the modifications of edge weights are done in parallel to save the computation time. Nevertheless, the whole iterative procedure can be time consuming as it requires many repetitions of min-cut segmentation (exact number depends on the boundary length of the regions modified). However, since T is usually quite small, so are the boundary lengths of the regions modified. The algorithm can also be made faster through the use of dynamic graph cuts [103].

For performance evaluation, we plotted the false alarm rate (type I error) against the miss detection rate (type II error), see Figure 3.7. Each experiment was repeated 500 times to achieve the averaging effect. The cluster size threshold was set to 10. According to these results, cluster size thresholding can substantially improve the segmentation performance. In case of thresholding of individual pixel intensities it reduces the miss detection error from

Figure 3.7: Performance comparison of the segmentation techniques

0.7 to 0.5, when the false alarm rate is set to 0.1%. Improvements for MRF segmentation are more modest - reduction in miss detection error is 0.05 – 0.1 on average, which can be explained by the non-optimality of simple region removal in the MRF context. The proposed approach resolves this problem and leads to the best overall performance. The reduced false alarm rate and high miss detection rate are due to high number of background pixels and very low number of foreground pixels (activated pixels).

We have also investigated the performance of our method on the statistical parametric map (SPM) obtained from a real fMRI experiment which seeks to find the regions of activation when the subject either attends or ignores a stimulus. The result of MRF segmentation and the retrospective correction approach is shown in Figure 3.8. Figures 3.8(a) and 3.8(b) show the SPM and the corresponding MR image slice respectively. Smoothing of the SPM followed by intensity thresholding results in a number of spurious regions (Figure 3.8(c)) which could be removed by cluster size thresholding (Figure 3.8(d)). The result

Figure 3.8: (a) Statistical parametric map (SPM), (b) corresponding structural MR image slice, (c) thresholding of the SPM, (d) cluster size thresholding of (c) (cluster size > 10), (e) thresholding result (d) overlaid on (b), (f) MRF segmentation of the SPM, (g) cluster size thresholding of (f) (cluster size > 10), (h) MRF segmentation with retrospective cluster size thresholding (cluster size > 10) - one small region grown beyond cluster size highlighted in red, (i) result of retrospective correction approach (h) overlaid on (b)

is overlaid on the structural MR image slice to map the activations with the structural data and also to distinguish the true activations from the spurious ones. True activations occur in the gray matter and the spurious ones occur in the white matter and regions outside the brain. MRF segmentation produces smoother activations and devoid of spurious regions (Figure 3.8(f)). The application of cluster size thresholding to the output of MRF segmentation results in the elimination of one small activation (Figure 3.8(g)). This result is equivalent to the result of thresholding followed by cluster size thresholding (Figure 3.8(d)). Our proposed retrospective correction approach outperforms the other methods in preserving the small activation (highlighted in red) by growing it beyond the cluster size threshold.

3.7 Discussion

We proposed a novel approach to improve the detection of small regions in fMRI. The idea is based on cluster size thresholding, which aims to remove spurious regions without affecting the truly activated ones. We have shown that, in the context of MRF segmentation, simple removal of small regions after the segmentation is not optimal. The developed retrospective correction finds the best modification (removal or growth) and achieves superior performance to both thresholding and MRF segmentation.

The current choice of cluster size threshold was made ad-hoc. Clearly, smaller values will be worse for removing the noise but better for preserving the activated regions. In our future work, we plan to investigate the choice of the optimal region size threshold and also consider practical aspects of extending the developed theory to the three dimensions.

Chapter 4

Removal of narrow connections: Critical analysis of existing approaches

Image segmentation partitions the image into several regions so that the resultant regions of interest have pixels of similar properties and convey meaningful information. The segmentation problem is often made complicated because of the lack of a clear boundary between the regions, i.e., when the regions to be isolated possess similar properties or are connected to each other. Such connections are very common in the case of medical images (heart ventricles [104–106], abdominal fat [107], brain [108–110], liver [111], and microscopic cells [112]) because of the noise, artifacts, and overall image geometry. Figure 4.1 shows some examples.

Though extensive work has been done to devise methods to break connections, much less attention has been paid to compare the performance of these methods. This gives the motivation to analyze each of the methods critically and compare their performances.

4.1 Narrow connection removal: Problem definition

Segmentation of medical images often involves solving the problem of removing connections between regions with similar properties. The definition of narrowness of the connection

Figure 4.1: Examples of narrow connections in medical images: row 1-abdominal fat segmentation, row 2-brain segmentation, row 3-liver segmentation

is subjective and usually depends on application. Even though the original images are usually grayscale, the problem of narrow connection removal is most commonly defined on binary images. This means that grayscale images are first converted to binary using, for example, simple thresholding of pixel intensities, although more complex conversion procedures are also possible. Figure 4.1 shows the binary images obtained from their corresponding original grayscale images by intensity thresholding, see narrow connections between the outer fat layer and the inner fat layer in the abdominal fat image, brain tissues connected narrowly connected with the non-brain tissues in brain image and liver connected to other tissues by narrow connections. The binarization (foreground F and

background B) provides initial elimination of the most obvious unwanted regions and often leads to creation of narrow connections between the regions of interest and the rest of the image, thus motivating our problem of narrow connection removal.

It is important to note that it is possible to devise segmentation methods that bypass binary mask creation and seek to find regions that are already disconnected from the others with similar properties. These techniques are usually more complex and some of them will be discussed in the context of skull stripping later in Chapter 6. However, using simple segmentation techniques like thresholding or region growing is very likely to result in segmentation that contains narrowly connected “extra” regions. This necessitates an additional step in the segmentation procedure to eliminate the narrow connections and separate the regions.

4.2 Review of existing methods

In this section, we review some of the methods used for narrow connection removal, such as morphological processing [109,110], distance transform followed by the watershed algorithm [113] and isoperimetric algorithm [75,76].

4.2.1 Morphological processing

Morphological processing (MORPH) is the simplest and the most common approach to break narrow connections [109,110]. Refer to Section 2.3.3 for details on morphological processing. Morphological operations can sufficiently break any narrow connection by employing suitable sized/shaped structuring elements. The procedure consists of erosion, selection of the largest connected component and dilation. The disk structuring element is the most commonly used structuring element for narrow connection removal. Figure 4.2 illustrates how morphological processing can be used to cut connections on a sample image.

Erosion operation shrinks the object and cuts narrow connections and dilation expands the object. The combination of these two operations with dilation followed by erosion is termed as opening operation. The side effects of morphological processing are distortion of the original shape of the object (smoothing of object’s contour) and elimination of thin

Figure 4.2: Morphological processing (a) sample image (b) erosion of (a) by disk structuring element, size 25 (c) Selection of the largest connected component (d) dilation of (c) by disk structuring element, size 25

protrusions. The maximum width of the connection that can be removed by morphological processing is controlled by the size and shape of the structuring element.

4.2.2 Distance transform followed by watershed algorithm

Another popular approach used to remove narrow connections is the distance transform followed by watershed algorithm (DWAT) [113]. Refer to Section 2.2.3 for details on watershed algorithm. The working of DWAT is illustrated in Figure 4.3. Distance transform is applied to the binary image, resulting in a grayscale image where the intensities of each pixel in the foreground represent the shortest distance of that pixel to the nearest boundary [114]. The distance transform value $D(x)$ for a foreground pixel x can be defined as

$$D(x) = \min_{y \in B} (d(x, y)) \quad (4.1)$$

where y is any pixel in the background B and d is any distance measure. The distance transform values are minimum at the boundary and larger towards the center of the foreground region. The values are smaller at the narrow connections and largest at the center, see Figure 4.3(b).

The watershed transform acts on the inverted distance transformed image. The inversion is done to create valleys corresponding to the foreground region. The watershed transform interprets the intensities (distance transform values) as height information and the high and low values as peaks and valleys, see Figure 4.3(d). Two regions are split from each other if they are separated by a hill. The higher value at the narrow connection cre-

Figure 4.3: Distance transform followed by watershed algorithm (a) sample image (b) distance transform of (a) (c) distance transform of (a) - inverted (d) plot of (c) (e) watershed transform of (c)

ates a hill between two valleys and thus gets separated on the application of the watershed transform.

4.2.3 Isoperimetric segmentation

The most recent approach proposed to cut narrow connections is isoperimetric segmentation (ISO) [76]. This approach works by setting the problem within a graph theoretic context. Here, the foreground region of the image is represented as a weighted graph $G = (V, E, W)$, where V , E , and W are sets of vertices, edges, and edge weights respectively.

The goal of ISO is to partition the vertices of the graph into two connected sets R_1 and R_2 so that isoperimetric ratio is minimized. The isoperimetric ratio is defined as a ratio of the cut value, i.e., the sum of edge weights along the boundary between the partition $\pi = (\pi_1, \pi_2)$, $|\partial\pi_{12}| = \sum_{i \in \pi_1} \sum_{j \in \pi_2} w_{i,j}$, and the area (the number of vertices) of the smaller partition element, which we assume is π_2 , where π_1 and π_2 are two partitions.

The final partition is the one that achieves the smallest isoperimetric ratio:

$$c = \inf_{\pi_1} \frac{|\partial\pi_{12}|}{Area(\pi_2)}, \quad Area(\pi_2) \leq \frac{1}{2}Area(total) \quad (4.2)$$

Figure 4.4: Cutting narrow connections with ISO

In the case of removal of narrow connections, ISO algorithm works as follows. We start with an initial mask, consisting of foreground region F and background region B , see Figure 4.4. The ISO algorithm defines the graph only on the region F and then examines all possible ways to cut F into π_1 and π_2 , choosing the cut with the smallest isoperimetric ratio. For example, between the two cuts shown in the Figure 4.4, $Area(\pi_2) < Area(\pi'_2)$. Assuming that all edge weights are equal to 1, $|\partial\pi_{12}| = 6 > |\partial\pi'_{12}| = 3$, because six edges need to be cut to separate π_1 from π_2 , but only three are cut in separating π'_1 and π'_2 . Hence $c_1 = \frac{|\partial\pi_{12}|}{Area(\pi_2)} > c_2 = \frac{|\partial\pi'_{12}|}{Area(\pi'_2)}$ and the second cut will be preferred over the first.

This segmentation procedure is performed iteratively with a cut having the smallest isoperimetric ratio determined at every step. Iterations terminate when this isoperimetric ratio is less than a predefined threshold c_T . The steps of the procedure are detailed in Algorithm 1.

As the edge weights of G do not all have to be equal to 1, the cut value will depend not only on the number of edges cut, but also on the weight assigned to each edge. Hence, we refer to $|\partial\pi_{12}|$ as the weighted boundary length between π_1 and π_2 , highlighting the dependence of the cut value on both the number and weights of the edges.

Algorithm 1 Isoperimetric segmentation (ISO)

-
- 1: Isoperimetric segmentation(initial binary mask (F, B) , c_T , seed point)
 - 2: Construct a graph G only on F
 - 3: Determine a cut with the minimum isoperimetric ratio c to break G into two partitions π_1 (containing the seed) and π_2
 - 4: Compute c of the cut using (4.2)
 - 5: **repeat**
 - 6: $F \leftarrow \pi_1$
 - 7: Construct a new graph G on F
 - 8: Determine a cut with the minimum isoperimetric ratio c to break G into π_1 and π_2
 - 9: Compute c of the cut using (4.2)
 - 10: **until** c is less than c_T
-

4.3 Performance comparison of existing methods

In order to set a basis for theoretical evaluation of performance of the methods described in Section 4.2, we consider a model image containing two arbitrarily shaped regions R_1 and R_2 connected to each other by a bridge of width w . The goal is to estimate region R_1 , which is equivalent to cutting the foreground along the bridge that separates R_1 from R_2 , see Figure 4.5(a). A “good” algorithm should be able to separate the regions along the bridge.

It is expected that narrower connections (smaller w) would be easier to cut than the wider connections. Hence, we evaluate the performance of the methods based on the maximum width of the connection that can be broken by each of them. If a method involves setting any parameter, i.e., size of structural element in morphological processing or isoperimetric ratio threshold in ISO, it is important to differentiate the following two cases:

1. *Performance under fixed parameter value.* Since the value of the parameter is usually unknown, it has to be fixed at a specific value that is decided based on some prior knowledge about the shapes and sizes of the regions involved and the width of the connection. Performance under this condition probably best reflects the real world performance.
2. *Performance under optimized parameter value.* This case assumes possessing complete knowledge of region shapes, sizes, and connection width. Hence, the best

parameter value can be chosen to accomplish successful narrow connection removal. This case provides the theoretical bound on algorithm's performance.

We also expect that the performance will depend not only on the width of the connection, but also on the shapes and sizes of R_1 , R_2 and w in Figure 4.5(a), which makes the performance evaluation difficult. Instead we aim to capture the pattern of performance by looking at a set of two model images consisting of simple shapes, as in Figure 4.5(b) and Figure 4.5(c), while preserving arbitrary sizes.

Figure 4.5: Model images (a) Arbitrarily shaped regions (b) Two circular regions (c) Circle-rectangle

The first of the two images (Figure 4.5(b)) contains two circular regions (R_1 and R_2) of radius r_1 and r_2 respectively. R_1 and R_2 are connected to each other to form a bridge of width w . Without loss of generality, we assume that $r_1 > r_2$. We also assume that the connection width can be as large as the diameter of R_2 , $w < 2r_2$. The second image (Figure 4.5(c)) consists of a circular region R_1 connected to a rectangular region R_2 of dimensions $l \times b$. In this case, we assume that the diameter of R_1 is larger than the length l and $w < l$. The choice of the shapes was made so that they approximately resemble the real world cases discussed in the introduction. For example, the second model image may approximate either dura adjacent to GM surface or external fat adjacent to internal fat.

Let S_1 , S_2 be the areas and P_1 , P_2 be the perimeters of R_1 and R_2 respectively and d_1 and d_2 be their respective region depths (the largest distance transform value within the region).

4.3.1 Morphological processing

Removing narrow connections using MORPH depends on the shape and size of the structuring element used. In order to sever the connection, the structuring element should be larger than the connection width. Assuming the use of a disk structuring element, w should be less than twice the region depth d_s of the structuring element, $w < 2d_s$. If the chosen structuring element satisfies this condition, then MORPH can be used to break almost any connection.

Since the performance of MORPH depends only on d_s , we compute the value of two performance metrics for the following two cases of parameter d_s :

1. **Fixed value of d_s .** With the value of d_s fixed, successful outcome occurs only when d_s satisfies the condition $w < 2d_s$.
2. **Optimized value of d_s .** In this case, since we know the connection width w , d_s can be chosen to be larger than w so that the bridge is broken. However, there is a maximum limit to the value of d_s (d_{max}) beyond which MORPH may exceed the allowable erosion limit thereby damaging the shape of the desired output region R_1 .

MORPH can cut connections as wide as $2r_2$ ($w < 2r_2$) for the two circles image and l ($w < l$) for circle-rectangle image, provided suitable sized structuring element is chosen.

Figure 4.6: MORPH using disk structuring element of different sizes (a) sample image (b) size 25 - little smoothening (c) size 50 - original shape damaged

Limitations

The main side effect of MORPH is smoothening of the contour of the object and elimination of thin protrusions. This might not be a problem when the connections are very narrow and the size of structural element is small (Figure 4.6(b)), but large structuring elements are likely to damage the object shape (Figure 4.6(c)). Overall, we can conclude that MORPH is suitable for narrow connection removal only when connections are sufficiently narrow.

4.3.2 Distance transform followed by watershed transform

Unlike MORPH, DWAT does not depend on any parameter. It only depends on the distance transform values interpreted as peaks and valleys. In the case of model images, there will be two valleys with depths (d_1 and d_2 for R_1 and R_2 respectively) and a peak corresponding to the connection, with height equal to half the width w of the connection. In order for the regions to be separated, the peak height should be less than the depth of the valley d_2 .

$$\frac{w}{2} < d_2 \quad \Rightarrow \quad w < 2d_2 \quad (4.3)$$

- **Two circles.** Here, d_2 is equal to r_2 and DWAT can cut connections as wide as $2r_2$ ($w < 2r_2$). This means DWAT can cut connections of any width for this case.
- **Circle-rectangle.** Here, d_2 is equal to $\frac{b}{2}$ and this limits the maximum width of the connection that can be broken to b which could be very small compared to l .

Limitations

The success of DWAT is determined by condition $w < 2d_2$. While this does not present a limiting factor for the two-circles case, it is a major limitation for the circle-rectangle where the breadth b is usually much less compared to its l . Even more serious drawback of DWAT can be observed only on model image in Figure 4.5(a) due to the complex shape appearance of region R_1 , see illustration in Figure 4.7. The inversion of distance transform results in undesired peaks within R_1 and watershed transform produces multiple regions. Real world images are unlikely to have such simple shapes as a circle or rectangle, and

Figure 4.7: DWAT - oversegmentation (a) original image (b) distance transform of (a) (c) distance transform of (a) - inverted (d) watershed transform of (c)

furthermore contain noise, which would make the oversegmentation problem even more serious, defeating the whole benefit of region separation.

4.3.3 Isoperimetric segmentation

The functioning of ISO can be summarized as follows:

1. Determine a cut to partition G with the smallest c .
2. If $c < c_T$, then select the largest partition element.
3. Repeat steps 1 and 2 until $c < c_T$.

In our example images, let $\pi = (\pi_1, \pi_2)$ be the partition of G , where the size of π_2 is smaller than the size of π_1 , i.e., $S(\pi_2) < S(\pi_1)$. Then, the ratio of weighted boundary length π_{12} between π_1 and π_2 to the area of π_2 should be minimized. For now, we assume that all the weights of the edges are equal to 1, hence

$$c = \min \frac{|\partial\pi_{12}|}{S(\pi_2)} \quad (4.4)$$

The failure of ISO can be decided based on the location of the cut. For example, if the cut occurs within R_1 in any iteration, then the results can be declared as a failure. But cutting within the smaller region is not yet a failure because the cut can still be improved in the next iteration. The following two lemmas are helpful in the performance analysis of ISO.

Lemma 3. *The isoperimetric ratio of a cut within a circular region of radius r is given by*

$$c = \frac{4 \sin \frac{\theta}{2}}{r(\theta - \sin \theta)}$$

where $0 < \theta < \pi$ is the angle subtended by the cut (chord) at the center of the circle. The cut with the smallest isoperimetric ratio occurs at $\theta = \pi$, i.e., along the diameter of the circle.

Proof. Refer to Appendix B. ■

Lemma 4. *Consider a circular region R_1 of radius r_1 connected to another circular region R_2 of radius r_2 by a bridge of width w . Then, (i) if $r_2 < r_1$ or $S_2 < \pi r_1^2$, cut with the minimum isoperimetric ratio occurs either along the connection or along the diameter of R_1 (ii) if $r_2 > r_1$ or $S_2 > \pi r_1^2$, no cut with the minimum isoperimetric ratio can occur within R_1*

Proof. Refer to Appendix B. ■

Lemma 3 states that the cut within an isolated circular region always occurs along its diameter. If there are two circular regions connected together, it is proved in Lemma 4 that the cut to separate those two regions can occur only along the connection or inside the larger region and no cuts can happen inside the smaller region. However, if the attached region is rectangular, the cut can still occur within the smaller region. Using the conclusions from Lemma 3 and Lemma 4, we can determine the expressions and values for the isoperimetric ratio and the conditions under which the correct cut can be obtained.

The ISO procedure is repeated again if the smallest isoperimetric ratio of the cut obtained in the first step is still greater than c_T or if the desired segmentation result is not obtained. The segmented portion obtained in the previous iteration is considered as the

new foreground region in the current step. The iterations terminate when the criterion ‘isoperimetric ratio of the cut should be less than c_T ’ is not satisfied. Thus the selection of the right value of c_T becomes important to control the end of the iterative procedure. The value of c_T should be properly selected so that the segmentation results in the correct cut. Very high and very low values of c_T will make ISO select undesired cuts. c_T should be just greater than the isoperimetric ratio of the correct cut.

Model image 1 - two circles

Consider the model image consisting of two circular regions as in Figure 4.5(b). Here $S_1 > S_2$ and from Lemma 4, it follows that the cut with minimum isoperimetric ratio can not occur within R_2 . The two possible cuts for the two circles image are shown in Figure 4.8. Following is the working of ISO on the two circles image.

Figure 4.8: Possible cuts for two-circles object

- *First iteration.* There are two possible cuts, through diameter of the larger region $c_1 = \frac{4}{\pi r_1}$ or through the bridge $c_2 = \frac{w}{\pi r_2^2}$. If $c_1 < c_2$, this means failure of the algorithm at the first iteration. If $c_1 > c_2$ (no initial failure), but $c_2 > c_T$ (threshold is too low), then no cuts will be made and it is a failure. So to successfully exit the first iteration, we need two conditions, $c_1 > c_2$ and $c_2 < c_T$. To satisfy the first condition, we need the width to be sufficiently small, $w < \frac{4r_2^2}{r_1}$. To satisfy the second condition, we need $\frac{w}{\pi r_2^2} < c_T$.
- *Second iteration.* In this case we are left only with a single circle and hence only one cut is possible with $c_1 = \frac{4}{\pi r_1}$. The failure will occur if this cut is permitted, so for success we need $c_1 = \frac{4}{\pi r_1} > c_T$.

So, the two conditions for success are $w < \frac{4r_2^2}{r_1}$ and $\frac{w}{\pi r_2^2} < c_T < \frac{4}{\pi r_1}$. Now let's consider what happens when c_T is fixed or optimized. First of all, for both of these cases $w < \frac{4r_2^2}{r_1}$ remains the upper bound on the widest connection that can be cut. However, improper choice of c_T can make the performance worse.

- *Fixed c_T .* If c_T is fixed and happens to fall in the range $\frac{w}{\pi r_2^2} < c_T < \frac{4}{\pi r_1}$, then the widest connection that can be cut is $w < \frac{4r_2^2}{r_1}$. If out of caution c_T is chosen too low, this will decrease the maximum width of the connection to $w < c_T \pi r_2^2$. If c_T is chosen too high, above $\frac{4}{\pi r_1}$, the algorithm will fail.
- *Optimized c_T .* In this case c_T can be chosen anywhere in the range $\frac{w}{\pi r_2^2} < c_T < \frac{4}{\pi r_1}$ and the maximum width that can be cut is $w < \frac{4r_2^2}{r_1}$.

Model image 2 - circle-rectangle

In the case of a rectangular region attached to the circular region, apart from the possibility of cuts happening along the diameter of R_1 and along the bridge, the cut can occur within the rectangular region if $w > 2b$. Figure 4.9 shows the three possible occurrences of the cut. We therefore shall consider the two cases, $w < 2b$ and $w > 2b$.

Figure 4.9: Possible cuts for circle-rectangle object

1. $w < 2b$.

- *First iteration.* Two possible cuts can occur, one along the diameter of R_1 with $c_1 = \frac{4}{\pi r_1}$ or through the connection with $c_2 = \frac{w}{\pi b}$. If $c_1 < c_2$, then it is a failure in the first iteration. $c_1 > c_2$ has no initial failure, but a low threshold with

$c_2 > c_T$ will not make any cuts and it is a failure. For the success of the first iteration, two conditions $c_1 > c_2$ and $c_2 < c_T$ have to be satisfied. This gives the expression for the width as $w < \frac{4lb}{\pi r_1}$ and c_T as $c_T > \frac{w}{lb}$.

- *Second iteration.* In the second iteration, only a single circle R_1 remains and the only possible cut is through its diameter with $c_1 = \frac{4}{\pi r_1}$. In order to prevent the cut, c_T should be $c_1 = \frac{4}{\pi r_1} > c_T$.

2. $w > 2b$.

- *First iteration.* When we have a wider connection, the two possible occurrences of cuts are through the diameter of R_1 with $c_1 = \frac{4}{\pi r_1}$ or through R_2 with $c_2 = \frac{2}{l-w}$. If $c_1 < c_2$, it is a failure. But $c_1 > c_2$ is still not a failure until $c_2 < c_T$. Therefore, it follows that two conditions $c_1 > c_2$ and $c_2 < c_T$ should be satisfied for successful first iteration. That gives $w < l - \frac{\pi r_1}{2}$ and $c_T > \frac{2}{l-w}$.
- *Second iteration.* In the second iteration, we are left with the circle attached to a small portion of the rectangle. Now, there are two possible cuts, through the diameter of the circle with $c_1 = \frac{4}{\pi r_1}$ or through the bridge with $c_2 = \frac{1}{b}$. Usually b is smaller than r_1 and hence $c_1 < c_2$. Therefore, for larger widths $w > 2b$, the algorithm produces wrong segmentation.

When $w < 2b$, the conditions for success are $w < \frac{4lb}{\pi r_1}$ and $\frac{w}{lb} < c_T < \frac{4}{\pi r_1}$ and when $w > 2b$, the method fails to produce correct segmentation. Therefore, we disregard the condition $w > 2b$ and consider only cases when $w < 2b$. We now explain the effect c_T has on the expression for width.

- *Fixed value of c_T .* In this case, the maximum width that can be severed is $w < \frac{4lb}{\pi r_1}$. It also has to be noted that the method fails when $w > 2b$. The maximum width that can be cut is $w < \min\left(2b, \frac{4lb}{\pi r_1}\right)$. However, selecting a low c_T gives $w < c_T lb$ and high c_T results in failure.
- *Optimized value of c_T .* In this case, any value of c_T within the range $\frac{w}{lb} < c_T < \frac{4}{\pi r_1}$ can be chosen and the maximum width that can be cut is $w < \min\left(2b, \frac{4lb}{\pi r_1}\right)$.

Limitations

The functioning of ISO is dependent on the value of c_T , which can be difficult to predict. Successful removal of narrow connections can be expected to require some fine tuning of c_T . However, fine tuning does not always guarantee successful removal of the bridge if the width is large. The severity of this limitation can be partially reduced by the use of different edge weight assignment which will be discussed in Section 5.1.

ISO has limitations that makes its use very limited in real world applications. The main limitation of ISO is its approximate optimization procedure, where binary segmentation labels (0 for background and 1 for foreground) are allowed to take arbitrary real values. This real-valued solution is then binarized by thresholding. While this procedure usually works well, its accuracy is somewhat difficult to predict. Another important factor that limits the use of ISO is the dependence of segmentation on the position of the seed point/ground point. Grounding a point corresponds to removing the row and column pertaining to the seed point in the matrix. This affects the solution of the optimization problem. Though this problem has been discussed earlier [77], it was illustrated using only a simple example image (two circles image) and has not been treated in sufficient detail. It has been shown that when the seed point is placed close to the bottleneck, i.e., in very close proximity to the bridge connecting the two regions, the resulting segmentation will differ from what is expected. We executed ISO on the circle-rectangle image several times varying the position of the seed point. We found that the segmentation result is affected not only when the seed point is close to the connection, but also when it is far (Figure 4.10).

Though ISO may have the advantage of the need to specify only one point as the seed, the ambiguity in finding its correct position poses a serious drawback and may make it difficult to use in practical applications. This drawback motivated us to develop a new solution to the cutting problem that uses graph cuts (GCUT) which has a superior optimization procedure that achieves globally optimal solution in polynomial time (Chapter 5).

Figure 4.10: Effect of position of seed point on ISO

4.4 Summary

Table 4.1 provides an overview of the maximum width of the connection that each of the methods can break. MORPH performs better than other methods when only the maximum width of the connection is considered for comparing the methods. It can cut almost any connection with a suitable choice of structuring element. The side effect is that the region gets smoothed and in some worst cases, the shape gets completely damaged. DWAT works very well on compact shaped images. For the two circles image, the maximum width of the bridge can go till the diameter of R_2 . But for the circle-rectangle object, only very narrow connections, i.e., with widths less than the width of the rectangle can be broken. Moreover, for arbitrary shaped regions, DWAT is likely to result in oversegmentation, which defeats the benefits of connection cutting. ISO's performance is worse than MORPH and DWAT in the case of circle-circle object and for circle-rectangle image, it can cut widths up to $2b$ compared to only b for DWAT.

Another aspect that is used to compare the methods is the need to set the value of any external parameters that they might require for successful narrow connection removal.

Table 4.1: Performance comparison: Maximum width of the connection that can be cut by different methods

Method	Two circles	Circle-rectangle
MORPH	$w < 2d_s$	$w < 2d_s$
DWAT	$w < 2r_2$	$w < b$
ISO (fixed c_T)	$w < \min\left(\frac{4r_2^2}{r_1}, c_T\pi r_2^2\right)$ when $c_T < \frac{4}{\pi r_1}$	$w < \min\left(2b, \frac{4lb}{\pi r_1}, c_T lb\right)$ when $c_T < \frac{4}{\pi r_1}$
ISO (optimized c_T)	$w < \frac{4r_2^2}{r_1}$ when $c_T < \frac{4}{\pi r_1}$	$w < \min\left(2b, \frac{4lb}{\pi r_1}\right)$ when $c_T < \frac{4}{\pi r_1}$

DWAT is the most advantageous of the three, since it is free of any external parameters. It does not require any prior knowledge about the size and shape of the region. Successful functioning of ISO depends on the value of c_T . Determining the right value of c_T is a difficult task and fixing it might not always produce satisfactory results. Obtaining the optimum value of c_T might require some prior knowledge about the width of the connection as well as the shape and size of the regions. Therefore, it has to be fine tuned depending on the application. MORPH's performance depends on the suitable choice of structuring element used. Another limitation of MORPH is that it damages the shape of the region regardless of the shape and size of the structuring element used. The effects are more pronounced if a completely unsuitable structuring element is chosen. Other drawbacks of ISO are its approximate optimization procedure with difficulty to predict accuracy and its dependence on the location of the seed point. Despite these limitations, ISO can be preferred over the others due to its excellent performance when assumptions on region size and shape are satisfied and c_T is properly chosen. The drawbacks of ISO can be at least partially alleviated by changing the edge weights of the graph, which will be discussed in Section 5.1.

Chapter 5

Distance transform based edge weight assignment for graph theoretic removal of narrow connections

After reviewing the existing techniques for narrow connection removal in Chapter 4, we concluded that ISO is superior to MORPH and DWAT, despite several limitations. Some of these limitations can be overcome by allowing the graph edge weights to have non-constant weights. In this chapter, we propose a modified version of ISO, we call generalized ISO (Gen ISO), which assigns weights based on distance transform. ISO can be considered as a special case of Gen ISO when all the edge weights are equal to one. Despite improvement in performance that follows the new weight assignment, the isoperimetric algorithm still suffers from limitations such as sensitivity to location of seed point and the approximate optimization procedure. In order to overcome these remaining drawbacks, we propose a novel approach based on graph cuts algorithm, which uses similar weight assignment to that of Gen ISO. We show that graph cuts approach provides an alternative to ISO with performance equivalent to or better than ISO, while lacking its drawbacks.

Both generalized ISO and the new graph cuts algorithm are based on a new edge weight assignment which makes use of the distance transform of the image. The distance

transform gives the distance of every foreground pixel to the nearest boundary. The pixels which have low distance transform values are either located closer to the boundary or inside narrow connections. The distance transform value $D(x)$ for a foreground pixel x can be defined as

$$D(x) = \min_{y \in B} (d(x, y)) \quad (5.1)$$

where y is any pixel in the background B and d is any distance measure. We define the new edge weight assignment as

$$w_{ij} = \max_{v_i, v_j \in \pi_1} (D(v_i), D(v_j)) \quad (5.2)$$

where $D(v_i)$ and $D(v_j)$ are the distance transform values of the foreground vertices v_i and v_j respectively as defined in (4.1). In the following sections, we show the benefits of the new weight assignment for graph theoretic narrow connection removal. These methods make use of distance transform based weight assignment. In the case of ISO, this new weight assignment makes the algorithm less sensitive to the choice of isoperimetric constant, whereas in graph cuts, it forms the basis of a novel fully automated algorithm for narrow connection removal.

5.1 Generalized ISO

ISO as in [76] used weights equal to one for all the edges in the graph. In general, the weights of the edges can be of any value and need not be equal. The expression for the width of the connection between arbitrary shaped regions for ISO with all edge weights equal to one is given in (4.2). In this section, we describe the generalized ISO algorithm and discuss how the weights can be defined more general.

5.1.1 Polynomial case

We assume all weights to be non-negative and real and try to formulate a generalized weight function that can best describe the edge weights of the graph for the ISO algorithm. Recall that ISO is based on minimization of isoperimetric ratio - the ratio of the cut value and the volume of the smaller partition region. Here, only the cut value can be influenced

by a change in edge weight assignment. Let $f(w)$ designate the cut value as a function of the connection width. Then the following inequality need to be specified for the cut to happen:

$$f(w) < c_T A_2 \quad (5.3)$$

where A_2 denotes the number of pixels in the smaller partition region. Since the narrower cuts should have smaller weight compared to wider cuts, it is natural to require $f(w)$ to be monotonically increasing.

5.1.2 Polynomial weight assignment

Practically any function $f(w)$ can be represented as a linear combination of simple polynomials w^k . In this section, we show how weighted cut value $f(w)$ can be made proportional to w^k by a weight assignment using distance transform raised to appropriate non-negative power p .

$$w_{ij} = \left[\max_{v_i, v_j \in \pi_1} (D(v_i), D(v_j)) \right]^p \quad (5.4)$$

We can achieve equal weights (equal to one) for all edges, as used in [76] by assigning $p = 0$.

If the weights are defined as in (5.4), then the weighted common boundary length $|\partial\pi_{12}|$ (which is equal to the sum of the weights of the edges along the width of the connection) becomes

$$|\partial\pi_{12}| = 1^p + 2^p + 3^p + \cdots + \left(\frac{w}{2}\right)^p + \left(\frac{w}{2}\right)^p + \cdots + 3^p + 2^p + 1^p \quad (5.5)$$

assuming w is even.

$$|\partial\pi_{12}| = 2 \sum_{x=1}^{\frac{w}{2}} x^p$$

This sum of series can be approximated using the following integral:

$$|\partial\pi_{12}| \simeq 2 \int_1^{\frac{w}{2}} x^p dx$$

$$|\partial\pi_{12}| = 2 \left[\frac{\left(\frac{w}{2}\right)^{p+1} - 1}{p+1} \right]$$

$$|\partial\pi_{12}| \propto w^{p+1} \quad (5.6)$$

The coefficient of proportionality depends only on p and hence is the same for all cuts. This also means it can be absorbed into c_T . Using this value of $|\partial\pi_{12}|$ in (4.4), we get the following cut condition:

$$\frac{w^{p+1}}{S(\pi_2)} < c_T \quad (5.7)$$

The following two lemmas represent generalization of lemmas 3 and 4 for weights assignment (5.4).

Lemma 5. *For a circular region of radius r as in Figure B.1, if the weights of the graph are assigned using (5.4), the isoperimetric ratio of a cut within this circular region is given by*

$$c = \frac{2^{p+2} r^{p-1} \sin^{p+1} \frac{\theta}{2}}{\theta - \sin \theta}$$

where $0 < \theta < \pi$ is the angle subtended by the cut (chord) at the center of the circle. The cut with the smallest isoperimetric ratio occurs at $\theta = \pi$ for $p \leq 2$, i.e., along the diameter of the circle.

Proof. Refer to Appendix B. ■

Lemma 6. *Consider a circular region R_1 of radius r_1 connected to another circular region R_2 of radius r_2 by a bridge of width w . If the weights of the graph are assigned using (5.4), then (i) if $r_2 < r_1$, cut with the minimum isoperimetric ratio occurs either along the connection or along the diameter of R_1 (ii) if $r_2 > r_1$, no cut with the minimum isoperimetric ratio can occur within R_1 .*

Proof. Refer to Appendix B. ■

We use these results in deriving the conditions for obtaining correct cut using Gen ISO for the model images 1 and 2.

Model image 1 - two circles

The two possible cuts for the two circles image are shown in Figure 4.8. From Lemma 6, it follows that the cut with the minimum isoperimetric ratio can happen either along the diameter of R_1 or along the bridge between R_1 and R_2 .

- *First iteration.* There are two possible cuts, through diameter of the larger region

$$c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi}$$

or through the bridge

$$c_2 = \frac{w^{p+1}}{\pi r_2^2}$$

$c_1 < c_2$ indicates failure of the algorithm at the first iteration. $c_1 > c_2$ is not still a failure, but if $c_2 > c_T$ (threshold is too low), then no cuts will be made and it is a failure. So to successfully exit the first iteration, we need two conditions, $c_1 > c_2$ and $c_2 < c_T$. To satisfy the first condition, we need the width to be sufficiently small,

$$w^{p+1} < 2^{p+2}r_1^{p-1}r_2^2$$

and to satisfy the second condition, we need

$$\frac{w^{p+1}}{\pi r_2^2} < c_T$$

For special case $p = 1$, this gives $w < 2\sqrt{2}r_2$ and $\frac{w^2}{\pi r_2^2} < c_T$ respectively.

- *Second iteration.* In this case we are left only with a single circle and hence only one cut is possible with $c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi}$. The failure will occur if this cut is permitted, so for success we need $c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi} > c_T$.

So, the two conditions for success are $w^{p+1} < 2^{p+2}r_1^{p-1}r_2^2$ and $\frac{w^{p+1}}{\pi r_2^2} < c_T < \frac{2^{p+2}r_1^{p-1}}{\pi}$.

We discuss the effect of c_T on these expressions.

- *Fixed c_T .* If c_T is fixed and happens to fall in the range $\frac{w^{p+1}}{\pi r_2^2} < c_T < \frac{2^{p+2}r_1^{p-1}}{\pi}$, then the widest connection that can be cut is $w^{p+1} < 2^{p+2}r_1^{p-1}r_2^2$ which is $w < 2\sqrt{2}r_2$ for $p = 1$. If c_T is chosen too low, this will decrease the maximum width of the connection to $w^{p+1} < c_T\pi r_2^2$ which is $w < r_2\sqrt{c_T\pi}$ for $p = 1$. If c_T is chosen too high, above $c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi}$, the algorithm will fail. If $p = 1$, then c_1 should not exceed $\frac{8}{\pi}$, a value independent of the region sizes and can be easily satisfied.
- *Optimized c_T .* In this case c_T can be chosen anywhere in the range $\frac{w^{p+1}}{\pi r_2^2} < c_T < \frac{2^{p+2}r_1^{p-1}}{\pi}$.

$\frac{2^{p+2}r_1^{p-1}}{\pi}$ and the maximum width that can be cut is $w < 2\sqrt{2}r_2$ when $p = 1$.

Model image 2 - circle-rectangle

Three possible cuts for the circle-rectangle object are shown in Figure 4.9. We have proved in 4.3.3 that ISO fails when $w > 2b$. Therefore, we disregard this condition where the cuts can happen within R_2 .

- *First iteration.* Two possible cuts can occur, one along the diameter of R_1 with $c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi}$ or through the connection with $c_2 = \frac{w^{p+1}}{lb}$. If $c_1 < c_2$, then it is a failure in the first iteration. $c_1 > c_2$ has no initial failure, but a low threshold with $c_2 > c_T$ will not make any cuts and it is a failure. For the success of the first iteration, two conditions $c_1 > c_2$ and $c_2 < c_T$ have to be satisfied. This gives the expression for the width as $w^{p+1} < \frac{2^{p+2}r_1^{p-1}lb}{\pi}$ and c_T as $c_T > \frac{w^{p+1}}{lb}$. For $p = 1$, it is $w < 2\sqrt{\frac{2lb}{\pi}}$ and $c_T > \frac{w^2}{lb}$.
- *Second iteration.* In the second iteration, only a single circle R_1 remains and the only possible cut is through its diameter with $c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi}$. In order to prevent the cut, c_T should be $c_1 = \frac{2^{p+2}r_1^{p-1}}{\pi} > c_T$. For $p = 1$, $c_T < \frac{8}{\pi}$.

The conditions for success are $w^{p+1} < \frac{2^{p+2}r_1^{p-1}lb}{\pi}$ and $\frac{w^{p+1}}{lb} < c_T < \frac{2^{p+2}r_1^{p-1}}{\pi}$. Following explains how c_T affects the expressions for the maximum width of the connection that can be cut by Gen ISO.

- *Fixed value of c_T .* In this case, the maximum width that can be severed is $w < 2\sqrt{\frac{2lb}{\pi}}$. However, selecting a low c_T gives $w < \sqrt{c_T lb}$ and high c_T results in failure.
- *Optimized value of c_T .* In this case, any value of c_T within the range $\frac{w^2}{lb} < c_T < \frac{2^{p+2}r_1^{p-1}}{\pi}$ can be chosen and the maximum width that can be cut is $w < \min\left(2b, 2\sqrt{\frac{2lb}{\pi}}\right)$. Since $l > b$ and probably $l \gg b$, then $\sqrt{lb} \gg b$ and hence the minimum is always $2b$.

Discussion

The generalized version of ISO increases the maximum width of the connection that can be cut, see Table 5.1. For two circles case, the maximum width that can be cut is increased

up to $2r_2$ for Gen ISO from ISO's $\frac{r_2^2}{r_1}$, which is much smaller than r_2 , since $r_2 \ll r_1$. In the case of circle-rectangle object, both ISO and gen ISO can cut connections up to $2b$ in width. In the case of ISO, c_T has to be chosen depending on the dimensions of the regions and improper choice may lead to worse performance. This drawback is resolved in Gen ISO where for $p = 1$, c_T does not depend on the region size and can always be chosen safely, i.e., slightly below $\frac{8}{\pi}$. Furthermore, Gen ISO gives additional flexibility in p whose value can be appropriately altered to suit different applications. However, $p = 1$ appears to be the best alternative since it provides easy selection of c_T and ability to cut sufficiently wider connections.

Table 5.1: Performance comparison: Maximum width of the connection that can be cut by different methods

Method	Two circles	Circle-rectangle
MORPH	$w < 2d_s$	$w < 2d_s$
DWAT	$w < 2r_2$	$w < b$
ISO (fixed c_T)	$w < \min\left(\frac{4r_2^2}{r_1}, c_T\pi r_2^2\right)$ when $c_T < \frac{4}{\pi r_1}$	$w < \min\left(2b, \frac{4lb}{\pi r_1}, c_T lb\right)$ when $c_T < \frac{4}{\pi r_1}$
ISO (optimized c_T)	$w < \frac{4r_2^2}{r_1}$ when $c_T < \frac{4}{\pi r_1}$	$w < \min\left(2b, \frac{4lb}{\pi r_1}\right)$ when $c_T < \frac{4}{\pi r_1}$
Gen ISO (optimized c_T)	$w < 2\sqrt{2}r_2$	$w < 2b$

5.2 Removing narrow connections using graph cuts

Graph cuts based segmentation (GCUT) treats the entire image as a weighted undirected graph. It relies on the specification of foreground f and background b seeds, and finds a partition $\pi = (\pi_1, \pi_2)$ with a minimum cut value $|\partial\pi|$, subject to the constraint that f and b are located in different partition elements:

$$\hat{\pi} = \arg \inf_{\pi} |\partial\pi|, \quad f \subseteq \pi_1, \quad b \subseteq \pi_2 \quad (5.8)$$

The most common practice of using graph cuts is to manually select foreground and

Figure 5.1: Selecting background seeds in GCUT manually provides good segmentation

background seeds and then perform the mincut [56,57], see example in Figure 5.1. Here the user selects the background seed region b inside the unwanted region, which is then removed using minimum cut on the graph. This procedure can be repeated several times until the correct cut occurs. However, such interactive routine cannot be automated, primarily because it is difficult (and often impossible) to know where to place the background seeds within F without user input. Instead, we propose to use graph cuts in such a way that the background seeds are trivially selected outside the preliminary mask, e.g., as the whole set B or its subset. In this case, the whole procedure of graph cuts based narrow connection removal can be made fully automated, because foreground seeds can be selected automatically for most applications, as will be demonstrated in Section 5.4.1. We are not aware of any other approach that uses graph cuts in this fashion, so the proposed approach, we call GCUT, represents a completely novel algorithm for narrow connection removal. GCUT's operation requires a specific way to set the edge weights. This is discussed in the remainder of this section.

Figure 5.2: Cutting narrow connections with GCUT

Let $F = (\pi_1, \pi_2)$ be the correct partition. The following necessary conditions have to be satisfied for GCUT to correctly estimate this partition.

1. We need $f \cap \pi_2 = \emptyset$ and $b \cap \pi_1 = \emptyset$, or else no weight assignment can guarantee $\hat{\pi} = \pi_1$.
2. The cut value corresponding to $\hat{\pi} = \pi_1$ should be less than the cut value corresponding to $\hat{\pi} = \pi_1 \cup \pi_2$, which yields $|\pi_1| < |\pi_1| + |\pi_2| - 2|\partial\pi_{12}| \Rightarrow |\partial\pi_{12}| < \frac{|\pi_2|}{2}$, where $|\cdot|$ stands for the cut value surrounding the region. Otherwise, the graph cut will return the original foreground region, without any cutting.
3. For any region π , $f \subset \pi \subseteq \pi_1$ we should have $|\pi_1| < |\pi|$. Otherwise, the minimum cut will yield region π rather than π_1 .

The goal is to choose edge weights, so that the desired cut has the minimum value of all admissible cuts, i.e. cuts that separate the seed regions. Assigning weights $w_{ij} = 1$ will satisfy condition 2, since the cross-sections of narrow connections can be expected to be much smaller than the boundary lengths of unwanted regions they connect. However, condition 3 will be violated, as the smaller than π_1 regions still containing f will have smaller boundary length. To avoid this we need to increase the weights of the edges connecting voxels located further from the boundary, effectively increasing the weighted boundary lengths of regions smaller than π_1 .

For example, assume that (π'_1, π'_2) is the desired cut, see Figure 5.2. First, we need the foreground and background seeds to be fully contained inside π_1 and B respectively, since the cut cannot be made through the seeds. To make the value of the cut (π'_1, π'_2) smaller than that of (π_1, π_2) or any other admissible cut, the weights of the edges surrounding π'_1 should be made small while the weights of the edges surrounding other cuts large.

We propose using distance transform function from (5.2)(or any of its monotonically increasing functions) to accomplish such a weight assignment. Such an assignment can satisfy both condition 2 and 3. Condition 2 will be satisfied because when the connection is sufficiently narrow, the distance transform values inside it will be small and hence the correct cut value is likely to be smaller than the weighted boundary length of the unwanted region (more details are provided in the next section). Condition 3 will be satisfied because

the weights of the edges deeper inside π_1 will become larger, thus increasing the cut value along any region contained within π_1 . This weight assignment is different from the one used for ISO in the sense that, in the case of ISO, it made the algorithm more general whereas in the case of graph cuts, it naturally follows from the requirement that the background seeds are chosen outside the preliminary mask. In our work, we use distance transform values raised to a non-negative real power p which will give the most general way to describe the weight function.

5.2.1 Performance evaluation

Lets consider Figure 4.5(a) where $R_1 > R_2$ is the desired region of interest. For GCUT to produce a partition (π_1, π_2) where $\pi_1 = R_1$ $\pi_2 = R_2$, it has to satisfy the following condition.

$$\begin{aligned} |\pi_1| &< |\pi_1| + |\pi_2| - 2|\partial\pi_{12}| \\ |\partial\pi_{12}| &< \frac{|\pi_2|}{2} \end{aligned} \tag{5.9}$$

We can understand the working of GCUT better if we further simplify the expression in (5.9) by using definite shapes. Consider the objects in Figure 4.5(a) and 4.5(b). The weighted length of the connection is equal to sum of the edge weights along the width of the connection. From (5.6), we get

$$|\partial\pi_{12}| \propto w^{p+1}$$

Using this value in (5.9), we get

$$w^{p+1} < \frac{|\pi_2|}{2} \tag{5.10}$$

(5.9) gives the expression for the width of the connection for arbitrary shaped regions. Lets now consider specific shapes, the two circles case and the circle-rectangle case.

- **Model image-1: Two circles image.** The perimeter has weights equal to one throughout except along the intersection between the regions. For the two circles case, (5.10) simplifies to

$$w^{p+1} < \frac{2\pi r_2 - w + w^{p+1}}{2}$$

$$w^{p+1} < 2\pi r_2 - w$$

For $p = 1$, this simplifies to

$$w^2 + w - 2\pi r_2 < 0$$

$$w < \frac{-1 + \sqrt{1 + 8\pi r_2}}{2}$$

Assuming $r_2 \gg 1$, this can be approximated as, $w < 2.5\sqrt{r_2}$

- **Model image-2: Circle-rectangle image.** We consider two cases for the circle-rectangle object.

1. Connection width $w < 2b$. Consider the object in Figure 4.9(a). The expression for the width is

$$w^{p+1} < \frac{2(l+b) - w + w^{p+1}}{2}$$

$$w^{p+1} < 2(l+b) - w$$

When $p = 1$,

$$w < \frac{-1 + \sqrt{1 + 8(l+b)}}{2}$$

Assuming $l+b \gg 1$, this can be approximated as, $w < \sqrt{2(l+b)}$

2. Connection width $w > 2b$. Consider the object in Figure 4.9(c). For a cut along the intersection of the two regions, the sum of the edge weights along the cut can be computed as

$$|\partial\pi_{12}| = 2(1^p + 2^p + \dots + b^p) + (w - 2b)b^p$$

$$|\partial\pi_{12}| = 2 \sum_{x=1}^b x^p + (w - 2b)b^p$$

Approximating this to a definite integral and simplifying, we get

$$|\partial\pi_{12}| \simeq 2 \int_1^b x^p dx + (w - 2b)b^p$$

$$|\partial\pi_{12}| = 2 \left[\frac{b^{p+1} - 1}{p+1} \right] + wb^p - 2b^{p+1}$$

$$|\partial\pi_{12}| = \frac{(p+1)wb^p - (p-1)2b^{p+1} - 2}{p+1}$$

When $p = 1$,

$$|\partial\pi_{12}| = \frac{2wb - 2}{2}$$

$$|\partial\pi_{12}| \approx wb \quad (5.11)$$

For a cut in Figure 4.9(c), the sum of edge weights along the cut will be equal to

$$|\partial\pi'_{12}| = 2b^2 + w \quad (5.12)$$

We can find from (5.11) and (5.12) that $|\partial\pi'_{12}|$ is clearly less than $|\partial\pi_{12}|$. Therefore, for the circle-rectangle object with $w > 2b$, parallel cuts within π_2 as shown in Figure 4.9(c) will be chosen over the cut along the intersection of π_1 and π_2 and is therefore considered as a failure.

5.3 Discussion

Table 5.2 compares the performance of Gen ISO and GCUT. Though the expression for Gen ISO for the two circles case implies that it can cut very wide connections, twice the radius still remains the limit. GCUT provides a reasonable but comparatively smaller value for the width of the connection that can be cut, but as we show in Section 5.4 the connection width is sufficient for most applications. For the circle-rectangle case, the two algorithms are roughly equivalent and cut connections up to approximately $2b$. Figure 5.3 and Figure 5.4 show the results of GCUT on the model images with varying connection widths. It is clear from the results that the cutting ability of GCUT follows the expressions in Table 5.2.

Table 5.2: Performance comparison: Maximum width of the connection that can be cut by different methods

Method	Two circles	Circle-rectangle
Gen ISO (optimized c_T)	$w < 2\sqrt{2}r_2$	$w < 2b$
GCUT	$w < 2.5\sqrt{r_2}$	$w < \min\left(2b, \sqrt{2(l+b)}\right)$

The result of ISO has a strong dependence on the location of the seed point, see Figure 4.10. Further, the publicly available code, described in detail in [77], scales poorly with image size, limiting its application to 2D images. A more efficient solution based on tree representation of sparse matrices [76] has not been made publicly available yet. The GCUT's cutting ability coupled with its simple optimization procedure, which allows reaching a globally optimal (vs. locally optimal in case of ISO) solution in a relatively short time (linear or weakly polynomial in the number of voxels) and moderate memory requirements [59] turns the decision in favor of GCUT.

Figure 5.3: GCUT on the two-circles image (a) model image with $w < 2.5\sqrt{r_2}$ (b) GCUT on (a), (c) model image with $w > 2.5\sqrt{r_2}$ (d) GCUT fails to remove the narrow connection in (c)

The model images are considered as simplifications to the real world images rather than approximations. The topic of generalization of performance from simple to complex examples is a tricky one. Though good performance on simple images does not guarantee good performance on complex images, it would also be unusual for an algorithm to be

Figure 5.4: GCUT on the circle-rectangle image (a) model image with $w < 2b$ (b) GCUT on (a), (c) model image with $w > 2b$ (d) GCUT on (c)

failing on a simple image and then suddenly start to perform well on a complex image. Our idea of bringing the simple images first was to make sure that all simple failure causes are removed. Only by looking at simple images first we could tell that standard ISO would fail for some sizes of larger region and is too sensitive to the choice of parameter c_T , thus motivating the modified version of ISO with distance transform based edge weights. Finally, finding out that GCUT performs similar to ISO on simplified images gave us the license to proceed to more complex images. Otherwise, we could have pursued the ISO path. In the following section, we shall design fully automated GCUT algorithms for narrow connection removal in the case of two real world segmentation problems.

5.4 Automated seed selection for GCUT

Critical analysis performed in Chapter 4 and in Sections 5.1 and 5.2 has shown that graph cuts presents an attractive alternative to existing approaches that aim to remove narrow connections. The proposed graph cut approach differs from standard interactive graph cuts segmentation primarily in the requirement that the background seeds are trivially selected as the set of all background pixels. In this section, we show how this requirement can allow making the seed selection (both background and foreground) fully automated, on the example of two common applications.

Fat segmentation

Adipose tissues control a variety of biological processes including energy metabolism. The presence of excess adipose tissues (subcutaneous adipose tissues (SAT) and visceral adipose tissues (VAT)), particularly VAT is linked with increased risk of metabolic and cardiovascular diseases such as atherosclerosis, hypertension, and diabetes [107, 115–118]. Separate quantification of VAT and SAT provides a better indicator of cardiovascular risk factors than total fat volume. In order to accomplish this, accurate segmentation of the VAT and SAT is required. Visually SAT appears as a continuous external layer of fat and VAT is the internal fat, see Figure 5.5. Despite the fact that the two fat depositions are separated by muscles, narrow connections can be occasionally formed between the two.

Figure 5.5: (a) Original MR abdominal image slice (b) SAT (red) and VAT (green) highlighted

The two most widely used approaches for separating abdominal fat compartments are morphological processing [119, 120] and active contours [116, 121, 122]. Other methods include thresholding and region growing [123] and morphon registration and segmentation

[124]. However, none of the existing approaches are ideal. Morphological processing tends to fail when connections formed between fat tissue compartments are too wide. It is also difficult to use on very thin subjects. Snakes grown from outside in tend to get stuck on the fibres within the subcutaneous fat tissue. This motivated us to consider the use of a new approach for the separation of SAT and VAT.

Skull stripping

The problem of separating brain tissues (white matter (WM), grey matter (GM), cerebral spinal fluid (CSF)) from non-brain tissues (skull, scalp, dura, eye sockets, neck tissues, etc.) is called skull stripping (see Figure 5.6). Tissue classification, registration, volumetric analysis of the brain, and cortical surface reconstruction all benefit from accurate removal of these non-brain tissues [125, 126]. The absence of clear boundaries between the brain and the non-brain tissues results in these tissues being connected to each other, thus creating the problem of narrow connection removal.

Figure 5.6: (a) Original MR brain slice (b) image after skull stripping

5.4.1 Seed selection

As explained earlier in the section 5.2, standard graph cuts approach requires initial selection of foreground and background seeds. The background seed selection is simple as the whole background region, i.e., pixels that were not included in the preliminary mask, can be defined as such. Choosing the foreground seed region is less trivial, as it must be sufficiently large to avoid trivial solutions, and at the same time should be fully contained

within the region of interest and do not “leak” into unwanted regions. Though it may appear that this would require manual user intervention, here on the examples of two chosen applications, we show how such selection can be accomplished fully automatically, utilizing only prior information about image structure.

Fat segmentation

Initial thresholding results in a binary segmentation (F, B) , where the foreground region F is assumed to contain the fat tissues (VAT and SAT). Note that in this application, both VAT and SAT can be considered as region of interest, as they both need to be estimated. However, it is convenient to treat only SAT as the region of interest and VAT as the region to be removed. This is due to simpler procedure of defining seeds on SAT, compared to VAT. After SAT is estimated, the subtraction of the result from the initial mask will give VAT.

Figure 5.7: Fat segmentation (a) original abdominal MR image slice (b) preliminary mask (c) foreground seed (d) segmented output - SAT (e) segmented output - VAT

To define the seeds we use the fact that SAT is the outermost continuous layer of fat enclosing the VAT. Here, we assume the fact that SAT layer is thick and continuous and the distance between the outer boundary of SAT and outer boundary of VAT is more than three pixels. We therefore determine the three outermost layers of the image and set it as the foreground seed f , see Figure 5.7(c). The outermost layers are obtained by

determining the perimeter of the image. We set the background seed b to be equal to the background of the initial mask ($b = B$). This seed selection helps to achieve good segmentation results, see Figure 5.7(d) and Figure 5.7(e).

Skull stripping

The foreground region F after initial thresholding is assumed to fully contain the brain as well as an unknown number of non-brain tissue voxels. We set the background seed b to be equal to the background of the initial mask ($b = B$). Choosing the foreground seed f is more challenging, as there is no a priori information concerning where actual brain tissue is located within the initial mask. The foreground seed selection can be based on the fact that in a T1W MR image, WM constitutes the largest region with uniformly high intensity, completely surrounded by darker gray matter. We begin by partitioning the image into cubes of size $5 \times 5 \times 5$ voxels. We then select a cube that best fulfills the combination of brightness (high mean intensity) and uniformity (low variance of intensity values). We found that this always results in a cube located inside the WM, (Figure 5.8).

Figure 5.8: Selection of foreground seed is performed by finding the brightest and the most uniform cube (b), followed by conservative region growing (c)

We can then obtain an estimate of the region bounded by WM using region growing, initialized by the cube. Note that taking the cube alone as the foreground seed can lead to a trivial solution containing only this seed, as the number of edges surrounding the cube is small. The stopping criterion for region growing was conservative in that the grown region will still be separated from non-brain structures by a rim of GM, (Figure 5.8). In other words, being completely surrounded by GM guarantees that regions growing with

sufficiently conservative stopping criteria would never “leak” outside GM and thus would not include non-brain structures.

5.5 Discussion

We proposed a new edge weight assignment for isoperimetric algorithm and graph cuts based removal of narrow connections. We found that graph cuts provides an alternative to other approaches for narrow connection removal. We also proposed graph cuts based automated algorithms for the removal of narrow connections by considering two real world problems, fat segmentation and skull stripping. The seed selection procedure in graph cuts is always seen as an impediment towards automated method of graph cuts. Here, we have shown with two examples that seed selection in graph cuts can be made automatic by using the information extracted from the input images. In the case of skull stripping, it should be noted that the output obtained after the separation of the brain tissues have to undergo a post-processing procedure to get the final brain mask. We therefore allot Chapter 6 for a complete discussion on skull stripping.

Chapter 6

Skull stripping

Skull stripping is the technique of separating brain tissues (white matter (WM), gray matter (GM), cerebral spinal fluid (CSF)) from non-brain tissues (skull, scalp, dura, eye sockets, neck tissues, etc.), see Figure 5.6. Figure 6.1(a) shows an example of an MR brain image with different tissues labeled. Tissue classification, registration, volumetric analysis of the brain, and cortical surface reconstruction all benefit from the accurate removal of these non-brain tissues. In particular, the removal of dura while leaving brain tissue untouched is especially important when estimating cortical thickness ([127], Constrained Laplacian Anatomic Segmentation using Proximity (CLASP) [128]) or grey matter volume (Voxel Based Morphometry (VBM) [125], Structural Image Evaluation, using Normalization, of Atrophy for cross-sectional measurement (SIENAX) [129], part of Functional MRI of the Brain (FMRIB) Software Library (FSL) [126]). Unintended removal of the cortical surface cannot be reversed downstream in the processing pipeline and will result in underestimation of cortical thickness. Inclusion of non-brain structures can result in reduced VBM sensitivity [130], while dural attachments can cause overestimation of cortical thickness [13]. Figure 6.1(b) highlights the estimated pial surface.

Skull stripping is more difficult than traditional segmentation because of the substantial overlap between the intensity ranges of the cortex and the surrounding tissues. Due to lack of difference in T1 relaxation times, no sharp boundaries exist between the brain and some non-brain regions in T1-weighted MR brain images. In this chapter, we propose a new skull stripping technique based on graph cuts. The proposed method achieves

Figure 6.1: MR brain image (a) tissue labels (b) pial surface highlighted in red

better dura removal and facilitates accurate estimation of cortical thickness. To establish the advantages of our approach, we evaluate its performance on legacy standardized test images as well as images collected using a more contemporary MR scanner. We compare our results with those obtained using current state-of-the-art techniques.

We review the existing 3D skull stripping techniques under three categories, region-based approaches, boundary-based approaches, and hybrid approaches in Section 6.1. We then provide the motivation for our research in Section 6.2. We discuss the limitations of the existing techniques and the need for a better technique. We then explain our proposed fully automatic 3D skull stripping approach detailing every step in the pipeline. In Section 6.4 and Section 6.5, we provide information on the data used in our experiments and the metrics used to evaluate the performance of our approach. We then describe the experiments performed, evaluate the performance of different techniques and discuss the results obtained in Section 6.6 and Section 6.7.

6.1 Review of existing skull stripping techniques

Skull stripping methods can be classified mainly into region-based [35, 110], boundary-based [131, 132], and hybrid approaches [133, 134]. Region-based approaches work on individual pixels/voxels without considering the overall structure or shape. Boundary-based approaches work with an initial boundary location. They deform the initial contour based on some criteria to fit the region of interest in the image. Refer to Section 2.2 for the review of region-based and boundary-based segmentation techniques. Hybrid approaches use a combination of region-based and boundary-based approaches to isolate the brain from the non-brain tissues. In this section, we describe some of the most popular and commonly used skull stripping algorithms.

6.1.1 Region-based approaches

Brain surface extractor (BSE)

Brain Surface Extractor (BSE) [110] extracts the brain region using a combination of anisotropic filter, Marr-Hildreth edge detector, and morphological operators. The steps involved in the skull stripping procedure using BSE are as follows:

Figure 6.2: Typical results of existing skull stripping techniques (a) original image, (b) ground truth mask, (c) BSE, (d) WAT, (e) BET, (f) HWA

- *Anisotropic diffusion filtering.* Anisotropic diffusion filtering [60, 135] smooths the noisy regions while retaining sharp boundaries. Two parameters control this filtering operation, diffusion constant D_c and diffusion iterations D_i . The diffusion constant controls how large the gradient should be to be not affected by the filter, thus removing small gradients while preserving larger gradients. The larger the value of the diffusion constant, the more blurring is introduced. Diffusion iterations controls the number of times, the diffusion process will be applied. More iterations produce more blurring. Typical values of D_c and D_i are 25 and 3 respectively.
- *Edge detection.* The edges are detected using Marr-Hildreth edge detector [136]. The parameter edge constant E_c associated with this process controls the scale space of the edges detected in the image. Typical value of E_c is 0.62.
- *Morphological processing.* Erosion operation is performed on the non-edge pixels of the resulting edge detected image. This breaks the narrow connections between regions, if any. The erosion operation is controlled by the erosion size parameter E_s .

that decides the size of the erosion operation. Typical value of E_s for most of the images is 1. Images with high resolution may need a higher value of E_s .

- *Extraction of the component corresponding to the brain.* The largest connected component resulting from the erosion operation is chosen and this represents the brain. Apart from the size, other parameters considered for selection are location within the frame and intensity.

Typical of region-based approaches, BSE may lead to inappropriate removal of the brain tissue, contributed in part by reduced brain signal intensity adjacent to false connections, see Figure 6.2(c). The weaknesses of BSE also include its high sensitivity to image inhomogeneities and strong dependence on empirically chosen parameters. Figure 6.4 illustrates the impact of the change in the value of these parameters on the output. Prediction of the best combination of these parameters is often difficult.

3D Watershed transform (WAT)

The Watershed Algorithm (WAT) [35], an intensity-based approach, relies on a 3D algorithm with pre-flooding performed on the intensity inverted image, selecting the basin to represent the brain. It may fail to remove dura, skull, and various non-brain structures in the neck/eye area (Figure 6.2(d)).

Watershed transform (WAT) based approach for skull stripping [35] relies on two main assumptions. First assumption is that the white matter forms a connected region and the second assumption is that the cerebral spinal fluid, which has lower intensity compared to WM and GM, surrounds the brain and separates the non-brain tissues that are brighter than CSF. Figure 6.3 illustrates the basic idea why watershed transform can be applied to the problem of skull stripping. It shows a conceptual representation of the brain, where WM is surrounded by darker GM and even darker CSF. All these are surrounded by the bright skull and scalp tissues. We consider the variation of intensities along a one dimensional strip and represent it as height information (see Figure 6.3(b)). In this interpretation, the WM can be considered as the top of a hill and CSF as the bottom of a valley. Two regions are said to be disconnected if they are separated by a valley, otherwise they are connected. By inverting the intensity interpretation (see Figure 6.3(c)),

Figure 6.3: Watershed transform - basic idea (a) conceptual representation of brain image (b) interpretation of intensity as height information (c) inverted intensity interpretation

we can find the brain region being clearly separated by the hills (CSF) and forming a basin. The watershed transform segments the image into basins and the task is to identify and isolate the basin corresponding to the brain. Refer to Section 2.2.3 for details on watershed segmentation. See Figure 6.2(d) for the result of watershed transform on a MR brain.

Skull stripping is performed by applying watershed transform on the intensity inverted image and choosing the basin representing the brain. The method is simple as it does not require any preprocessing step, but it often results in oversegmentation due to the strict connectivity condition. This problem is overcome by a post-processing step called

preflooding. Preflooding relaxes the connectivity condition and merges the regions to create a larger region. Refer to Section 2.3.4 for details on preflooding. There is a fast implementation of this method called the fast watershed algorithm that achieves a linear time complexity in the number of pixels. Since this method works based only on the intensity values, the segmented volume after the watershed computation contains some non-brain tissue such as CSF or some parts of the skull/scalp. This indicates the failure of the method in the presence of connections between the brain and the non-brain tissues, which occur quite often in MR brain images. Furthermore, the method also depends on the chosen preflooding height which has a significant impact on the output, see Figure 6.4.

Some of the other region-based approaches that have been proposed are the following: Park et al. [137] use histogram analysis and 2D region growing to segment the brain. Though the method is said to be robust, its performance evaluation is confined only to a limited set of data. Shan et al. [138] propose a histogram based method that uses a combination of thresholding and morphological processing to isolate the brain. Chiverton et al. [139] use statistical techniques like fitting of probabilistic functions and morphological operations for skull stripping. The method works on both adult and infant MRI data.

6.1.2 Boundary-based approaches

The irregular, anatomically implausible brain masks sometimes generated by BSE and WAT can be avoided by imposing additional smoothness constraints with a deformable surface model, which is then fitted onto the brain surface by a set of internal and external forces. This method is utilized in the boundary-based approaches. In this section, we shall review some boundary-based approaches where the deformable surface is guided by intensity information, not the edge information.

Brain extraction tool (BET)

Brain Extraction Tool (BET) [131] is a boundary-based approach based on region intensity. The first step involves some preprocessing to remove 2% of the brightest and the darkest pixels. Then, the intensity value of a pixel lying at 10% between the resulting pixels is chosen as the threshold to roughly distinguish brain and non-brain tissues from the background. The center of gravity and the radius of the brain are estimated by taking

Figure 6.4: Impact of change in the value of parameters on skull stripping. Left: Original MR brain slices (b) BSE with parameters (3,25,0.8,2) (c) BSE with parameters (3,25,0.75,2) (e) BET with default parameter value = 0.5 (f) BET with parameter value = 0.6 (h) WAT with default preflooding height = 0.145 (i) WAT with preflooding height = 0.15

into account all voxels greater than this threshold value. The resulting brain surface is then modeled by a tessellated sphere using connected triangles. This forms the deformable model which is then updated based on a set of forces. They are guided by constraints on surface smoothness and voxel intensities in the vicinity of the surface position. Refer to Section 2.2.4 for details on deformable models.

First, the mean position of all vertices neighboring the given vertex is calculated. This is then used to find the difference vector $\mathbf{s}$, the vector that takes the current vertex to the mean position of its neighbors. The normal ($\mathbf{s}_n$) and tangential ($\mathbf{s}_t$) components of $\mathbf{s}$ are used in the formulation of the movement vector $\mathbf{u}$, that updates each vertex in the sphere to a new position iteratively. Vector $\mathbf{u}$ consists of three components, $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$:

- *Within surface vertex spacing.* This component keeps the vertices equally spaced. Vector $\mathbf{u}_1$ is given by the expression

$$\mathbf{u}_1 = \frac{\mathbf{s}_t}{2} \quad (6.1)$$

- *Surface smoothness control.* This component moves the vertex in line with the neighboring vertices, thus increasing the smoothness of the surface. Vector $\mathbf{u}_2$ is given by the expression

$$\mathbf{u}_2 = f_2 \mathbf{s}_n \quad (6.2)$$

where f_2 is the fractional update constant obtained by applying a sigmoid function of the radius of curvature r .

$$f_2 = \frac{1 + \tanh\left(F * \left(\frac{1}{r} - E\right)\right)}{2} \quad (6.3)$$

$$r = \frac{l^2}{2|\mathbf{s}_n|}$$

E and F are constants that control the scale and offset of the sigmoid,

$$E = \frac{\left(\frac{1}{r_{min}} + \frac{1}{r_{max}}\right)}{2}$$

$$F = \frac{6}{\left(\frac{1}{r_{min}} - \frac{1}{r_{max}}\right)}$$

and r_{min} and r_{max} are the minimum and maximum radii of curvature. From (6.3), it is clearly implied that more smoothing takes place below r_{min} and light smoothing takes place above r_{max} .

- *Brain surface selection term.* This component contains the intensity term. It forces the surface model to fit to the real brain surface. Vector $\mathbf{u}_3$ is given by the expression

$$\mathbf{u}_3 = 0.05f_3l\mathbf{s}_n \quad (6.4)$$

where l is the mean inter-vertex distance and 0.05 indicates the relative weighting constant that sets the balance between the smoothness term f_2 and the intensity-based term f_3 and is found empirically. The update fraction f_3 is given by the expression

$$f_3 = \frac{2(I_{min} - t_l)}{I_{max} - t_2} \quad (6.5)$$

where I_{min} and I_{max} are the minimum and maximum intensity values along the line pointing inwards from the current vertex, t_2 is the intensity that excludes 2% of the darkest pixels and t_l is an intensity threshold that distinguishes between brain and the background. If I_{min} is lower than the local threshold t_l , f_3 becomes negative, causing the surface to move inwards at the current vertex. If it is higher, then the surface moves outwards. The intensity threshold t_l is given by the expression

$$t_l = (I_{max} - t_2) * b_t + t_2$$

where b_t is a preset constant that takes a default value of 0.5.

The total update vector $\mathbf{u}$ is,

$$\mathbf{u} = 0.5\mathbf{s}_t + f_2\mathbf{s}_n + 0.05f_3l\mathbf{s}_n \quad (6.6)$$

Since BET takes into consideration the geometric information (spacing between the vertices in the initial tessellated sphere and curvature at every point on the surface), the

Figure 6.5: Problem in boundary determination in BET: A single dark voxel (pointed by arrow) can alter the intensity threshold and therefore the shape of the surface of the deformable model

results produced are better than that obtained by region-based approaches (see Figure 6.2(e)). However, BET is sensitive to noise/intensity inhomogeneity in the image. The value of the intensity threshold determined from the maximum and minimum intensities along a line inwards from the current vertex in the tessellated sphere can be altered significantly by the presence of a single noise pixel (highlighted in Figure 6.5(b)). The output is also affected by the change in the value of the parameter b_t that controls the intensity threshold, see Figure 6.4.

Level set based methods

Level sets, proposed by Osher and Sethian [40], are implicit representation of deformable models designed to handle topological changes in the image. The formulation of appropriate external forces combined with different pre and post processing steps makes level sets suitable for skull stripping applications. The following are some of the skull stripping methods that use level sets.

Suri [140] proposed an active contour algorithm that uses the level set method for contour evolution. The method uses a fuzzy membership function to classify the image into four components: WM, GM, CSF, and background, and then uses a gradient detector and a deformable model to evolve an active contour to fit the surface between the CSF and

GM. Baillard et al. [141] used a registration procedure to align the brain to an atlas and use the brain surface obtained from the atlas as the initial contour. The initial contour is then evolved using the level set method, in which the speed term is determined by the curvature of the evolving contour. Zhuang et al. [132] used a model based level set to solve the skull stripping problem. Here, the speed function is formulated as a combination of forces, brain surface attraction force and the morphological smoothing force that are derived from the corresponding forces in BET. Mikheev et al. [142] used thresholding and edge detection to obtain the initial mask and then grew the mask to fit the boundary of the brain.

6.1.3 Hybrid approaches

Hybrid approaches are techniques which share the characteristics of both region-based and boundary-based approaches. Skull stripping using hybrid approach consists of employing a set of segmentation algorithms on the original brain image. It is usually a suitable combination of region-based and boundary-based algorithms, so that a better segmentation performance is achieved due to the combined benefits of both approaches.

The most widely used of the hybrid approaches is the Hybrid Watershed Algorithm (HWA) proposed by Segonne et al. [134]. HWA is a hybrid approach that combines watershed algorithm and deformable model, where the latter adds atlas-based shape constraints in order to guarantee an anatomically meaningful brain mask. The method follows the same preprocessing steps as BET. The result of the watershed transform is used as initialization for the deformable model. The deformable model used here differs from that in BET in the formulation of the set of forces to deform the initial curve. This set of forces is governed by the smoothness constraint, intensity constraint and atlas-based shape constraint. The forces that are comprised in this set are the following:

- *Curvature reducing force.* The curvature reducing force controls the smoothness of the brain surface and is the same as the surface smoothness control force in BET.
- *MRI-based force.* The MRI-based force includes the intensity constraint and drives the template towards the true brain boundary. It helps to avoid bright regions and moves the surface towards the desired intensity.

- *Atlas-based force.* The atlas-based force models the shape of the brain. The spherical atlas is constructed from a set of training samples (manually segmented brain images). Two parameters, distance of the brain surface to the center of gravity and the curvature of the surface are computed for all the samples in the training set.

Typical result of HWA is shown in Figure 6.2(f). It can be seen that even additional constraints do not always resolve the problem of eliminating dura attachments, because segmentation can be smooth and still lead to brain loss (Figure 6.2(e)) or include non-brain tissues (Figure 6.2(e-f)).

Some of the other hybrid approaches which have been proposed are the following. Atkins et al. [109] used thresholding, anisotropic diffusion filtering and morphological operations for the preprocessing step and used active contours to localize and segment the brain. Huang et al. [133] used thresholding and active contours for the segmentation. EM algorithm and morphological operations are applied as post-processing steps to refine the result. Zhang et al. [143] used a hidden Markov random field model and expectation-maximization algorithm for the segmentation of the brain.

6.2 Motivation

Skull stripping is difficult because there is an intensity overlap in the boundary regions between brain and non-brain structures (see Figure 6.6). This violates the main assumption in almost all the skull stripping algorithms that the brain is a single connected region separated from non-brain tissues by a rim of CSF. Even with high resolution T1W MR images, thin connections between the brain and other cranial structures exist in the form of dura and connective tissue lining venous sinuses [108], (Figure 6.2(a)) and these connections reduce the accuracy of skull stripping (Figure 6.2(b-c,e-f)).

The problem of elimination of dura and other non-brain regions is addressed in HWA using shape constraints extracted from a training set of manually segmented brain images and hence HWA produces the minimum brain loss [134]. Despite the fact that among existing solutions HWA is the only approach that is very careful at preserving the brain, suiting it for subsequent cortical thickness estimation, it can greatly benefit from further stripping of the dura [144]. Hence, our motivation is to develop an alternative method of

Figure 6.6: Intensity overlap in the boundary region of the brain

skull stripping to achieve further dura reduction while maintaining the same low rate of brain loss, in order to facilitate more accurate estimation of cortical thickness.

6.3 Proposed approach

Brain region falsely connecting with the non-brain regions is one of the most commonly occurring artifacts in the output of the existing skull stripping techniques. This occurs due to overlapping intensities in the region connecting the brain and the non-brain regions. These false connections affect both region-based methods (quite naturally) and boundary-based methods (because the edges in the false connection are weak). One solution to this problem is to impose higher order constraints into the segmentation process. This is done in BET and HWA, where these constraints (smoothness and intensity-based constraints in BET and smoothness, intensity-based and atlas-based constraints in HWA) are merged into a single function and the segmentation is formulated as the minimization of that function. However, the output of BET and HWA (Figure 6.2(e) and Figure 6.2(f)) may still include non-brain tissues (dura mater, non-brain tissues in the anterior and posterior fossa) as their inclusion or exclusion might result in the similar values of optimization criteria. Here, we propose an alternative solution that does not involve complicated geometric constraints and the training set or an atlas.

We propose to approach the skull stripping problem in a two-step procedure where the

first step is to provide an initial segmentation and the second step involves a retrospective procedure to remove narrow connections and post-processing. Our method is motivated by earlier work [108, 109, 138, 145] that suggested segmenting the brain by using simple intensity thresholding followed by morphological opening operations to cut the narrow connections. Instead of morphological operations, which can only remove sufficiently narrow connections, we rely on graph theoretic image segmentation technique to position the cuts that serve to isolate and remove dura. The proposed method of skull stripping (GCUT) consists of the following group of operations (Figure 6.7) that are performed in sequence.

- Thresholding to obtain preliminary mask
- Removal of narrow connections using graph cuts
- Post-processing

Figure 6.7: The pipeline of proposed skull stripping approach

6.3.1 Obtaining preliminary mask

The preliminary mask is obtained by thresholding the original image using a suitable threshold. The initial thresholded mask must satisfy two conditions:

1. The brain should be weakly connected to non-brain structures.
2. The mask should preserve as much brain as possible, since the subsequent narrow connection removal can only further reduce the mask.

For T1W images, the appropriate intensity threshold lies somewhere between the mean intensities of GM and CSF. Threshold values that are too low may lead to the inclusion of

Figure 6.8: Effect of different threshold values on the quality of initial mask. Too low threshold (second column) leads to insufficient separation between brain and non-brain structures, too high threshold (right column) results in brain loss

CSF and dura, resulting in the appearance of strong connections between brain and the cranial vault (Figure 6.8). Values that are too high may provide a clearer demarcation between brain and non-brain structures but at the expense of brain erosion. The desired threshold results in a mask with sufficiently narrow connections and acceptable brain loss that can be compensated for during post-processing (Figure 6.8).

Existing methods of threshold selection utilize the image histogram - using the histogram's first valley [109] or fitting a function of Otsu's threshold [108, 138]. However, histogram features may not identify an appropriate threshold. For example, choosing the first valley of the histogram may result in a threshold that is too low (Figure 6.8). In this example, an appropriate threshold is located in between the first valley and peak of GM tissue distribution.

For T1W images, we found that a good threshold lies within 32–40% of WM intensity, i.e. $0.32I_{WM} \leq T \leq 0.4I_{WM}$ and empirically chose $T = 0.36I_{WM}$ for subsequent tests. 'WM intensity' is estimated by averaging intensities within WM seed voxels, as explained in Section 5.4.1.

6.3.2 Removal of narrow connections

This step of our skull stripping pipeline assumes that the initial brain mask has non-brain tissues connected to the brain only by narrow bridges. In our approach (GCUT), we use graph cuts to cut the narrow bridges between the brain and non-brain. Refer to

5.2 for details on the distance transform based edge weight assignment and seed selection for graph cuts. This assignment increases the weight of edges located deep within the foreground region, making cuts here less likely. On the other hand, weights inside the narrow connections remain small.

In the case of an occasional wide connection, this procedure fails because of high distance transform values inside the connection (Figure 6.9, middle column). A simple solution to this is to use the fact that appropriate cuts between brain and non-brain structures usually go through voxels of relatively lower intensity, e.g. partial volume voxels between GM, CSF and dura. The weights of the edges that connect these voxels should thus be reduced, to favor cutting through them. We found the following assignment to satisfy the above criteria and work well (Figure 6.9, right column):

$$w_{ij}^* = w_{ij} \cdot \left[\exp \left(k \frac{\min_{v_i \in F} (I(v_i), I(v_j)) - T}{I_{WM} - T} \right) - 1 \right] \quad (6.7)$$

where w_{ij} is given by (5.2), T is the threshold used to obtain the initial mask and k is a parameter that controls the contribution of voxel intensities. For T1W images, we obtained good results when $1 \leq k \leq 3$ and chose $k = 2.3$ empirically for subsequent tests. The mean intensity of WM (I_{WM}) can be estimated by averaging intensities of the voxels inside the foreground seed.

6.3.3 Postprocessing

The initial thresholding procedure can inadvertently remove some darker partial volume voxels at the GM/CSF boundary. To improve results further, we applied a post-processing step to recover partial volume voxels and CSF. This was accomplished by performing morphological closing operation (10mm voxel dilation and 10mm voxel erosion, sizes rounded to the nearest integer) on the final mask. It is then followed by an addition of a layer of voxels at the cuts and two layers of dark voxels elsewhere. This smoothes the mask and fills in the ventricles (Figure 6.7).

Figure 6.9: Original image (left column), graph cuts with weight assignment based on distance transform (middle), graph cuts with weight assignment based on distance transform and intensity (right). First row is an image from data set 3, second row - from data set 1

6.4 Data sets

We used the following four data sets for performance evaluation:

- **Data Set 1:** 18 T1W volumes from the Internet Brain Segmentation Repository (IBSR¹), resolution $1 \times 1 \times 1.5$ mm, 128 slices.
- **Data Set 2:** 20 T1W volumes of normal subjects from IBSR¹, resolution $1 \times 1 \times 3.1$ mm, 64 slices.
- **Data Set 3:** 15 healthy subjects (age 56 – 71, 9 males), each scanned once on a Siemens Allegra 3T scanner using the following parameters: TR = 2300 ms, TE = 2.91 ms, TI = 900 ms, FA = 9 degrees, resolution $1 \times 1 \times 1.1$ mm, 160 slices.
- **Data Set 4:** 15 healthy subjects (age 56 – 71, 4 males), each scanned once on a

¹<http://www.cma.mgh.harvard.edu/ibsr/>

Siemens Allegra 3T scanner using the following parameters: TR = 2300 ms, TE = 2.91 ms, TI = 900 ms, FA = 9 degrees, resolution $1 \times 1 \times 1.1$ mm, 160 slices.

In contrast to data set 3, segmentation of images from data set 4 gave rise to a variety of problems, e.g. inclusion of dura with GM, underestimation of WM surface, etc., which entailed substantial manual editing. The purpose of including this data set was to illustrate the potential benefit of the proposed approach for the subsequent segmentation of brain tissues, using FreeSurfer segmentation pipeline as an example.

Figure 6.10: Example of ground truth masks, top row - data sets 1 and 2 (IBSR), cerebellum included, bottom row - data sets 3 and 4 (Siemens Allegra 3T), cerebellum excluded

6.4.1 Ground truth

For all four data sets, the ground truth was defined as GM+WM [108, 146, 147]. Manual, expert segmentation containing GM, WM, and subcortical structures (inclusive of cerebellum) was already included in the data sets 1 and 2 (Figure 6.10, top row). To obtain ground truth for data sets 3 and 4, these were processed using FreeSurfer 3.04 (Marti-

nos Centre, Charlestown MA)², which is a set of automated tools for reconstruction of the brain's cortical surface from structural MRI data [127,148]. It contains both volume-based and surface-based analysis, which primarily use the white matter surface. It includes tools for the reconstruction of topologically correct and geometrically accurate models of pial surfaces and measurement of cortical thickness. The resultant pial surfaces obtained from FreeSurfer were edited by an expert and converted to volume masks. Note that the ground truth for data sets 3 and 4 excluded cerebellum (Figure 6.10, bottom row).

Table 6.1: Estimated CNR between GM and dura/CSF, and coefficient of variation within WM for tested data sets

Data set	CNR	CV
	Mean (SD) [range]	Mean (SD) [range]
1	28.61 (11.70) [10.36 52.49]	0.10 (0.02) [0.06 0.13]
2	34.94 (12.15) [13.21 56.85]	0.13 (0.08) [0.08 0.33]
3	35.14 (6.0) [23.84 47.82]	0.09 (0.01) [0.08 0.11]
4	34.37 (6.98) [19.90 46.60]	0.09 (0.01) [0.07 0.10]

6.4.2 Image quality

To further highlight the differences between data sets, we evaluated two image quality metrics, namely contrast-to-noise ratio (CNR) between GM and dura/CSF and coefficient of variation (CV) of WM (Table 6.1). To estimate CNR, we first defined dura/CSF region by selecting one voxel thin layer external to GM. CNR was then defined as the difference between median intensities of GM and dura/CSF divided by the standard deviation of the noise, where the latter was estimated from a manually selected region of interest (ROI) in the air space outside the head. Images with dark CSF, good separation between GM and dura/CSF, and low noise have higher CNRs and should be easier to skull strip. To estimate the CV of WM, we divided the standard deviation of image intensities within

²<http://surfer.nmr.mgh.harvard.edu/fswiki/recon-all>

WM region (as defined by ground truth) by the mean intensity within the same region. This metric is often used in the evaluation of non-uniformity correction algorithms as it is sensitive to slow variations in image intensities. However, it is also sensitive to noise and presence of imaging artifacts, sharp intensity variations, ghosting, all of which may be detrimental to skull stripping.

According to Table 6.1, data set 2 on average has better delineation between brain and non-brain structures and hence higher CNR, compared to data set 1. However, its images exhibited a large variety of imaging artifacts (strong intensity nonuniformity, ghosting), which is reflected in the 30% increase in $CV(WM)$ compared to data set 1. Images from data set 3 and 4 had CNR similar to that of images in data set 2, but better uniformity. Their quality was also much more consistent, which is reflected in halving of the metric variability compared to data sets 1 and 2. Data set 4 had slightly lower CNR compared to data set 3, possibly explaining why it was more problematic.

6.5 Evaluation metrics

In order to evaluate the performance of the skull stripping techniques, we computed the similarity coefficients (Jaccard similarity index and Dice similarity index) [110,134,144,149] and the segmentation error using false positive (FP) rate and false negative (FN) rate [134,150–152]. Refer to Appendix A for details on these performance evaluation metrics.

One serious deficiency of these metrics in the setting of skull stripping is that they may lead to unfair comparisons between different approaches. This is because in addition to GM, WM, and subcortical structures that must be preserved in a valid brain mask, there are also ‘optional’ structures, e.g. brain stem and CSF, inclusion or exclusion of which does not materially affect the quality of the tissue segmentation that usually follows. If the ground truth contains only GM and WM, as was the case for all our data sets, these ‘optional’ regions would contribute to artificial increase in FP rate and reduction in similarity coefficients. Hence skull stripping that preserves less CSF or cuts more of brain stem can be falsely declared superior to skull stripping which does the opposite, even if the two approaches are equivalent in terms of utility for subsequent processing or analysis.

Another problem that primarily affects existing formulation of FP rate is its insensitiv-

Figure 6.11: Skull stripping result 1, FP=27%, FP_adj=8% (left), skull stripping result 2, FP=31%, FP_adj=6% (right)

ity to the types of preserved non-brain structures, potentially making it a poor measure of subsequent segmentation performance. For example, consider two qualitatively different skull stripping results shown in Figure 6.11. The mask on the left contains large chunks of skull and dura; their proximity to a large area of the brain surface would likely to cause pial surface overgrowth. The mask on the right preserves a large portion of orbital contents/skull that are located further from the brain surface and are less likely to cause segmentation problems, despite causing a higher FP rate (31% vs. 27%).

To provide results that are less affected by CSF voxels, we followed suggestions made by prior investigators Boesen et al. [153], Rex et al. [154], and Shattuck et al. [110] and

included JS, DS, and FP metrics calculated without ‘dark’ voxels, in addition to the traditional metrics. The dark voxels were intended to be a coarse estimate of CSF voxels and were classified as those voxels that had intensity below $0.36I_{WM}$, the same threshold that was used to obtain preliminary mask, and did not belong to the ground truth. The new metrics can be calculated the same way as the normal metric, through exclusion of dark voxels from segmented mask. Note that FN rate did not need to be recalculated because ground truth mask did not contain dark voxels by their definition. To reduce the influence of brain stem and cerebellum (for data sets 3 and 4 only), and to differentiate the skull/dura from other non-brain structures, we also provided an ‘adjusted FP rate’ that excluded from consideration non-brain voxels located further than 5mm external to the ground truth boundary. As shown in Figure 6.11, this new metric is more sensitive to the amount of preserved dura and neighboring skull, resulting in a lower estimate for skull stripping result on the right.

6.6 Quantitative performance evaluation

We conducted several experiments to evaluate the performance of our skull stripping approach (GCUT) over other existing approaches. In the first experiment, we performed skull stripping using GCUT on different data sets and evaluated its performance. Then, we evaluated the effect of GCUT on subsequent brain segmentation performed using FreeSurfer 3.0.4. The final experiment involved determining the robustness and sensitivity to the algorithm’s parameters.

6.6.1 Comparison with existing skull stripping approaches

In this experiment, we compared GCUT with BET, BSE, WAT, and HWA on four chosen data sets, see results in Table 6.2 - Table 6.5. BET, WAT, and HWA were used with default parameters. For BSE we changed the default parameter values as suggested in [150] (diffusion constant = 35, diffusion iterations = 3, edge constant = 0.62, erosion size = 2); this resulted in better performance compared with the default values on the four chosen data sets.

Substantial disparity between similarity coefficients and FP rates of data sets 1 and 2

Table 6.2: Comparison of graph cuts skull stripping approach (GCUT) with existing skull stripping approaches, Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), and Hybrid Watershed Algorithm (HWA), using data set 1 (18 1.5mm scans, IBSR). GCUT_HWA stands for mask obtained by intersecting GCUT and HWA mask.

Method	DS (w/o dark pixels) Mean (SD) [range]	JS (w/o dark voxels) Mean (SD) [range]	FN (%) Mean (SD) [range]	FP (w/o dark voxels, %) Mean (SD) [range]	FP_adj (%) Mean (SD) [range]
BSE	0.92* (0.04) [0.84 - 0.98]	0.86* (0.08) [0.72 - 0.95]	5.87* (7.82) [0.44 - 22.41]	7.94 (7.97) [0.6 - 30.9]	2.17* (1.98) [0.10 - 7.40]
BET	0.94 (0.04) [0.80 - 0.98]	0.89 (0.07) [0.67 - 0.95]	1.93* (2.25) [0.11 - 6.73]	8.10 (7.46) [1.9 - 35.4]	2.27* (1.28) [0.56 - 5.42]
WAT	0.94 (0.03) [0.85 - 0.97]	0.89 (0.05) [0.73 - 0.94]	2.45** (1.79) [0.08 - 7.08]	7.44 (8.19) [1.5 - 35.9]	2.50* (2.08) [0.61 - 8.94]
HWA	0.94** (0.02) [0.90 - 0.96]	0.89** (0.03) [0.82 - 0.93]	0.015 (0.02) [0 - 0.07]	9.04** (5.43) [2.9 - 22.4]	4.12** (2.27) [1.25 - 8.90]
GCUT	0.95 (0.02) [0.93 - 0.97]	0.91 (0.03) [0.86 - 0.95]	0.029 (0.04) [0.0001 - 0.15]	7.09 (4.47) [1.5 - 15.8]	3.02 (2.05) [0.49 - 7.01]
GCUT_HWA	0.95* (0.01) [0.93 - 0.97]	0.91* (0.03) [0.87 - 0.95]	0.044** (0.05) [0.0004 - 0.17]	6.42* (4.28) [1.4 - 15.4]	2.72* (1.97) [0.46 - 6.94]

Bold emphasis designates the best value among all rows.

* $0.001 < p < 0.05$, where p designates the statistical significance of the difference between the current value and GCUT's result.

** $p < 0.001$.

compared to data sets 3 and 4 was due to ground truth definition; the ground truth for data sets 1 and 2 contained cerebellum whereas the ground truth for data sets 3 and 4 did not. Since all tested approaches preserved cerebellum as part of the brain, this led to higher FP rates and lower DS, JS for data sets 3 and 4. The adjusted FP rate was affected to a lesser degree, as it only counted the voxels within immediate vicinity of the brain surface.

Our performance evaluation of existing algorithms was consistent with previous findings. BSE's reported FN rate ranged from 2 – 12% [110, 151, 153, 154], depending on whether the parameters were fixed for the whole set or tuned for each individual volume [153]. The FN rates for BET and WAT were reported to be 2.7 – 4.3% [151, 154] and 2% [154] respectively. Our findings fell in the same range, except that BSE and WAT performed poorly on data sets 2-4 in terms of FN rate. Our tests confirmed that HWA

Table 6.3: Comparison of GCUT with existing skull stripping approaches using data set 2 (20 normal subjects, IBSR)

Method	DS (w/o dark pixels) Mean (SD) [range]	JS (w/o dark voxels) Mean (SD) [range]	FN (%) Mean (SD) [range]	FP (w/o dark voxels, %) Mean (SD) [range]	FP_adj (%) Mean (SD) [range]
BSE	0.80** (0.22) [0 - 0.95]	0.70** (0.22) [0 - 0.91]	27.01** (24.05) [3.51 - 100]	4.04* (3.08) [1.49 - 12.65]	0.69** (0.78) [0 - 3.12]
BET	0.80** (0.14) [0.59 - 0.95]	0.68** (0.18) [0.42 - 0.91]	0.10** (0.11) [0.01 - 0.39]	57.78* (47.71) [9.48 - 139.27]	6.42** (2.94) [2.52 - 13.02]
WAT	0.80* (0.15) [0.52 - 0.96]	0.70* (0.19) [0.35 - 0.92]	24.53** (22.74) [0.12 - 62.73]	7.46 (4.17) [3.42 - 17.04]	2.23** (1.70) [0.68 - 6.31]
HWA	0.90 (0.13) [0.51 - 0.97]	0.83 (0.17) [0.35 - 0.94]	1.91 (6.53) [0 - 28.87]	26.58 (49.85) [4.81 - 188.81]	4.89 (3.82) [1.15 - 16.31]
GCUT	0.93 (0.09) [0.56 - 0.97]	0.87 (0.12) [0.39 - 0.94]	0.011 (0.02) [0 - 0.058]	18.48 (32.5) [5.03 - 155.34]	4.29 (2.84) [1.21 - 12.21]
GCUT_HWA	0.92 (0.09) [0.56 - 0.97]	0.87 (0.13) [0.39 - 0.94]	1.92 (6.53) [0 28. - 88]	16.65* (32.78) [4.54 - 155.34]	3.47* (2.47) [0.95 - 12.21]
Excluding 5 failed volumes for HWA and GCUT_HWA and 1 failed volume for GCUT and BSE					
BSE	0.84** (0.11) [0.62 - 0.95]	0.74** (0.15) [0.45 - 0.91]	23.17** (17.29) [3.51 - 54.53]	3.59** (2.38) [1.49 - 9.65]	0.72** (0.78) [0.19 - 3.12]
HWA	0.95* (0.01) [0.92 - 0.97]	0.90* (0.02) [0.85 - 0.94]	0.012 (0.04) [0 - 0.17]	10.83* (3.37) [4.81 - 16.89]	3.89** (1.56) [1.15 - 6.55]
GCUT	0.94 (0.02) [0.89 - 0.97]	0.89 (0.03) [0.81 - 0.94]	0.012 (0.02) [0 - 0.058]	11.28 (4.45) [5.03 - 23.17]	3.87 (2.20) [1.21 - 10.59]
GCUT_HWA	0.95* (0.01) [0.93 - 0.97]	0.91* (0.02) [0.87 - 0.94]	0.02 (0.05) [0 - 0.20]	9.70* (3.26) [4.54 - 15.37]	3.17** (1.52) [0.95 - 5.75]

Bold emphasis designates the best value among all rows.

* $0.001 < p < 0.05$, where p designates the statistical significance of the difference between the current value and GCUT's result.

** $p < 0.001$.

offers a favorable trade-off between FN and FP rates, leading to negligible (almost zero) brain loss at the expense of slightly higher FP rates [144]. This trend was violated only for data set 2, where HWA resulted in substantial brain loss. Further investigation revealed that this poor performance was due to five subjects with severe artifacts in data set 2 for which HWA either led to up to 50% brain loss or returned a running error. After exclusion of these subjects (see the bottom of Table 6.3), HWA's performance became consistent with that on other data sets. Note also that BET and BSE were similar to HWA in terms

of adjusted FP rate but 2-4 times worse in terms of FP rates (data sets 3-4). This suggests that HWA tends to preserve smaller non-brain structures in the vicinity of the brain surface, e.g. skull and dura, while BET and BSE preserve larger non-brain structures in the eye and neck areas.

Table 6.4: Comparison of GCUT with existing skull stripping approaches using data set 3 (Siemens Allegra 3T scanner, good quality)

Method	DS (w/o dark pixels) Mean (SD) [range]	JS (w/o dark voxels) Mean (SD) [range]	FN (%) Mean (SD) [range]	FP (w/o dark voxels, %) Mean (SD) [range]	FP_adj (%) Mean (SD) [range]
BSE	0.71* (0.22) [0 - 0.90]	0.58** (0.21) [0 - 0.82]	9.54 (25.03) [2.34 - 100]	54.04** (31.22) [16.2 - 102.7]	3.45 (1.44) [0.11 - 6.26]
BET	0.77** (0.09) [0.62 - 0.85]	0.63** (0.11) [0.45 - 0.74]	1.75** (1.47) [0.04 - 4.30]	51.16** (31.78) [19.4 - 112.2]	5.50** (2.68) [2.93 - 12.15]
WAT	0.83* (0.11) [0.53 - 0.89]	0.72* (0.13) [0.36 - 0.80]	13.1* (18.2) [1.18 - 57.32]	14.28* (5.48) [2.4 - 22.6]	2.63 (0.89) [1.02 - 4.52]
HWA	0.88** (0.01) [0.86 - 0.90]	0.79** (0.02) [0.76 - 0.81]	0.013 (0.02) [5e-4 - 0.05]	18.21** (2.10) [14.9 - 21.5]	3.63** (0.54) [2.86 - 4.40]
GCUT	0.89 (0.01) [0.87 - 0.90]	0.80 (0.02) [0.77 - 0.82]	0.025 (0.03) [5e-4 - 0.14]	17.23 (1.99) [14.2 - 20.6]	2.95 (0.42) [2.22 - 3.55]
GCUT_HWA	0.89** (0.01) [0.87 - 0.90]	0.80** (0.02) [0.77 - 0.82]	0.035* (0.04) [0.006 - 0.16]	16.91** (1.91) [14.0 - 20.3]	2.81** (0.37) [2.16 - 3.33]

Bold emphasis designates the best value among all rows.

* $0.001 < p < 0.05$, where p designates the statistical significance of the difference between the current value and GCUT's result.

** $p < 0.001$.

Overall, compared to HWA, our approach led to substantial decrease (10 – 30%) in adjusted FP rate, which was statistically significant for data sets 1 and 3 ($p < 0.001$). The FN rates of two approaches were negligible; HWA did slightly better on data sets 1 and 3 while GCUT did better on data set 2 and 4. None of the differences was statistically significant. Note that HWA and GCUT performed similar to each other on data set 2, but only after excluding 5 subjects for HWA and one subject for GCUT. GCUT was superior when all subjects in data set 2 were used for performance evaluation.

The results discussed here relate to images that were not corrected for intensity non-uniformity. This said, intensity correction using N3 algorithm [15] with default parameters had little effect on the subsequent skull stripping performance of our algorithm, consistent

Table 6.5: Comparison of GCUT with existing skull stripping approaches using data set 4 (Siemens Allegra 3T scanner, poor quality)

Method	DS (w/o dark pixels) Mean (SD) [range]	JS (w/o dark voxels) Mean (SD) [range]	FN (%) Mean (SD) [range]	FP (w/o dark voxels, %) Mean (SD) [range]	FP_adj (%) Mean (SD) [range]
BSE	0.72* (0.22) [0 - 0.89]	0.60* (0.20) [0 - 0.80]	8.59 (25.30) [1.12 - 100]	52.86** (30.08) [16.3 - 103.4]	5.10* (2.14) [0.03 - 9.24]
BET	0.78** (0.08) [0.61 - 0.88]	0.64** (0.10) [0.44 - 0.79]	1.46* (1.79) [0.01 - 5.98]	49.81** (28.66) [16.3 - 116.2]	6.21* (2.62) [2.79 - 11.87]
WAT	0.83* (0.09) [0.57 - 0.91]	0.72* (0.12) [0.40 - 0.84]	12.22* (14.89) [0.29 - 49.25]	16.95 (4.65) [8.40 - 23.55]	3.76 (1.35) [1.55 - 6.15]
HWA	0.88 (0.01) [0.85 - 0.90]	0.79 (0.02) [0.75 - 0.82]	0.055 (0.07) [0.0015 - 0.26]	19.55 (2.61) [13.3 - 23.2]	4.28 (0.79) [3.12 - 5.79]
GCUT	0.88 (0.02) [0.83 - 0.90]	0.79 (0.03) [0.71 - 0.82]	0.038 (0.04) [0.0010 - 0.13]	18.67 (3.64) [10.4 - 27.1]	3.92 (1.40) [1.80 - 8.02]
GCUT_HWA	0.89* (0.01) [0.86 - 0.90]	0.80* (0.02) [0.76 - 0.82]	0.089* (0.09) [0.0025 - 0.38]	17.36* (2.54) [10.2 - 20.2]	3.19* (0.56) [1.77 - 4.0]

Bold emphasis designates the best value among all rows.

* $0.001 < p < 0.05$, where p designates the statistical significance of the difference between the current value and GCUT's result.

** $p < 0.001$.

with previous findings [144]. For example, there was no change in JS, DS and FN rate, but a slight increase in adjusted FP rate from 2.95 (without N3 correction) to 3.2 (with N3 correction) for images in data set 3. For data set 4, correction helped improve JS from 0.79 to 0.8 and DS from 0.88 to 0.89. The FN rate was also slightly improved from 0.038 to 0.034, but at the expense of a slight increase in FP adjusted rate (from 3.92 to 4.09).

Since both GCUT and HWA had very low FN rates, an obvious way to further decrease the FP rate is by using the intersection of the two masks; the results are shown in Table 6.2 - Table 6.5 under the name GCUT_HWA. The intersection led to a small but tolerable increase in FN rate and further 5 – 20% decrease in the adjusted FP rate.

6.6.2 Effect on FreeSurfer segmentation pipeline performance

To evaluate the practical usefulness of GCUT to the subsequent brain tissue segmentation, we applied it in the context of FreeSurfer segmentation pipeline. In FreeSurfer, the pial surface localization is limited to the brain mask obtained after the skull stripping

procedure. An imperfect skull strip can affect the resultant pial surface in two ways. Inclusion of non-brain structures, such as dura matter, may result in overestimation of the pial surface. The opposite problem, brain loss inside the mask, will logically lead to underestimation of the pial surface. Given the ground truth, one way to assess the quality of segmentation is to determine the distance between corresponding points on the ground truth and test pial surfaces. However, we have found this difficult to implement in practice. First, this does not distinguish between underestimation and overestimation. Second, the surface mesh used in FreeSurfer has varying distances between vertices depending on the local surface complexity, making it difficult to find unique pairs of matching vertices. To circumvent these problems, we converted the pial surfaces into volume masks using FreeSurfer’s built-in function ‘mri_surfmask’ and evaluated performance in volume space. We also excluded from computation of FP and FN a layer of voxels one voxel thick closest to the ground truth volume boundary. This made the computation more robust, as even small sub-voxel changes in the surface position can lead to inclusion or exclusion of an entire voxel. We also excluded a 10mm thick slab centered on the midsagittal plane because FreeSurfer segmentation is known to include dura at this location. The underestimation and overestimation of the pial surface were then expressed using standard FN and FP rates.

Table 6.6: Effect of brain masks on subsequent estimation of pial surface position in 15 hemispheres with prior overestimation problem

	FP (%)				FN (%)		
	HWA	GCUT	GCUT_HWA	GCUT_HWA - HWA	HWA	GCUT	GCUT_HWA
AVG	0.06	0.17	0.03	-0.03	0.12	0.15	0.13
STD	0.02	0.34	0.03	0.03	0.17	0.16	0.14
p-value				< 0.001			

Our experiment was performed as follows. We first visually examined the HWA segmentation results of data set 4, processing the left and right hemispheres separately. We selected 15 hemispheres where the pial surface was overestimated (later referred to as problematic) and 15 where this did not occur (referred to as non-problematic). Finally, we compared FN and FP rates of the three masks, HWA, GCUT, and GCUT_HWA across the two sets of hemispheres, see results in Table 6.6 and Table 6.7.

Table 6.7: Effect of brain masks on subsequent estimation of pial surface position in 15 hemispheres without prior overestimation problem

	FP (%)			FN (%)		
	HWA	GCUT	GCUT_HWA	HWA	GCUT	GCUT_HWA
AVG	0.02	0.02	0.02	0.47	0.42	0.31
STD	0.01	0.01	0.01	0.38	0.30	0.26

Contrary to our expectations, GCUT brain mask performed poorly; it led to a 2.5 fold increase in overestimation in problematic hemispheres. However, the intersection of HWA and GCUT masks, GCUT_HWA, performed very well, halving the FP rate of problematic hemispheres. For non-problematic hemispheres, the intersection mask did not change the FP rate (as expected) and also resulted in a small decrease in FN rate. The reason for such a decrease is unclear; it appears that the large FN rates reported in Table 6.6 and Table 6.7 were not a result of inappropriate masking but a result of residual intensity non-uniformity that often causes pial surface underestimation on the medial and inferior surfaces of the temporal lobe. The negligible increase in FN rate after using GCUT_HWA mask suggests that the new mask does not lead to greater brain loss than HWA and that used alone, it is fine for non-problematic brains.

Visual examination of pial surfaces segmented with the help of GCUT_HWA showed that overestimation problem was completely resolved in 11 out of the 15 problematic hemispheres. Figure 6.12 shows several examples of the successful use of GCUT_HWA.

6.6.3 Robustness and sensitivity to algorithm's parameters

In this work, robustness refers to the ability of the algorithm to successfully process images whose appearance or quality substantially deviated from the norm. We applied our algorithm to examples of such images. For example, data set 1 contained a number of images with very poor contrast between gray matter and CSF. Images in data set 2 exhibited various imaging artifacts, such as strong intensity inhomogeneity and ghosting. Data set 4 was selected because it comprised brains that had strong connections between dura and GM as well as pronounced intensity inhomogeneity, characteristics that usually lead to segmentation errors.

To highlight robustness of GCUT compared to HWA we calculated the number of

Figure 6.12: Problematic FreeSurfer segmentation performance using HWA brain mask (left) is improved using GCUT_HWA (right), see pial surface (highlighted in red) overgrown on the left

Table 6.8: Sensitivity of GCUT performance to intensity threshold parameter for $k = 2.3$

Data set	T (% of I_{WM})	30	32	34	36	38	40
1	FN (%)	0.015	0.021	0.018	0.031	0.0389	0.0726
	FP_adj (w/o dark voxels, %)	5.22	4.57	3.73	2.94	2.37	1.81
2	FN (%)	0.004	0.006	0.010	0.012	0.015	0.02
	FP_adj (w/o dark voxels, %)	16.30	14.43	12.81	11.28	9.73	8.76
3	FN (%)	0.023	0.023	0.023	0.025	0.029	0.0369
	FP_adj (w/o dark voxels, %)	4.28	3.74	3.41	2.95	2.68	2.45
4	FN (%)	0.030	0.030	0.033	0.038	0.058	0.086
	FP_adj (w/o dark voxels, %)	6.04	5.15	4.47	3.92	3.49	3.09

Table 6.9: Sensitivity of GCUT performance to parameter k controlling the influence of voxel intensity on cut positions for $T = 0.36I_{WM}$

Data set	k	1	1.5	2	2.3	2.5	3
1	FN (%)	0.129	0.086	0.041	0.029	0.023	0.085
	FP_adj (w/o dark voxels, %)	2.80	2.86	2.93	3.02	3.14	3.30
2	FN (%)	0.016	0.013	0.011	0.012	0.013	0.029
	FP_adj (w/o dark voxels, %)	3.39	3.37	3.60	3.87	3.91	3.81
3	FN (%)	0.034	0.027	0.025	0.025	0.023	0.019
	FP_adj (w/o dark voxels, %)	2.91	2.92	2.96	2.95	2.98	3.14
4	FN (%)	0.092	0.052	0.049	0.038	0.035	0.069
	FP_adj (w/o dark voxels, %)	3.64	3.78	3.91	3.92	4.01	4.30

failures of each approach on the four data sets. ‘Failure’ was defined qualitatively as either gross brain loss or preservation of majority of non-brain structures. Using this criterion, GCUT was more robust than HWA; the failures were observed only for data set 2, five for HWA (substantial brain loss or running code errors) and one for GCUT (preservation of almost all non-brain structures). GCUT’s failure was on subject “7_8” and was caused by erroneous choice of seed position, inside the neck region rather than in WM. Usually, the neck region is highly heterogeneous, which avoids this problem, but in this particular subject it had a very homogeneous appearance. Interestingly, both HWA and BSE failed on the same subject, the latter due to wrong selection of the largest connected component (the neck instead of the brain).

Motivated by the cortical thickness estimation problem, we have found it more informative to define failure quantitatively in terms of a brain mask with either moderately

Figure 6.13: ROC curves (false positive rate vs. true positive rate) for $k = 2.3$ (Table 6.8) for (a) data set 1 (b) data set 2 (c) data set 3 (d) data set 4

high FN rate ($FN > 0.1\%$) or moderately high adjusted FP rate ($FP_{adj} > 7\%$). Such masks are often unsuitable for cortical thickness measurements. By this definition, GCUT was more robust than HWA on data sets 1 (two failures for GCUT vs. three for HWA) and 2 (two failures for GCUT vs. six for HWA) but less robust on data set 3 (one failure for GCUT vs. none for HWA). On data set 4 the two approaches were similar (two failures each).

GCUT was found to be robust to the choice of parameters: the intensity threshold T that is used to obtain preliminary mask and parameter k that controls the contribution of voxel intensities in deciding cut positions. This follows, first of all, from the fact that using the same values resulted in excellent performance across all four data sets. Table 6.8 and Table 6.9 and the corresponding receiver operating characteristic (ROC) curves in Figure 6.13 and Figure 6.14 show that the best values are $T = 0.36I_{WM}$ and $k = 2.3$. Further, changes in the parameter values had limited influence on the overall performance.

Figure 6.14: ROC curves (false positive rate vs. true positive rate) for $T = 0.36I_{WM}$ (Table 6.9) for (a) data set 1 (b) data set 2 (c) data set 3 (d) data set 4

Changing T led to reciprocal changes in FN and FP rates - decreasing T led to reduction in FP rate and increase in FN rate as more voxels were lost in the preliminary mask (Table 6.8). Smaller k (higher influence of voxel intensities) led to lower FP rate but at the expense of larger FN rate, due to increased chance of cutting within the brain. Increasing k initially reduced but subsequently increased the FN rate, as a result of wrongly positioned cuts that were no longer guided by voxel intensities.

6.7 Discussion

Overall, in terms of similarity indices, FP and FN rates (both conventional and modified), our method (GCUT) performed better than all other approaches. GCUT was superior to BET according to all metrics on all data sets except the first one, where BET achieved somewhat better false positive rates and similarity indices at the expense of almost 2% false negatives compared to almost zero for our approach. Notably, the slight advantage

of BET was observed with only one data set out of four.

Similar statements can be made about BSE and WAT. BSE was inferior to GCUT on data sets 3-4 according to all metrics, and superior to GCUT in terms of false positives on data sets 1-2, at the expense of unacceptable brain loss of 6 – 27%. WAT was consistently superior to GCUT in terms of false positives, but again due to very high FN rate of 2.5 – 24%. WAT’s FN rate exceeded 12% on three data sets out of four.

Our approach was superior to HWA on all data sets and metrics, excluding FN rate, where the two approaches were roughly equivalent and achieved negligible brain loss much lower than that produced by BET, BSE, and WAT. On FP adjusted rate, which is most sensitive to preserved dura, GCUT achieved 10–30% reduction compared to HWA. Taking into account other factors as well as effect on the subsequent surface estimation performance, the new skull stripping approach GCUT offers a different balance of advantages and disadvantages compared to HWA.

Advantages

1. Cleaner skull strip with less remaining dura (at least 10-30% reduction).
2. Better robustness of output using legacy data.
3. Freedom from dependence on shape priors, suggesting possible deployment of this method to developmental studies on humans and for studies involving animal brains.
4. No need for alignment to standard space, which may fail for strongly misaligned brains.

Disadvantages

1. Trivial increase in brain loss for some data sets, which can still be considered negligible for practical applications.
2. Despite less dura preservation, subsequent segmentation can be more problematic.

A caveat regarding the current study is that performance analyses are based on data obtained from relatively healthy adults. Further work is advisable to determine if the ad-

Figure 6.15: Typical errors exhibited by HWA (middle) and GCUT (right). HWA is often confused by double boundary between scalp/dura/GM, resulting in inclusion of large chunks of skull/dura mater, rows 1-2. GCUT fails to cut connections where there is no noticeable intensity separation between dura and GM, row 3

vantages outlined here will generalize to data obtained in clinical settings where movement, tissue abnormalities, and artifacts could be problematic.

However, the main value of our approach is not its standalone performance but rather its effect on subsequent segmentation when used in conjunction with HWA. As illustrated in Section 6.6.2, the intersection of the two masks can solve the overestimation problem (11 out of 15 problematic cases in our study; see Figure 6.12). This effect can be attributed to GCUT's mask having complementary properties to that of HWA, as illustrated in Figure 6.15.

HWA's main problem appears to be the double boundary between the scalp/skull/dura/GM structures, often resulting in inclusion of large chunks of skull/dura mater that run parallel to brain surface. This happens because image intensity information and smoothness (or shape) constraints are combined in a single energy function [134]. In the presence of a double boundary, inclusion or exclusion of skull/dura mater results in similar mask shapes and equally dark boundary voxels, making it likely for the algorithm to choose the wrong mask, (top two rows of middle column in Figure 6.15). In GCUT, the intensity information is used to create a preliminary mask after which the shape constraints are imposed by performing cuts. As the dura mater connections with GM are irregular in the preliminary mask, they are likely to be cut, reducing the likelihood of dural inclusion with HWA (top two rows of right column in Figure 6.15).

On the other hand, HWA's ability to explicitly impose smoothness and shape constraints results in more regular brain mask shapes, which is particularly important when closely adherent skull/dura shares the same signal intensity as GM and becomes inseparable from GM (row 3 of Figure 6.15). In such cases GCUT would make a wrong cut, resulting in somewhat irregular mask shape. Intersecting the two masks appears to resolve both problems and significantly decrease the overestimation of the pial surface (Figure 6.12).

It is likely that multispectral segmentation techniques made more feasible by availability of fast 3D T2W imaging [155] will result in further improvement with respect to dura mater removal. An alternative strategy is to use multiple echoes to create intensity differences between dura adjacent to GM, particularly in the medial temporal region [13]. While this is clearly promising, analyzing the vast amount of legacy data available will still benefit from the technique we describe.

6.8 Conclusion

We proposed a new skull stripping approach that builds upon earlier work [108, 109], which used intensity thresholding followed by cutting of narrow connections to obtain a brain mask. Our approach consists of three steps: intensity thresholding, refining of the initial mask by cutting false connections between brain and non-brain structures, and

post-processing to recover CSF and partial volume voxels. Instead of using mathematical morphology for false connection removal [108, 109], a method that only cuts sufficiently narrow connections, our algorithm uses a superior graph cuts approach that is capable of locating precise cut positions of varying widths.

By itself, our approach offers a good alternative to HWA - it is more robust on legacy data, can work on a larger variety of brain shapes and achieves at least 10–30% reduction in residual dura without significant increase in brain tissue removal. The greatest benefit of using our approach is realized when it is employed in conjunction with HWA, for example by using a simple intersection of the two masks. The errors produced by the two masks are complementary, resulting in significant improvement of subsequent segmentation performance when the masks are combined. In our experiments, the combined application of the two techniques resulted in the successful segmentation of 11 out of 15 volumes that were not adequately segmented using HWA alone.

Chapter 7

Skull stripping of multimodal MR images

Skull stripping usually had been restricted to single modality images, i.e., mostly T1-weighted images due to their high white matter (WM)/grey matter (GM) contrast and difficulty in the acquisition of high resolution T2-weighted images, although there were few cases when T2W images had also been used because of their high gray matter/cerebrospinal fluid (CSF) contrast [109, 156]. There have been several early approaches to deal with brain tissue segmentation (separate estimate of WM, GM, CSF) using multimodal imaging, but to the best of our knowledge, these have never been seriously used in application and have never been extended to skull stripping. The most recent approaches to tissue segmentation also rely on single T1 modality. The main reason is that up until 2005, T2-weighted data had low resolution in one of the directions. While most T1-weighted images had resolution of $1.5 \times 1.5 \times 1.5$ mm or less, T2-weighted data had resolution of $1.5 \times 1.5 \times 5$, which was too coarse for brain segmentation. Starting from 2005, with the advancement in MR sequences and hence the paired acquisition of both T1W and T2W images in high resolution during the same session has created an opportunity to exploit the information from both images to improve brain segmentation in general and skull stripping in particular. In this work, the multimodal images referred to include T1W and T2W images.

The T1/T2 contrast of MR images depends on the relaxation times of each of the

tissues. WM appears bright and CSF appears dark in T1W images and dark WM and bright CSF characterize T2W images, see Figure 7.1. Refer to Section 2.1.1 for details on MR image contrast. Skull and/or dura removal around the posterior fossa and basal temporal regions can be problematic [144] and could benefit from additional contrast provided by 3D T2W imaging methods [155]. In this chapter, we propose a new method based on GCUT, that utilizes T2W image information to provide increased benefits for skull stripping particularly in removing additional dura attachments. We start with the problem of brain segmentation and show that a significant decrease in misclassification rate can be achieved by the use of T2W image information in addition to the T1W image information. We then discuss the problem of skull stripping on T1W-T2W combinations. We also discuss and compare the use of multiecho images with the multimodal (T1W and T2W) data.

Figure 7.1: (a) T1-weighted image and (b) T2-weighted image

7.1 Previous work

Skull stripping approaches that work on multimodal images instead of the usual single modality image exploit the property of T2W that helps in better identification of dura and the high WM/GM/CSF contrast ratio of T1W images. Some methods have been proposed that suggest skull stripping on T2W images only. For example, the algorithm by Atkins et al. [109] works on either T1W or T2W image to create the initial brain mask. The method uses thresholding of the original MR image intensities; the computation of the

threshold value depends on the image modality. Active contours algorithm is then applied on the initial brain mask to deform the contour to fit the brain boundary. The method by Rajagopalan et al. [156] applied an intensity transformation on the T2W image to enable it to be used by the standard skull stripping algorithms. The T2W image intensities are projected onto a standardized T1 intensity scale resulting in a T1W like image, which can then be processed by any standard skull stripping algorithm.

Alyassin et al. [157] used a combination of proton density weighted (PDW) image and T2W image. Thresholding is done using values determined from the image histogram. Helms et al. [158] used BET (refer to Section 6.1.2) on T1W and T2W images for skull stripping. Registration of the images was accomplished by FLIRT (refer to Section 7.4). Then, FMRIB advanced segmentation tool (FAST) was applied on T1W and T2W to segment the brain into individual tissue classes. Finally, the T2W brain mask is applied to the T1W image to get the final output.

Recently, a new 4-echo MEMPRAGE sequence [13] has been proposed to provide better delineation between the dura and CSF by varying T2 weighting in the echoes. Thresholding was performed on the ratio image of the first and last echoes.

In this chapter, we propose a skull stripping algorithm for multimodal images (T1W and T2W images), which is an extension of GCUT proposed in Chapter 6. Before specifying the details of the algorithm, we evaluate the usefulness of this combination for general brain tissue estimation. We also evaluate the potential increase in dura cluster separability when a combination is involved. We also compare this T1W-T2W combination with MEMPRAGE.

7.2 Multimodal brain segmentation

Brain segmentation is the identification and separation of individual tissues of interest (WM, GM, CSF, etc.) in a brain image. What makes the problem difficult is that on most imaging modalities, i.e., T1W or T2W, there is substantial overlap between tissue intensity distributions. Figure 7.2 and 7.3 show the intensity distributions of WM, GM, and CSF for T1W, T2W images and their combination. The individual tissues were delineated based on the available ground truth. Since the intensities of ventricular CSF and external CSF

voxels are almost the same and since the ground truth contains only ventricles and not the external CSF, ventricular CSF was chosen to represent CSF for convenience and simplicity. There is substantial overlap between tissue intensity distributions when only a single modality is used (T1W or T2W) (Figure 7.3), but the clusters are more separable when the modalities are combined (Figure 7.2). A random selection of 1000 voxels from each tissue class is used to obtain the intensity distribution. The problem of brain segmentation on the combination of two modalities can be viewed as a pattern classification problem where suitable techniques can be used to determine decision boundaries to separate the three classes.

Figure 7.2: Intensity distributions of T1W and T2W images in Figure 7.1 and their combination

Table 7.1: Confusion matrices for unimodal-only T1W (left) and multimodal-T1W and T2W (right) brain segmentation

only T1W		Actual value		
		WM	GM	CSF
Predicted value	WM	362933	29589	298
	GM	39305	429951	4120
	CSF	19	56369	17722
Misclassification rate = 13.75%				

T1W and T2W		Actual value		
		WM	GM	CSF
Predicted value	WM	355429	22858	274
	GM	45644	476454	2983
	CSF	852	16103	18875
Misclassification rate = 9.45%

Figure 7.3: Histograms of the (a) T1W intensity distribution (b) T2W intensity distribution

Figure 7.4: (a) Thresholding based only on T1W image intensity information (b) Thresholding based on T1W and T2W image intensity information

If only T1W image information is present, then the intensity distribution reduces to a projection profile of the intensities along the x-axis. The simplest method is to use multiple thresholds to isolate the classes, which is equivalent to two vertical decision boundaries (Figure 7.4(a)). The misclassification rate obtained for a selected set of images from the data set (explained in Section 7.6) is shown in Table 7.2.

When T2W image is available, this additional information can be utilized to determine decision boundaries that produce better separation of the classes with reduced misclassi-

Table 7.2: Multimodal brain segmentation misclassification rate

Image	only T1W (%)	T1W and T2W (%)
1	14.26	8.30
2	13.83	9.69
3	13.74	9.73
4	12.73	10.62
5	13.44	8.29
6	14.47	10.08

fication rate. The simplest of the pattern classification techniques would be to project the voxels to a line that maximizes the separation between the classes as in linear discriminant analysis (LDA). An example of the decision boundaries obtained using T1W and T2W images is shown in Figure 7.4(b). Though accurate separation of the classes is very difficult, the use of T2W and an appropriate decision boundary can help reduce the misclassification rate by about 30%, see confusion matrices showing the classification of voxels between the three classes (WM, GM, and CSF) in Table 7.1 and the misclassification rates in Table 7.2.

7.2.1 Brain tissue segmentation using multiecho MPRAGE

The multiecho MPRAGE (MEMPRAGE) was developed with an intention to have more contrast between the tissues of interest and hence better separation of the classes (Section 2.1.3). MEMPRAGE has a higher bandwidth and less distortion, while achieving good signal to noise ratio (SNR) [13]. The MEMPRAGE data consist of four echoes. Each progressive echo contains a larger weighting of T2* contrast. Note that T2* contrast weighting is not completely equivalent to T2 contrast (refer to Section 2.1.1), so one can expect the segmentation based on MEMPRAGE to produce results that are different from those produced by a combination of T1W and T2W images.

Figure 7.5 shows an example of the images of first and last echo of MEMPRAGE sequence. The difference in intensities of the tissues between the echoes is very small and this is clearly reflected in the distribution of intensities in Figure 7.6. The intensity points are aligned along a line with slope slightly less than 1, highlighting that the second echo is slightly darker than the first. This means that there is little advantage in using

Figure 7.5: Multiecho MPRAGE (a) first echo (b) last echo

Figure 7.6: Distribution of intensities of first and last echo of MEMPRAGE

this combination for segmentation of main brain tissues. However, the main advantage of MEMPRAGE is the substantial darkening of the skull and some structures immediately neighboring the outer surface of GM, which is visible in Figure 7.5. This is not reflected in the intensity distribution because this structure is not a part of three main brain tissues, i.e., WM, GM, and CSF.

7.3 Multimodal dura segmentation

Skull stripping aims to segment the MR brain image into two classes, the brain (WM and GM) and the non-brain, and can be considered as a two class separation problem compared to the three class (WM, GM, and CSF) separation problem of brain tissue segmentation. As we have explained in Chapter 6, successful skull stripping is dependent on the ability to differentiate (and eliminate) dura from GM. Elimination of the dura and CSF voxels with the least effect on the brain voxels is a difficult task because of their varied intensities across locations and the absence of a clear boundary separating them from other brain tissues. The partial volume voxels between the dura and GM and between the CSF and GM pose a serious problem for skull stripping. WM is completely enclosed by GM and hence does not affect the determination of the decision boundary. Therefore, we only need to consider dura-CSF and GM voxels to separate the two classes. The dura-CSF layer appears as a two or three voxel thick layer immediately neighboring the outside surface of the GM. Here, we defined three layers of voxels outside the ground truth brain mask as dura-CSF region. In order to have an equal number of voxels of GM, we considered only three outer layers of GM for classification. The dura-CSF layer was obtained by dilating the ground truth image by three layers and the adjacent GM layer was obtained by eroding the ground truth by an equal number of iterations to get approximately equal number of dura-CSF and GM voxels.

Dura-CSF region is heterogeneous and consists of two distinct clusters (dura and CSF) and also partial volume voxels that are formed between these two classes and GM. Figure 7.7(a) shows the full distribution of the WM, GM, and dura-CSF tissue intensities of T1W-T2W combination and Figure 7.7(b) gives the distribution of dura-CSF and its approximate division into three classes, CSF (yellow), dura (cyan), and partial volume (green). Figure 7.8 shows the positions of the voxels belonging to dura and CSF clusters, overlaid on the T1W image. Due to heterogeneous properties of dura-CSF region, it is unlikely that one single decision boundary can separate it from WM and GM. Hence, in what follows, we discuss how each component of dura-CSF can be individually separated from the brain.

- *CSF*: CSF voxels have darker T1W intensity and brighter T2W intensity. Partial

Figure 7.7: Intensity distribution of T1W and T2W images in Figure 7.1 (a) plot of WM, GM, and dura-CSF (b) plot of only dura-CSF class

isolation of these pure CSF voxels from GM and WM using only T1W intensity information can be done by using a vertical decision boundary. If both T1W and T2W intensities are available, then suitable pattern classification methods like linear discriminant analysis (LDA) and support vector machines (SVM) can be used. The resulting decision boundaries (see Figure 7.9) and the misclassification rate (see Table 7.3) obtained in each case state that the vertical decision boundary provides better classification than that of LDA and SVM. There is a decrease of about 10 – 12% in the misclassification rate of vertical decision boundary compared to SVM and LDA. We found that LDA and SVM result in similar decision boundaries and their relatively poorer performance (compared to vertical decision boundary) can be attributed to their working characteristics and the strong overlap of voxels with the GM class. LDA works on the basis of maximizing the variance between the two classes and it assumes that the variables are normally distributed which is not true in the case of multimodal data. So we cannot expect LDA solution to minimize the misclassification rate. SVM locates a hyperplane that separates the training points with the maximum distance. It is usually very effective when the number of data points is small and dimensionality is large. But in our case, we only have

Figure 7.8: Typical results of dura-CSF overlay on the original T1W image (a) Original image (b) dura overlay (c) CSF overlay

two dimensions and hundreds of thousands of data points. Moreover, our classes are highly overlapping, which poses another difficulty for SVM. SVM classifier also does not guarantee minimum misclassification rate. Overall, we can conclude that using T2W modality in addition to T1W does not help to separate the pure CSF voxels from the brain.

Table 7.3: Multimodal skull stripping misclassification rate for CSF vs GM+WM separation on the intensity distribution in Figure 7.9

Image	only T1W (%)	LDA (%)	SVM (%)
1	4.18	4.80	4.84
2	4.40	5.19	5.09
3	3.93	4.34	4.05
4	3.50	3.87	3.81
5	3.69	4.24	4.27
6	3.66	4.08	3.87

- *Dura*: Dura voxels in general are darker in both T1W and T2W images and are concentrated in the lower left quadrangle of the intensity distribution. Their intensities do not have uniform characteristics like that of CSF or WM. Some of the dura voxels appear as dark as T1W CSF voxels while others are as bright as T1W WM voxels. From Figure 7.7(b), it is apparent that the dura class cannot be separated from the brain using only T1W intensity information or only T2W information. The intensity range of these voxels overlap strongly with WM in T2W intensity scale and with GM in the T1W intensity scale. Therefore, vertical or horizontal decision boundary

Figure 7.9: Classification of CSF voxels in the intensity distribution of Figure 7.7 (b)

will not help in the classification and will necessitate the combined use of T1W and T2W intensity information. Figure 7.10 shows an example of an empirically found “good” decision boundary that separates the dura voxels.

- *Partial volume voxels*: This class of voxels comprises mostly the partial volume dura voxels that border the GM and WM. Apart from that, it also comprises brighter dura voxels. The intensity profile of these voxels completely overlaps with GM and WM, which makes the separation based on T1W and T2W practically impossible.

Table 7.4: Multimodal skull stripping misclassification rate for the thresholding on the intensity distribution in Figure 7.11

Image	only T1W (%)	T1W and T2W (%)
1	7.54	5.76
2	8.33	6.57
3	7.17	5.99
4	7.22	5.55
5	7.45	5.72
6	6.75	5.64

We can conclude that only two out of three classes of dura-CSF voxels can be separated from the brain. From Table 7.3, we have shown that the vertical decision boundary is

Figure 7.10: Classification of dura voxels in the intensity distribution of Figure 7.7 (b)

sufficient for CSF/brain separation. Dura can be separated by a slanted decision boundary as in Figure 7.10. These two boundaries can be simply combined to separate dura-CSF from GM-WM, as shown in Figure 7.11(b).

Table 7.4 compares the misclassification rates obtained by using decision boundaries based only on T1W information and the combination of T1W and T2W information, see Figure 7.11. The combined use of T1W and T2W information helps in achieving better misclassification rate. It results in about 20% reduction of misclassification rate compared to using only T1W information. Note that the difference in the misclassification rates between Table 7.3 and Table 7.4 is because of the difference in the brain mask considered. The former uses a conservative brain mask, i.e., three layer of each of the tissues, whereas the latter uses a full brain mask.

7.3.1 Skull stripping of multiecho MPRAGE

In the case of MEMPRAGE sequences, [13] had suggested separating dura from GM by thresholding the ratio of intensities of the first and last echoes. The threshold for this segmentation was chosen at two standard deviations above the mean distribution of voxel intensity ratios. This is equivalent to a linear decision boundary that is constrained

Figure 7.11: (a) Segmentation based only on T1W image intensity information (b) segmentation based on T1W and T2W image intensity information

to go through the center of coordinates, see Figure 7.14(b). Our goal is to provide an interpretation of this technique and compare it with the segmentation using only T1W threshold.

Figure 7.13 shows the overlay of dura and CSF segmented using van der Kouwe threshold [13]. This threshold provides good segmentation because it separates dura voxels (cyan in Figure 7.12(b)) from the rest. On the other hand, pure CSF voxels (yellow) can be separated based on T1W threshold.

Unlike the T1W threshold of multimodal images, T1W threshold for MEMPRAGE is a slanted line. This represents the T1W intensity scale which is along the diagonal line since T1W image is obtained by averaging all the echoes. Figure 7.14 shows the intensity distributions and compares three different decision boundaries, ratio threshold using T1W threshold, van der Kouwe threshold, and the new threshold.

Though a good amount of dura elimination is achieved by using van der Kouwe threshold, it can be seen that this is achieved only at the cost of erosion of significant amount of GM and WM voxels. A better decision boundary can be obtained by slightly relaxing the condition (ratio < 1.2). This provides better classification than van der Kouwe threshold which can be seen from the quantitative results in Table 7.5. The new threshold gives

Figure 7.12: Intensity distributions of first and last echoes in Figure 7.5 (a) plot of WM, GM, and dura-CSF (b) plot of only dura-CSF voxels

about 50% reduction in misclassification rate compared to the van der Kouwe threshold.

Table 7.5: Skull stripping of MEMPRAGE: Misclassification rate

Image	only T1W (%)	van der Kouwe threshold (%)	New threshold (%)
1	7.70	12.09	6.77
2	7.67	14.31	7.39
3	7.49	13.82	7.25
4	7.59	12.50	7.13
5	7.47	12.34	6.96
6	7.70	17.07	7.06

When comparing multimodal vs. multiecho performance of dura-CSF vs. GM-WM classification, multiecho performs slightly worse (see Tables 7.15 and 7.4). We can see from the results in Figure 7.15 that dura delineation on multiecho images leads to erosion of a significant portion (often more than 3-4 layers) of GM along with the dura voxels. Since most of this erosion occurs in the boundary voxels, the post-processing procedure might not be able to recover these voxels. This leads us to conclude that using multiecho images is not effective in “safe” dura elimination, i.e., the one that leads to no brain erosion, when compared with multimodal images. Hence, in the sections that follow, we consider the use of only multimodal images (T1W and T2W) for skull stripping.

Figure 7.13: Typical results of dura-CSF overlay on the first echo of MEMPRAGE (a) original image (b) dura overlay (c) CSF overlay

7.4 Main issues in multimodal skull stripping

We concluded from Section 7.3 that multimodal skull stripping is advantageous over using only single modality images or multiecho images, particularly in the case of dura removal. Before using multimodal images for skull stripping, we have to address certain issues that are associated with the implementation of this approach.

1. **Alignment:** The most important of the issues concerned with the use of multimodal images is their alignment. T1W and T2W images are usually misaligned because of the time gap between their acquisitions. The images can also be locally distorted due to imperfect shimming in the regions where magnetic susceptibility changes rapidly leading to non-linear distortion. Depending on the nature of the misalignment, linear or non-linear registration techniques are necessary to align the images. In order to accomplish registration of two images, in our work we use FMRIB software library (FSL). FSL contains two tools that can be used for registration of images of different modalities, FLIRT (FMRIB's Linear Image Registration Tool) [159] and FNIRT (FMRIB's Non-linear Image Registration Tool) [160].

FLIRT: FMRIB's linear image registration tool performs affine registration of 3D images of different modalities using a global optimization method. It can be used with a number of different geometric transformation models (degrees of freedom) such as rigid body, affine, etc. Intensity based cost functions such as correlation ratio and mutual information are used for inter-modal registration and least squares

Figure 7.14: Thresholding based on ratio of first and last echo intensity information (a) only T1W threshold (b) van der Kouwe threshold (c) new ratio threshold

and normalized correlation are used for intra-modal registration [159,161,162]. These intensity-driven approaches try to match intensity patterns between the input image and the reference image by minimizing the cost function. The output of FLIRT is a linear transformation matrix that can be applied on the input image to align it with the reference image. An improvement to the FLIRT (addition of rigidity constraints) can be found in [163].

FNIRT: - FMRIB's non-linear image registration tool performs registration of images which have undergone non-linear deformations. It uses sum of squared differences as the cost function which probably limits the use of FNIRT to only

Figure 7.15: Typical results of dura elimination in multiecho and multimodal images (a) original image slices (b) dura layer classified from MEMPRAGE overlaid on original image (c) dura layer classified from multimodal skull stripping overlaid on original image

intra-modality images. The optimum of the function is determined by Levenberg-Marquardt modification of the Gauss-Newton method. The result of FNIRT is dependent on a number of parameters and the use of the right value of these parameters requires expert knowledge on non-linear registration. The dependency of the value of one parameter on another parameter increases the complexity of choosing the right value of the parameters [160]. It also requires FLIRT to be performed first and the resultant affine transformation matrix be specified as one of inputs to the FNIRT procedure. The method is very sensitive to the change in the parameter values.

2. **Multiple parameters:** Another issue when performing thresholding on multimodal images is the problem of handling multiple parameters. The combined use of T1W and T2W images for skull stripping gives potential benefits at the cost of having two additional parameters. These parameters define the decision boundaries that separate the dura and CSF class from GM and WM classes based on both T1W and T2W intensity information. In order to attain better dura elimination, thresholding has to be done based on these decision boundaries (see Figure 7.11(b)).

7.5 Proposed approach

We propose to solve the problem of skull stripping and dura elimination in multimodal images using a two-step approach where we first determine an initial mask and then refine it in the next step. This approach (GCUT_T1T2) is similar to GCUT for monomodal images (GCUT_T1²) and consists of the following group of operations (see Figure 7.16).

- Classification using 2D decision boundary to obtain preliminary mask
- Removal of narrow connections using graph cuts
- Post-processing

Before proceeding with the skull stripping procedure, the images have to be corrected for any misalignment. We chose FLIRT over FNIRT because of its suitability to work on

²GCUT used in Chapter 6 and GCUT_T1 refer to the same method. The new notation is used to highlight the difference between monomodal and multimodal skull stripping

Figure 7.16: Pipeline of proposed multimodal skull stripping approach

multimodal images. We consider T1W image as reference and use FLIRT to compute the affine transformation matrix that aligns T2W image to T1W image. This affine transformation matrix is then used to transform T2W image and align it with the reference.

7.5.1 Obtaining preliminary mask

The preliminary mask is obtained by thresholding the registered input brain images by a suitable threshold. We had previously found that for GCUT_T1, a threshold $T = 36\%I_{WM}$ performed well in removing the right amount of CSF and dura, preserving as much brain as possible and creating narrow connections between the brain and non-brain tissues. Refer to Section 6.3.1 for details on threshold selection and creation of preliminary brain mask for GCUT_T1. For GCUT_T1T2, we use the 2D decision boundary proposed in 7.3, that helps to separate additional dura with the help of both T1W and T2W intensity information.

Recall that a single parameter T does not eliminate all the dura (see intensity distribution and output image in Figure 7.11(a)) requiring multiple parameters based on T1W and T2W intensity information to be used to remove additional dura (see Figure 7.11(b)). From the intensity distributions of the T1W and T2W intensities, it follows that the complete removal of dura without loss of brain tissues is impossible because of the strong overlap of some of the dura voxels with the GM class or their proximity to the brain voxels. These constitute the partial volume voxels occurring between the GM and the dura layer. Since GCUT includes a post-processing procedure to reinstate partial

volume voxels, the threshold levels can be relaxed slightly to exclude these partial volume voxels.

Figure 7.17: Intensity distribution and decision boundaries used for thresholding

We accomplish additional dura removal by using a decision boundary as in Figure 7.11(b). We empirically defined this decision boundary as a straight line with a slope equal to -1 which is at a perpendicular distance d_{WM} from the mean white matter intensity (I_{WM1}, I_{WM2}) , see Figure 7.17. The equation of the decision boundary can be determined by using simple analytical geometry as

$$I_1 + I_2 - I_{WM1} - I_{WM2} + \sqrt{2}d_{WM} = 0 \quad (7.1)$$

where I_1, I_2 denote the T1W and T2W intensities and I_{WM1}, I_{WM2} denote the mean white matter intensities of T1W and T2W images respectively.

The classification procedure of GCUT_T1T2 therefore will have two parameters, T and d_{WM} , each of which can be optimized. It has to be noted that the classification using T1W and T2W intensity information is equivalent to thresholding using only T1W information when d_{WM} is sufficiently large. Therefore, T and d_{WM} have to properly chosen so as to exploit the benefits of T2W intensity information. The result of thresholding gives a preliminary brain mask after dura erosion with sufficiently narrow connections and

acceptable brain loss that can be compensated for during post-processing.

Figure 7.18: ROC curve for $T = 0.36I_{WM1}$ and different values of d_{WM}

For the T1W and T2W images in our data set, we plotted the receiver operating characteristic (ROC) curves (see Figure 7.18) and found that a good classification can be obtained by using $T = 0.36I_{WM1}$ and $d_{WM} = 0.28I_{WM1}$ and chose these values for our experiments. The mean white matter intensities of T1W and T2W images, I_{WM1} and I_{WM2} are obtained by averaging the intensities within the WM seed voxels of the corresponding images. Refer to Section 5.4 for details on seed selection.

Shown in Figure 7.19 are the typical results of thresholding (THRES_T1 and THRES_T1T2) on monomodal and multimodal images. Compared to thresholding on only T1W image, the additional use of T2W intensity information helps in eliminating substantial amount of dura and other non-brain regions around the brain stem. THRES_T1T2 also results in narrower connections than those obtained by THRES_T1 thus promising improvement in the subsequent graph cut procedure.

7.5.2 Removal of narrow connections

Assuming that the preliminary brain mask has only narrow connections between the brain and non-brain, we use graph cuts to remove these connections. However, despite availability of both T1 and T2 contrasts, we have decided to rely only the T1W intensity information. We found that T1W intensity information by itself was sufficient to break these connections and used (6.7) as in GCUT_T1 for edge weight assignment. However, complex functions utilizing both T1W and T2W intensity information can be designed

Figure 7.19: Thresholding using multimodal (T1W and T2W) information (bottom row) results in additional dura elimination and narrower connections compared to thresholding using only monomodal (T1W) information (middle row)

and employed to possibly produce better results.

7.5.3 Post-processing

The presence of partial volume voxels poses a serious problem in the segmentation. This is compounded in the case of multimodal skull stripping because of the possible misalignment between T1W and T2W images. Though FLIRT registration corrects the misalignment to a certain extent, there could be some partial volume voxels that could have been removed during the thresholding procedure. In order to reinstate those partial volume voxels, we

perform morphological closing operation (10mm voxel dilation and 10mm voxel erosion, sizes rounded to the nearest integer) on the final mask followed by an addition of a layer of voxels at the cuts and a layer of “dark” voxels elsewhere. The “dark” voxels were defined as those outside the decision boundaries defined by the parameters T and d_{WM} . This procedure fills in the ventricles and holes and creates a smooth mask (see Figure 7.16).

7.6 Data set

We performed experiments on the data set containing images of 13 healthy subjects obtained on Siemens Tim Trio 3T scanner using the following parameters: T1W MPRAGE sequence: TR = 2530 ms, FA = 7 degrees, TI = 1200 ms, multiecho TE ranging from 1.64 to 7.22 ms. T2W sequence: TR = 3390 ms, TE = 388 ms, turbo factor = 115, resolution $1 \times 1 \times 1$ mm. The ground truth for the data set were obtained by processing the input T1W images using FreeSurfer 3.04 [127] and converting the pial surfaces into volume masks.

7.7 Performance evaluation

In order to compare the performance of our approach we used the metrics such as Jaccard similarity index (JS), Dice similarity index (DS), false negative (FN) rate, false positive (FP) rate, and false positive adjusted rate (FP adj). Refer to Section 6.5 and Appendix A for details on the metrics used for performance evaluation. We tested our proposed multimodal skull stripping approach GCUT_T1T2 on the images from the data set described in Section 7.6.

Table 7.6: Comparison of monomodal and multimodal skull stripping approaches

Method	DS (w/o dark pixels)	JS (w/o dark voxels)	FN (%)	FP (w/o dark voxels, %)	FP_adj (%)
	Mean (SD) [range]	Mean (SD) [range]	Mean (SD) [range]	Mean (SD) [range]	Mean (SD) [range]
GCUT_T1	0.87 (0.01) [0.86 - 0.88]	0.77 (0.01) [0.75 - 0.79]	0.006 (0.01) [0 - 0.03]	21.74 (1.97) [19.36 - 25.36]	4.02 (0.57) [3.24 - 5.25]
GCUT_T1T2	0.87 (0.01) [0.86 - 0.88]	0.77 (0.01) [0.75 - 0.79]	0.02 (0.03) [0.0001 - 0.12]	20.99 (1.65) [18.52 - 24.17]	3.69 (0.51) [2.90 - 4.61]

7.7.1 Quantitative performance evaluation

We compared GCUT_T1 and GCUT_T1T2, see results in Table 7.6. Contrary to our expectations, GCUT_T1T2 did not result in a substantial improvement in performance over GCUT_T1. Considering only the similarity indices, both the methods perform equally well. They provide a good trade-off between FN and FP resulting in similar JS and DS values. GCUT_T1T2 leads to a small decrease in the FP rate with a slight increase in the FN rate. Comparing false positive adjusted rate, GCUT_T1 resulted in a reduction of about 10% in dura elimination with a little but negligible brain loss. We consider a method to have failed if the resultant brain mask has a high FN rate ($FN > 0.1\%$). Both the methods are within the safe range of FN rate.

The results in Table 7.6 compare the two methods numerically. The subtle difference between the two methods (GCUT_T1T2 is expected to perform significantly better than GCUT_T1) can be attributed to insufficient accuracy of the boundary of the ground truth images which in some cases can preserve few layers of dura voxels located immediately outside the brain surface. Details regarding the ground truth and qualitative evaluation measures are explained in the following section.

7.7.2 Qualitative performance evaluation

Qualitative evaluation of skull stripping results has shown that some increase in FN rate following the proposed approach might not reflect brain erosion but can simply be due to imperfection in the ground truth. As shown in Figure 7.20(a) and (b), ground truth may occasionally include parts of dura that are indistinguishable from GM. In this case, the correct segmentation following the proposed skull stripping approach might be improperly identified as brain erosion.

This allows us to conclude that the quantitative results are not sufficiently reliable to draw conclusions regarding potential benefits of the proposed approach. Shown in Figure 7.21 are the typical results of skull stripping (GCUT_T1 and GCUT_T1T2) on monomodal and multimodal images. In these examples, multimodal skull stripping removes more dura and results in a cleaner skull strip than monomodal skull stripping. The increased performance is greatly due to the thresholding procedure that helped in

Figure 7.20: Inaccuracy in the ground truth (a) original image slices (b) ground truth overlaid on the original image

eliminating more dura and creating narrower connections. However, there are some dura attachments that are as bright as GM and do not have marked boundary between them. Even with the help of multimodal intensity information, these attachments are very difficult to be removed without potential loss of GM tissues.

7.8 Discussion

We proposed a new skull stripping approach for multimodal skull stripping. The use of T2W modality in addition to the T1W image contributes to differentiating dura from GM and producing cleaner brain. Current limitations of GCUT_T1T2 include the use of

Figure 7.21: Additional dura elimination in multimodal skull stripping (bottom row) compared to monomodal skull stripping (middle row)

a simple decision boundary to separate the brain and the non-brain classes and an edge weight assignment function based only on T1W intensity information. It also remains likely that some dura attachments will not be differentiated because of their close proximity to GM and strong overlap of intensities between the dura and the GM/WM.

Future directions for multimodal skull stripping might involve designing complex decision boundaries with the help of pattern recognition techniques and formulating new function for edge weight assignment for graph cuts that utilizes information from both the modalities. Further, a new post-processing procedure could be developed that handles partial volume voxels more effectively.

Chapter 8

Conclusions and future directions

In this thesis, we have proposed the use of retrospective techniques for the segmentation of MR brain images where the segmentation procedure is formulated as a minimization of an energy function. The segmentation is benefitted by the use of these techniques that divide the procedure in to two steps, first of which provides the initial segmentation and the second step incorporates the prior information. This division leads to simpler optimization, which otherwise becomes very complicated on the introduction of the constraints. We have showed that the retrospective procedure provides superior performance compared to standard segmentation.

In the case of detection of activations in fMRI, the retrospective cluster size thresholding approach allows the regions to grow before they are eliminated based on the cluster size. This approach finds the best modification and achieves superior performance to that of the standard MRF-based segmentation. Further development in this direction will focus on extending the algorithm to 3D, automating the selection of cluster size, and deciding the stopping criterion.

We have provided a theoretical analysis and comparison of performance of the existing methods for narrow connection removal. We have also provided generalized version of the isoperimetric algorithm and graph cuts based algorithm for cutting narrow connections. Though the analysis was performed on simple model images, it provides enough knowledge to choose the right method depending on the application. Further studies will involve complex shapes and sizes for analysis.

We have proposed a novel skull stripping technique using graph cuts for T1-weighted MR images. The algorithm provides at least 10-30% reduction in dura attachments when compared to the hybrid watershed algorithm (HWA) without significant loss of brain tissues. The algorithm is independent of shape priors and orientation and can reduce pial surface overestimation in FreeSurfer segmentation. The algorithm has several advantages that it has been incorporated in to FreeSurfer 5.0.0 under the name `mri_gcut`. The limitation of the algorithm is that it can result in trivial increase in brain erosion for some data sets and the subsequent segmentation can be problematic. Despite its limitation, it provides a good alternative to HWA. The algorithm has so far been tested only on data from healthy adults. Further studies are required to determine if the advantages outlined here will generalize to data obtained in clinical settings where movement, tissue abnormalities, and artifacts could be problematic.

We have also proposed a graph cuts based skull stripping algorithm for multimodal images utilizing both T1W and T2W intensity information. Though the results do not show improvement in false negative rate, it clearly shows a marked reduction in the dura attachments. Future studies are required to design algorithms which uses non-linear decision boundaries determined using advanced machine learning techniques to obtain the preliminary mask.

One of the problems in brain segmentation that can use the retrospective framework is the partial volume estimation of the brain. The first step is the initial discrete segmentation. To estimate the partial volume effect, only pixels in the boundary need to be considered. Statistical models can be used to estimate the volume where the cues come from the intensity information. The proposed skull stripping algorithms can be further improved to include estimation of partial volume voxels. Another potential problem where retrospective approach can be employed is brain tissue segmentation. Overall, a promising direction for future studies would be to identify characteristic graph based representation of images that can benefit from the proposed retrospective segmentation framework.

Appendix A

Performance evaluation metrics

We compute several metrics to evaluate and compare the performance of different segmentation techniques. These metrics give a numerical value of the matching between the ground truth and the segmentation output. Expert segmented result (M) is taken as the ground truth (gold standard) and the segmentation result (N) is then compared with M. The different metrics we consider for comparison are false positive rate (FP), false negative rate (FN), Jaccard similarity coefficient (JS), and Dice coefficient (DS).

False positive rate (type I error)

Consider a binary hypothesis test with null hypothesis H_0 and the alternative hypothesis H_1 . The possible outcomes of the hypothesis test can be put together as shown in Table A.1.

Table A.1: All possible outcomes of a hypothesis test

		Actual condition	
		True	False
Test result	Positive	True positive (TP) (correct result)	False positive (FP - type I error) (wrong result)
	Negative	False negative (FN - type II error) (wrong result)	True negative (TN) (correct result)

False positives (FP), also known as type I error or false alarm is the error of rejecting a null hypothesis when it is actually true. In other words, it is the error of accepting

an alternative hypothesis that should not have been accepted. Mathematically, it is the proportion of negative instances that were erroneously reported as being positive.

$$\text{False positive rate} = \frac{\text{Number of false positives}}{\text{Number of negative instances}} \quad (\text{A.1})$$

Usually, the type I error is set equal to 5%. This is termed as the level of significance. Increasing the type I error rate minimizes the type II error rate and vice versa.

False negative rate (type II error)

False negatives (FN), also known as type II error or miss detection is the error of accepting a null hypothesis when the alternative hypothesis is true. In other words, it is the error of rejecting an alternative hypothesis that should not have been rejected. It is the proportion of positive instances that were erroneously reported as being negative.

$$\text{False negative rate} = \frac{\text{Number of false negatives}}{\text{Number of positive instances}} \quad (\text{A.2})$$

Jaccard similarity coefficient

The Jaccard similarity coefficient (JS) is a statistic used to measure similarity between two samples. It is given as the ratio of the intersection between the two sets to the union of the two sets.

$$JS = \frac{|M \cap N|}{|M \cup N|} \quad (\text{A.3})$$

A JS value of 1 indicates perfect overlap (100% similar) and the index value of 0 indicates no overlap. In terms of FP, FN, and TP, JS can be written as $\frac{TP}{TP+FN+FP}$.

Dice similarity coefficient

Similar to Jaccard similarity coefficient, the Dice similarity coefficient (DS) also measures the agreement between two sets. DS of two sets is twice the number of the common elements divided by the sum of the two sets.

$$DS = \frac{2|M \cap N|}{|M| + |N|} \quad (\text{A.4})$$

A DS value of 1 indicates perfect overlap (100% similar) and the index value of 0 indicates no overlap. DS is related to JS by $DS = \frac{2JS}{1+JS}$. In terms of FP, FN, and TP, DS can also be written as $\frac{2TP}{2TP+FP+FN}$.

Appendix B

Lemmas

Lemma 3. *The isoperimetric ratio of a cut within a circular region of radius r is given by*

$$c = \frac{4 \sin \frac{\theta}{2}}{r(\theta - \sin \theta)}$$

where $0 < \theta < \pi$ is the angle subtended by the cut (chord) at the center of the circle. The cut with the smallest isoperimetric ratio occurs at $\theta = \pi$, i.e., along the diameter of the circle.

Proof. The isoperimetric ratio c is the ratio of the length of the cut (chord) to the area of the circular segment that is cut off by the intersection.

$$c = \frac{AB}{\text{Area of circular segment (shaded region)}}$$

$$c = \frac{AB}{\text{Area of sector } OAB - \text{Area of triangle } OAB}$$

Length of the chord $AB = 2r \sin \frac{\theta}{2} = r\sqrt{2 - 2 \cos \theta}$.

Area of sector $OAB = r^2 \frac{\theta}{2}$.

Area of triangle $OAB = 2 \frac{1}{2} r \sin \frac{\theta}{2} r \cos \frac{\theta}{2} = r^2 \sin \frac{\theta}{2} \cos \frac{\theta}{2} = r^2 \frac{\sin \theta}{2}$

Area of the circular segment (shaded region) is therefore

$$\frac{r^2}{2}(\theta - \sin \theta)$$

Figure B.1: Minimum isoperimetric ratio cut - illustration

The isoperimetric ratio c is

$$c = \frac{2r \sin \frac{\theta}{2}}{\frac{r^2}{2}(\theta - \sin \theta)} \Rightarrow c = \frac{4 \sin \frac{\theta}{2}}{r(\theta - \sin \theta)} \quad (\text{B.1})$$

Taking the derivative of c , we get

$$\frac{\partial c}{\partial \theta} = \frac{4}{r} \left(\frac{\frac{\theta}{2} \cos \frac{\theta}{2} - \frac{1}{2} \sin \theta \cos \frac{\theta}{2} - \sin \frac{\theta}{2} + \sin \frac{\theta}{2} \cos \theta}{(\theta - \sin \theta)^2} \right)$$

Figure B.2: Plot of the function in (B.1)

The derivative is negative and non-zero within the range $(0, \pi)$. The function has the minimum at $\theta = \pi$ in this range, see Figure B.2. This proves that the cut occurring along

the diameter of the circle has the smallest isoperimetric ratio. ■

Lemma 4. Consider a circular region R_1 of radius r_1 connected to another circular region R_2 of radius r_2 by a bridge of width w . Then, (i) if $r_2 < r_1$ or $S_2 < \pi r_1^2$, cut with the minimum isoperimetric ratio occurs either along the connection or along the diameter of R_1 (ii) if $r_2 > r_1$ or $S_2 > \pi r_1^2$, no cut with the minimum isoperimetric ratio can occur within R_1

Figure B.3: Minimum isoperimetric ratio cut (cut within larger region) - illustration

Proof. Consider the circular regions and a cut as shown in Figure B.3(a). The cut creates two partitions π_1 and π_2 and subtends an angle θ ($0 < \theta < \pi$) with the center of the circle. The width of the cut π_{12} is equal to

$$\pi_{12} = 2r_1 \sin \frac{\theta}{2}$$

The areas of the partitions are given by

$$S(\pi_1) = \pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta)$$

$$S(\pi_2) = \pi r_2^2 + \frac{r_1^2}{2}(\theta - \sin \theta)$$

The isoperimetric ratio of this cut can be given by

$$c = \frac{2r_1 \sin \frac{\theta}{2}}{\min \left(\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta), S_2 + \frac{r_1^2}{2}(\theta - \sin \theta) \right)} \quad (\text{B.2})$$

Now it is enough if we prove that c in (B.2) is minimum when $\theta = 0$ or $\theta = \pi$.

1. when $S(\pi_1) > S(\pi_2)$:

$$\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta) - \left(S_2 + \frac{r_1^2}{2}(\theta - \sin \theta) \right) > 0$$

$$\pi r_1^2 - r_1^2(\theta - \sin \theta) - S_2 > 0$$

$$\theta - \sin \theta < \pi - \frac{S_2}{r_1^2}$$

• $S_2 > \pi r_1^2$: Suppose $S_2 = k\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta < \pi(1 - k)$$

This equation has no solutions.

• $S_2 < \pi r_1^2$: Suppose $S_2 = \frac{1}{k}\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta < \pi(1 - \frac{1}{k})$$

The expression for the isoperimetric ratio is

$$c = \frac{2r_1 \sin \frac{\theta}{2}}{S_2 + \frac{r_1^2}{2}(\theta - \sin \theta)}$$

where $S_2 < \pi r_1^2$, $\theta - \sin \theta < \pi(1 - \frac{1}{k})$ and $k > 1$.

Substituting the values, we get

$$c = \frac{2r_1 \sin \frac{\theta}{2}}{\pi r_1^2 + \frac{r_1^2}{2}\pi \left(1 - \frac{1}{k}\right)} \Rightarrow c = \frac{4k \sin \frac{\theta}{2}}{\pi r_1(3k - 1)} \Rightarrow c = \frac{4 \sin \frac{\theta}{2}}{r_1(\theta - \sin \theta + 2\pi)} \quad (\text{B.3})$$

Equation (B.3) is minimum when $\theta = 0$ in the range $(0, \pi)$, see plot of (B.3) in Figure B.4.

2. when $S(\pi_1) < S(\pi_2)$:

$$S_2 + \frac{r_1^2}{2}(\theta - \sin \theta) - \left(\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta) \right) > 0$$

Figure B.4: Plot of the function in (B.3)

$$S_2 - \pi r_1^2 + r_1^2(\theta - \sin \theta) > 0$$

$$\theta - \sin \theta > \pi - \frac{S_2}{r_1^2}$$

- $S_2 > \pi r_1^2$: Suppose $S_2 = k\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta > \pi(1 - k)$$

This equation has no solutions.

- $S_2 < \pi r_1^2$: Suppose $S_2 = \frac{1}{k}\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta > \pi(1 - \frac{1}{k})$$

The expression for the isoperimetric ratio is

$$c = \frac{2r_1 \sin \frac{\theta}{2}}{\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta)}$$

where $S_2 < \pi r_1^2$, $\theta - \sin \theta < \pi(1 - \frac{1}{k})$ and $k > 1$.

Substituting the values, we get

$$c = \frac{2r_1 \sin \frac{\theta}{2}}{\pi r_1^2 - \frac{r_1^2}{2} \left(\pi \left(1 - \frac{1}{k} \right) \right)} \Rightarrow c = \frac{4k \sin \frac{\theta}{2}}{\pi r_1 (k+1)}, \Rightarrow c = \frac{4 \sin \frac{\theta}{2}}{r_1 \left(\theta - \sin \theta + \frac{2\pi}{k} \right)} \quad (\text{B.4})$$

Equation (B.4) is minimum when $\theta = 0$ in the range $(0, \pi)$, see plot of (B.4) in Figure B.5.

Figure B.5: Plot of the function in (B.4)

Figure B.6: Minimum isoperimetric ratio cut (cut within smaller region) - illustration

When $S_2 < \pi r_1^2$, we obtained $\theta = 0$ and $\theta = \pi$. This indicates that the cut with the minimum isoperimetric ratio can occur either along the connection/bridge or along the diameter of R_1 . Cut with the smallest isoperimetric ratio occurring within R_1 on the other side of the diameter (the farther side of the bridge in Figure B.3(b)), is not possible because

that will produce a partition π'_1 that is smaller than π_2 , i.e., $S(\pi'_1) < S(\pi_2)$ for a similar connection width $\pi'_{12} = \pi_{12}$ resulting in higher isoperimetric ratio. Similar explanations can be used to prove that the cut cannot occur on the smaller region, see Figure B.6. ■

Lemma 5. *For a circular region of radius r as in Figure B.1, if the weights of the graph are assigned using distance transform raised to the power p , where $p > 0$, the isoperimetric ratio of a cut within this circular region is given by*

$$c = \frac{2^{p+2} r^{p-1} \sin^{p+1} \frac{\theta}{2}}{\theta - \sin \theta}$$

where $0 < \theta < \pi$ is the angle subtended by the cut (chord) at the center of the circle. The cut with the smallest isoperimetric ratio occurs at $\theta = \pi$ for $p \leq 2$, i.e., along the diameter of the circle.

Proof. From (B.1), the isoperimetric ratio c of a cut within the circular region when all edges have identical weights equal to 1, is given by

$$c = \frac{2r \sin \frac{\theta}{2}}{\frac{r^2}{2}(\theta - \sin \theta)} \quad (\text{B.5})$$

From (5.6), the weighted boundary length of the connection of width w can be obtained as w^{p+1} . Using this in (B.5), we get the expression for c as

$$c = \frac{\left(2r \sin \frac{\theta}{2}\right)^{p+1}}{\frac{r^2}{2}(\theta - \sin \theta)} \Rightarrow c = \frac{2^{p+2} r^{p-1} \sin^{p+1} \frac{\theta}{2}}{\theta - \sin \theta} \quad (\text{B.6})$$

The function has the minimum at $\theta = \pi$ in the range $(0, \pi)$ when $p \leq 2$, see Figure B.7. This proves that the cut occurring along the diameter of the circle has the smallest isoperimetric ratio. ■

Lemma 6. *Consider a circular region R_1 of radius r_1 connected to another circular region R_2 of radius r_2 by a bridge of width w . If the weights of the graph are assigned using distance transform raised to the power p , where $p > 0$, then (i) if $r_2 < r_1$ or $S_2 < \pi r_1^2$, cut with the minimum isoperimetric ratio occurs either along the connection or along the*

Figure B.7: Plot of the function in (B.6) for $p \leq 2$

diameter of R_1 (ii) if $r_2 > r_1$ or $S_2 > \pi r_1^2$, no cut with the minimum isoperimetric ratio can occur within R_1 .

Proof. Consider a cut as shown in Figure B.3(a). The cut creates two partitions π_1 and π_2 and subtends an angle θ ($0 < \theta < \pi$) with the center of the circle. The width of the cut π_{12} is equal to

$$\pi_{12} = \left(2r_1 \sin \frac{\theta}{2}\right)^{p+1}$$

The areas of the partitions are given by

$$S(\pi_1) = \pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta)$$

$$S(\pi_2) = \pi r_2^2 + \frac{r_1^2}{2}(\theta - \sin \theta)$$

The isoperimetric ratio of this cut can be given by

$$c = \frac{\left(2r_1 \sin \frac{\theta}{2}\right)^{p+1}}{\min\left(\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta), S_2 + \frac{r_1^2}{2}(\theta - \sin \theta)\right)} \quad (\text{B.7})$$

Now it is enough if we prove that c in (B.7) is minimum when $\theta = 0$ or $\theta = \pi$.

1. when $S(\pi_1) > S(\pi_2)$:

$$\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta) - \left(S_2 + \frac{r_1^2}{2}(\theta - \sin \theta) \right) > 0$$

$$\pi r_1^2 - r_1^2(\theta - \sin \theta) - S_2 > 0$$

$$\theta - \sin \theta < \pi - \frac{S_2}{r_1^2}$$

- $S_2 > \pi r_1^2$: Suppose $S_2 = k\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta < \pi(1 - k)$$

This equation has no solutions.

- $S_2 < \pi r_1^2$: Suppose $S_2 = \frac{1}{k}\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta < \pi\left(1 - \frac{1}{k}\right)$$

The expression for the isoperimetric ratio is

$$c = \frac{\left(2r_1 \sin \frac{\theta}{2}\right)^{p+1}}{S_2 + \frac{r_1^2}{2}(\theta - \sin \theta)}$$

where $S_2 < \pi r_1^2$, $\theta - \sin \theta < \pi\left(1 - \frac{1}{k}\right)$ and $k > 1$.

Substituting the values, we get

$$c = \frac{\left(2r_1 \sin \frac{\theta}{2}\right)^{p+1}}{\pi r_1^2 + \frac{r_1^2}{2}\pi\left(1 - \frac{1}{k}\right)} \Rightarrow c = \frac{k2^{p+2}r_1^{p-1} \sin^{p+1} \frac{\theta}{2}}{\pi(3k - 1)} \Rightarrow c = \frac{2^{p+2}r_1^{p-1} \sin^{p+1} \frac{\theta}{2}}{\theta - \sin \theta + 2\pi + \varepsilon} \quad (\text{B.8})$$

Equation (B.3) is minimum when $\theta = 0$ in the range $(0, \pi)$, see plot of (B.8) in Figure B.8. For ease of comparison and simplicity, we use $p = 1$ for the equations.

Therefore, using $p = 1$ in (B.8), we get

$$c = \frac{8 \sin^2 \frac{\theta}{2}}{\theta - \sin \theta + 2\pi + \varepsilon} \quad (\text{B.9})$$

Figure B.8: Plot of the function in (B.8)

2. when $S(\pi_1) < S(\pi_2)$:

$$S_2 + \frac{r_1^2}{2}(\theta - \sin \theta) - \left(\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta) \right) > 0$$

$$S_2 - \pi r_1^2 + r_1^2(\theta - \sin \theta) > 0$$

$$\theta - \sin \theta > \pi - \frac{S_2}{r_1^2}$$

- $S_2 > \pi r_1^2$: Suppose $S_2 = k\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta > \pi(1 - k)$$

This equation has no solutions.

- $S_2 < \pi r_1^2$: Suppose $S_2 = \frac{1}{k}\pi r_1^2$ where $k > 1$, then

$$\theta - \sin \theta > \pi(1 - \frac{1}{k})$$

The expression for the isoperimetric ratio is

$$c = \frac{\left(2r_1 \sin \frac{\theta}{2}\right)^{p+1}}{\pi r_1^2 - \frac{r_1^2}{2}(\theta - \sin \theta)}$$

where $S_2 < \pi r_1^2$, $\theta - \sin \theta > \pi(1 - \frac{1}{k})$ and $k > 1$.

Substituting the values, we get

$$c = \frac{\left(2r_1 \sin \frac{\theta}{2}\right)^{p+1}}{\pi r_1^2 - \frac{r_1^2}{2}\pi(1 - \frac{1}{k})} \Rightarrow c = \frac{k2^{p+2}r_1^{p-1} \sin^{p+1} \frac{\theta}{2}}{\pi(k+1)} \Rightarrow c = \frac{2^{p+2}r_1^{p-1} \sin^{p+1} \frac{\theta}{2}}{\theta - \sin \theta + \frac{2\pi}{k} - \varepsilon} \quad (\text{B.10})$$

Equation (B.10) is minimum when $\theta = 0$ in the range $(0, \pi)$, see plot of (B.10) in Figure B.9. Using $p = 1$ in (B.10), we get

$$c = \frac{8 \sin^2 \frac{\theta}{2}}{\theta - \sin \theta + \frac{2\pi}{k} - \varepsilon} \quad (\text{B.11})$$

Figure B.9: Plot of the function in (B.10)

When $S < \pi r_1^2$, we obtained $\theta = 0$ and $\theta = \pi$. This indicates that the cut with the minimum isoperimetric ratio can occur either along the connection/bridge or along the

diameter of R_1 . Cut with the smallest isoperimetric ratio occurring within R_1 on the other side of the diameter (the farther side of the bridge in Figure B.3(b)), is not possible because that will produce a partition π'_1 that is smaller than π_2 , i.e., $S(\pi'_1) < S(\pi_2)$ for a similar connection width $\pi'_{12} = \pi_{12}$ resulting in higher isoperimetric ratio. Similar explanations can be used to prove that the cut cannot occur on the smaller region, see Figure B.6. ■

Bibliography

- [1] R.B. Buxton. *Introduction to functional magnetic resonance imaging: Principles and techniques*. Cambridge University Press, 2002.
- [2] K.S. Meacham. *The MRI study guide for technologists*. Springer, 1995.
- [3] E. Berry and A. Bulpitt. *Fundamentals of MRI: An interactive learning approach*. CRC Press, 2009.
- [4] R.H. Hashemi, W.G. Bradley, and C.J. Lisanti. *MRI: The basics*. Lippincott Williams and Wilkins, 2004.
- [5] Z.P. Liang and P.C. Lauterbur. *Principles of Magnetic Resonance Imaging: A signal processing perspective*. New York: IEEE Press, 2000.
- [6] D.G. Mitchell and M.S. Cohen. *MRI principles*. Saunders, 2004.
- [7] S. Webb. *The physics of medical imaging*. Taylor and Francis, 1988.
- [8] D.W. McRobbie, E.A. Moore, M.J. Graves, and M.R. Prince. *MRI: From picture to proton*. Cambridge University Press, 2007.
- [9] P. Jezzard, P.M. Matthews, and S.M. Smith. *Functional MRI: An introduction to methods*. Oxford University Press, 2002.
- [10] D. Weishaupt, V.D. Köchli, and B. Marincek. *How does MRI work?* springer, 2006.
- [11] S.A. Huettel, A.W. Song, and G. McCarthy. *Functional magnetic resonance imaging*. Sinauer Associates, 2009.

- [12] J. Jovicich, S. Czanner, X. Han, D. Salat, A. van der Kouwe, B. Quinn, J. Pacheco, M. Albert, R. Killiany, P. Maguire, D. Blacker, D. Rosas, N. Makris, R. Gollub, A. Dale, B.C. Dickerson, and B. Fischl. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, 46(1):177–192, 2009.
- [13] A.J.W. van der Kouwe, T. Benner, D.H. Salat, and B. Fischl. Brain morphometry with multiecho MPRAGE. *NeuroImage*, 40:559–569, 2008.
- [14] B. Fischl, D.H. Salat, A.J.W. van der Kouwe, N. Makris, F. Ségonne, B.T. Quinn, and A.M. Dale. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 22:1060–1075, 2004.
- [15] J.G. Sled, A.P. Zijdenbos, and A.C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998.
- [16] E.B. Lewis and N.C. Fox. Correction of differential intensity inhomogeneity in longitudinal MR images. *NeuroImage*, 23(1):75–83, 2004.
- [17] U. Vovk, F. Pernuž, and B. Likar. Intensity inhomogeneity correction of multispectral MR images. *NeuroImage*, 32(1):54–61, 2006.
- [18] E.L. Geda, D.L. Collins, and D.L. Arnold. Automated quality control of brain MR images. *Journal of Magnetic Resonance Imaging*, 28(2):308–319, 2008.
- [19] E.M. Haacke, R.W. Brown, M.R. Thompson, and R. Venkatesan. *Magnetic resonance imaging: Physical principles and sequence design*. John Wiley and Sons, Inc., 1999.
- [20] X. Han, J. Jovicich, D. Salat, A. van der Kouwe, B. Quinn, S. Czanner, E. Busa, J. Pacheco, M. Albert, R. Killiany, P. Maguire, D. Rosas, N. Makris, A. Dale, B. Dickerson, and B. Fischl. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194, 2006.

- [21] M. Tsurusaki, R.C. Semelka, M. Zapparo, J. Elias Jr., E. Altun, E. Pamuklar, and K. Sugimura. Quantitative and qualitative comparison of 3.0T and 1.5T MR imaging of the liver in patients with diffuse parenchymal liver disease. *European Journal of Radiology*, 72(2):314–320, 2009.
- [22] A. Geissler, A. Gartus, T. Foki, A.R. Tahamtan, R. Beisteiner, and M. Barth. Contrast-to-noise ratio CNR as a quality parameter in fMRI. *Journal of Magnetic Resonance Imaging*, 25(6):1263–1270, 2007.
- [23] A.E. Burgess. The Rose model revisited. *Journal of the Optical Society of America A: Optics, Image Science and Vision*, 16:633–646, 1999.
- [24] J.G. Sled, A.P. Zijdenbos, and A.C. Evans. A comparison of retrospective intensity non-uniformity correction methods for MRI. *Information Processing in Medical Imaging 1997, Lecture Notes in Computer Science*, 1230:459–464, 1997.
- [25] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. *ECCV 2002, Lecture Notes in Computer Science*, 2352:408–422, 2002.
- [26] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.
- [27] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–165, 2004.
- [28] P.J. Besl and R.C. Jain. Segmentation through variable-order surface fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(2):167–192, 1988.
- [29] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.
- [30] S.A. Hojjatoleslami and J. Kittler. Region growing: A new approach. *IEEE Transactions on Image Processing*, 7(7):1079–1084, 1998.

- [31] A. Chakraborty, L.H. Staib, and J.S. Duncan. Deformable boundary finding influenced by region homogeneity. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–627, 1994.
- [32] S. Beucher and C. Lantuejoul. Use of watersheds in contour detection. *International Workshop on Image Processing*, 1979.
- [33] J.B.T.M. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta Informaticae*, 41:187228, 2001.
- [34] F. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image representation*, 1(1):21–46, 1990.
- [35] H.K. Hahn and H.O. Peitgen. The skull stripping problem in MRI solved by a single 3D watershed transform. *MICCAI 2000, Lecture Notes in Computer Science*, 1935:134–143, 2000.
- [36] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.
- [37] C. Xu and J.L. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, 7(3):359–369, 1998.
- [38] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.
- [39] T. McInerney and D. Terzopoulos. Topologically adaptive snakes. *IEEE International Conference on Computer Vision*, pages 840–845, 1995.
- [40] S. Osher and J.A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.
- [41] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):266–280, 2000.

- [42] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: A survey. *Medical Image Analysis*, 1(2):91–108, 1996.
- [43] C. Xu and J. Prince. Gradient vector flow: A new external force for snakes. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 66–71, 1997.
- [44] B. Leroy, I. Herlin, and L.D. Cohen. Multi-resolution algorithms for active contour models. *Lecture Notes in Computer Science*, 219, 1996.
- [45] L.D. Cohen. On active contour models and balloons. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 53(2):211–218, 1991.
- [46] D. Cremers, F. Tischhäuser, J. Weickert, and C. Schnörr. Diffusion snakes: Introducing statistical shape knowledge into the Mumford-Shah functional. *International Journal of Computer Vision*, 50(3), 2002.
- [47] J. Tang, S. Millington, S.T. Action, J. Crandall, and S. Hurwitz. Ankle cartilage surface segmentation using directional gradient vector flow snake. *IEEE International Conference on Image Processing*, pages 2745–2748, 2004.
- [48] J. Cheng and S.W. Foo. Dynamic directional gradient vector flow for snakes. *IEEE Transactions on Image Processing*, 15:1563–1571, 2006.
- [49] R. Malladi, J.A. Sethian, and B.C. Vemuri. Shape modeling with front propagation: A level set approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):158–175, 1995.
- [50] C. Li, C.Y. Kao, J.C. Gore, and Z. Ding. Implicit active contours driven by local binary fitting energy. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.
- [51] S. Chen and R.J. Radke. Level set segmentation with both shape and intensity priors. *IEEE International Conference on Computer Vision*, pages 763–770, 2009.
- [52] D. Cremers. Nonlinear dynamical shape priors for level set segmentation. *Journal of Scientific Computing*, 35(2-3):132–143, 2008.

- [53] Z. Ximei, P. Zhenkuan, and Z. Shixiu. Segmentation of nerve fibers by introducing intensity priority in level set. *International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, pages 1–3, 2010.
- [54] L. Min, X. Xiangmin, Q. Min, and W. Zhuocai. A fast level set segmentation method based on the overall information of image. *International Symposium on Information Engineering and Electronic Commerce (IEEC)*, pages 1–4, 2010.
- [55] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of Royal Statistical Society Series B*, 51:271–279, 1989.
- [56] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. *IEEE International Conference on Computer Vision*, 1:105–112, 2001.
- [57] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.
- [58] L. Ford and D. Fulkerson. Flows in networks. *Princeton University Press*, 1962.
- [59] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.
- [60] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.
- [61] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [62] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

- [63] S. Wang and J.M. Siskind. Image segmentation with minimum mean cut. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 517–524, 2001.
- [64] D.S. Hochbaum. Polynomial time algorithms for ratio regions and a variant of normalized cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):889–898, 2010.
- [65] N. Xu, R. Bansal, and N. Ahuja. Object segmentation using graph cuts based active contours. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:46–53, 2003.
- [66] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, 23(3):309–314, 2004.
- [67] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [68] C. Couprie, L. Grady, L. Najman, and H. Talbot. Power watersheds: A new image segmentation framework extending graph cuts, random walker and optimal spanning forest. *IEEE Conference on Computer Vision*, pages 731–738, 2009.
- [69] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.
- [70] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. *IEEE Conference on Computer Vision*, pages 1–8, 2007.
- [71] C. Allene, J.Y. Audibert, and M. Couprie. Some links between extremum spanning forests, watersheds and min-cuts. *Image and Vision Computing*, 28(10):1460–1471, 2010.
- [72] J. Cousty, G. Bertrand, L. Najman, and M. Couprie. Watershed cuts: Minimum spanning forests and the drop of water principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1362–1374, 2009.

- [73] J. Cousty, G. Bertrand, L. Najman, and M. Couprie. Watershed cuts: Thinnings, shortest path forests, and topological watersheds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):925–939, 2010.
- [74] R.C. Gonzalez and R.E. Woods. *Digital image processing*. Pearson Education, Inc., 2002.
- [75] L. Grady and E.L. Schwartz. Isoperimetric partitioning: A new algorithm for graph partitioning. *SIAM Journal on Scientific Computing*, 27(6):1844–1866, 2006.
- [76] L. Grady. Fast, quality, segmentation of large volumes - Isoperimetric distance trees. *ECCV 2006, Lecture Notes in Computer Science*, 3(3953):449–462, 2006.
- [77] L. Grady and E.L. Schwartz. Isoperimetric graph partitioning for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):469–475, 2006.
- [78] J. Xiong, J.H. Gao, J.L. Lancaster, and P.T. Fox. Assessment and optimization of functional MRI analyses. *Human Brain Mapping*, 4:153–167, 1996.
- [79] J. Marchini and A. Presanis. Comparing methods of analyzing fMRI statistical parametric maps. *NeuroImage*, 22(3):1203–1213, 2004.
- [80] K.M. Petersson, T.E. Nichols, J.B. Poline, and A.P. Holmes. Statistical limitations in functional neuroimaging I. Non-inferential methods and statistical models. *Phil. Trans. R. Soc. London B*, 354:1239–1260, 1999.
- [81] T.E. Nichols and A.P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15:1–25, 2002.
- [82] T.E. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, 12:419–446, 2003.
- [83] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B*, 57:289–300, 1995.

- [84] C.R. Genovese, N.A. Lazar, and T.E. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4):870–878, 2002.
- [85] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- [86] N.V. Hartvig and J.L. Jensen. Spatial mixture modeling of fMRI data. *Human Brain Mapping*, 11(4):233–248, 2000.
- [87] B.R. Logan and D.B. Rowe. An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, 22(1):95–108, 2004.
- [88] K.J. Friston, K.J. Worsley, R.S.J. Frackowiak, J.C. Mazziotta, and A.C. Evans. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214–220, 1994.
- [89] J.B. Poline, K.J. Worsley, A.C. Evans, and K.J. Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, 5(2):83–96, 1997.
- [90] E.T. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M.J. Brammer. Global, voxel, cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(1):32–42, 1999.
- [91] S. Hayasaka and T.E. Nichols. Combining voxel intensity and cluster extent with permutation test framework. *NeuroImage*, 23(1):54–63, 2004.
- [92] N.A. Lazar, B. Luna, J.A. Sweeney, and W.F. Eddy. Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, 16(2):538–550, 2002.
- [93] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

- [94] K. Held, E.R. Kops, B.J. Krause, W.M.III. Wells, R. Kikinis, and H.-W. Muller-Gartner. Markov random field segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 16(6):878–886, 1997.
- [95] D.L. Pham, C. Xu, and J.L. Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2, 2000.
- [96] X. Descombes, F. Kruggel, and D.Y. Von Cramon. fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. *Neuroimage*, 8(4):340–349, 1998.
- [97] X. Descombes, F. Kruggel, and D.Y. Von Cramon. Spatio-temporal fMRI analysis using Markov random fields. *IEEE Transactions on Medical Imaging*, 17(6):1028–1039, 1998.
- [98] J.C. Rajapakse and J. Piyaratna. Bayesian approach to segmentation of statistical parametric maps. *IEEE Transactions on Biomedical Engineering*, 48(10):1186–1194, 2001.
- [99] E. Salli, H.J. Aronen, S. Savolainen, A. Korvenoja, and A. Visa. Contextual clustering for analysis of functional MRI data. *IEEE Transactions on Medical Imaging*, 20(5):403–414, 2001.
- [100] E.R. Cosman, J.W. Fisher, and W.M. Wells. Exact map activity detection in fMRI using a GLM with an Ising spatial prior. *MICCAI 2004, Lecture Notes in Computer Science*, 2004.
- [101] M.W. Woolrich, T.E.J. Behrens, C.F. Beckmann, and S.M. Smith. Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. *IEEE Transactions on Medical Imaging*, 24(1):1–11, 2005.
- [102] W. Ou and P. Golland. From spatial regularization to anatomical priors in fMRI analysis. *IPMI 2005, Lecture Notes in Computer Science*, 3565:88–100, 2005.
- [103] P. Kohli and P.H.S. Torr. Efficiently solving dynamic Markov random fields using graph cuts. *IEEE International Conference on Computer Vision*, 2:922–929, 2005.

- [104] N. Paragios. A variational approach for the segmentation of the left ventricle in cardiac image analysis. *International Journal of Computer Vision*, 50(3):345–362, 2002.
- [105] P.C. Tay, B. Li, C.D. Garson, S.T. Acton, and J.A. Hossack. Left ventricle segmentation using model fitting and active surfaces. *Journal of Signal Processing Systems*, 2008.
- [106] Marie-Pierre Jolly. Automatic segmentation of the left ventricle in cardiac MR and CT images. *International Journal of Computer Vision*, 70(2):151–163, 2006.
- [107] Q. Peng, R.W. McColl, Y. Ding, J.Wang, J.M. Chia, and P.T. Weatherall. Automated method for accurate abdominal fat quantification on water-saturated magnetic resonance images. *Journal of Magnetic Resonance Imaging*, 26(3):738–746, 2007.
- [108] L. Lemieux, G. Hagemann, K. Krakow, and F.G. Woermann. Fast, accurate and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. *Magnetic Resonance in Medicine*, 42(1):127–135, 1999.
- [109] M.S. Atkins and B.T. Mackiewicz. Fully automatic segmentation of the brain in MRI. *IEEE Transactions on Medical Imaging*, 17(1):98–107, 1998.
- [110] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg, and R.M. Leahy. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856–876, 2001.
- [111] J. Lee, N. Kim, H. Lee, J.B. Seo, H.J. Won, Y.M. Shin, Y.G. Shin, and S.H. Kim. Efficient liver segmentation using a level-set method with optimal detection of the initial liver boundary from level-set speed images. *Journal of Computer Methods and Programs in Biomedicine*, 88:26–38, 2007.
- [112] X. Yang, H. Li, and X. Zhou. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy. *IEEE Transactions on Circuits and Systems*, 53(11):2405–2414, 2006.

- [113] J. Park and J.M. Keller. Snakes on the watershed. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1201–1205, 2001.
- [114] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344–371, 1986.
- [115] V. Positano, T. Christiansen, M.F. Santarelli, S. Ringgaard, L. Landini, and A. Gastaldelli. Accurate segmentation of subcutaneous and intermuscular adipose tissue from MR images of the thigh. *Journal of Magnetic Resonance Imaging*, 29(3):677–684, 2009.
- [116] V. Positano, A. Gastaldelli, A.M. Sironi, M.F. Santarelli, M. Lombardi, and L. Landini. An accurate and robust method for unsupervised assessment of abdominal fat by MRI. *Journal of Magnetic Resonance Imaging*, 20(4):684–689, 2004.
- [117] P. Ranefall, A.W. Bidar, and P.D. Hockings. Automatic segmentation of intra-abdominal and subcutaneous adipose tissue in 3D whole mouse MRI. *Journal of Magnetic Resonance Imaging*, 30(3):554–560, 2009.
- [118] Y. Jin, C.Z. Imielinska, A.F. Laine, J. Udupa, W. Shen, and S.B. Heymsfield. Segmentation and evaluation of adipose tissue from whole body MRI scans. *MICCAI 2003, Lecture Notes in Computer Science*, 2878:635–642, 2003.
- [119] J. Kullberg, H. Ahlstrom, L. Johansson, and H. Frimmel. Automated and reproducible segmentation of visceral and subcutaneous adipose tissue from abdominal MRI. *International Journal of Obesity*, 31(12):1806–1817, 2007.
- [120] T-H. Liou, W.P. Chan, L-C Pan, P-W. Lin, P. Chou, and C-H. Chen. Fully automated large-scale assessment of visceral and subcutaneous abdominal adipose tissue by magnetic resonance imaging. *International Journal of Obesity*, 30(5):844–852, 2006.
- [121] V. Positano, K. Cusi, M. F. Santarelli, A. Sironi, R. Petz, R. Defronzo, L. Landini, and A. Gastaldelli. Automatic correction of intensity inhomogeneities improves unsupervised assessment of abdominal fat by MRI. *Journal of Magnetic Resonance Imaging*, 28(2):403–410, 2008.

- [122] D.L. Sussman, J. Yao, and R.M. Summers. Automated measurement and segmentation of abdominal adipose tissue in mri. *IEEE international conference on Biomedical imaging: from nano to Macro*, pages 936–939, 2010.
- [123] D. D. Brennan, P. F. Whelan, K. Robinson, O. Ghita, J. M. O’Brien, R. Sadleir, and S. J. Eustace. Rapid automated measurement of body fat distribution from whole-body MRI. *AJR American Journal of Roentgenology*, 185(2):418–423, 2005.
- [124] O.D. Leinhard, A. Johansson, J. Rydell, Ö. Smedby, F. Nyström, P. Lundberg, and M. Borga. Quantitative abdominal fat estimation using MRI. *International Conference on Pattern Recognition*, pages 1–4, 2008.
- [125] J. Ashburner and K.J. Friston. Voxel-based morphometry: The methods. *NeuroImage*, 11(6):805–821, 2000.
- [126] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H.J. Berg, P.R. Bannister, M.D. Luca, I. Drobnjak, D.E. Flitney, R.K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N.D. Stefano, J.M. Brady, and P.M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):S208–S219, 2004.
- [127] A.M. Dale, B. Fischl, and M.I. Sereno. Cortical surface-based analysis i. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.
- [128] J.S. Kim, V. Singh, J.K. Lee, J. Lerch, Y.A. Dab’bagh, D. MacDonald, J.M. Lee, S.I. Kim, and A.C. Evans. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage*, 27(1):210–221, 2005.
- [129] S.M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P.M. Matthews, A. Federico, and N.D. Stefano. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479–489, 2002.
- [130] G. Fein, B. Landman, H. Tran, J. Barakos, K. Moon, V.D. Sclafani, and R. Shumway. Statistical parametric mapping of brain morphology: Sensitivity is dramatically

- increased by using brain-extracted images as inputs. *NeuroImage*, 30(4):1187–1195, 2006.
- [131] S.M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.
- [132] A.H. Zhuang, D.J. Valentino, and A.W. Toga. Skull-stripping magnetic resonance brain images using a model-based level set. *NeuroImage*, 32(1):79–92, 2006.
- [133] A. Huang, R. Abugharbieh, R. Ram, and A. Traboulsee. MRI brain extraction with combined expectation maximization and geodesic active contours. *IEEE International Symposium on Signal Processing and Information Technology*, pages 107–111, 2006.
- [134] F. Segonne, A.M. Dale, E. Busa, M. Glessner, D. Salat, H.K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22:1060–1075, 2004.
- [135] G. Gerig, O. Kubler, R. Kikinis, and F.A. Jolesz. Nonlinear anisotropic filtering of MRI data. *IEEE Transactions on Medical Imaging*, 11(2):221 – 232, 1992.
- [136] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of Royal Society of London*, 207(B):187–217, 1980.
- [137] J.G Park and C. Lee. Skull stripping based on region growing for magnetic resonance brain images. *NeuroImage*, 47(4):1394–1407, 2009.
- [138] Z.Y. Shan, G.H. Yue, and J.Z. Liu. Automated histogram-based brain segmentation in T1-weighted three-dimensional magnetic resonance head images. *NeuroImage*, 17(3):1587–1598, 2002.
- [139] J. Chiverton, K. Wells, E. Lewis, C. Chen, B. Podda, and D. Johnson. Statistical morphological skull stripping of adult and infant MRI data. *Computers in Biology and Medicine*, 37(3):342–357, 2007.
- [140] J.S. Suri. Two-dimensional fast magnetic resonance brain segmentation. *IEEE Engineering in Medicine and Biology Magazine*, 20(4):84–95, 2001.

- [141] C. Baillard, P. Hellier, and C. Barillot. Segmentation of brain 3D MR images using level sets and dense registration. *Medical Image Analysis*, 5(3):185–194, 2001.
- [142] A. Mikheev, G. Nevsky, S. Govindan, R. Grossman, and H. Rusinek. Fully automatic segmentation of the brain from T1-weighted MRI using Bridge Burner algorithm. *Journal of Magnetic Resonance Imaging*, 27(6):1235–1241, 2008.
- [143] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20:45–57, 2001.
- [144] C.F. Notestine, I.B. Ozyurt, C.P. Clark, S. Morris, A.B. Grethe, M.W. Bondi, T.L. Jernigan, B. Fischl, F. Segonne, D.W. Shattuck, R.M. Leahy, D.E. Rex, A.W. Toga, K.H. Zou, Morphometry BIRN, and G.G. Brown. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. *Human Brain Mapping*, 27(2):99–113, 2006.
- [145] T. Kapur, W.E.L. Grimson, W.M. Wells III, and R. Kikinis. Segmentation of brain tissue from magnetic resonance images. *Medical Image Analysis*, 1(2):109–127, 1996.
- [146] Y. Hata, S. Kobashi, S. Hirano, H. Kitagaki, and E. Mori. Automated segmentation of human brain MR images aided by fuzzy information granulation and fuzzy inference. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 30(3):381–395, 2000.
- [147] R. Stokking, K.L. Vincken, and M.A. Viergever. Automatic morphology-based brain segmentation (MBRASE) from MRI-t1 data. *NeuroImage*, 12(6):726–738, 2000.
- [148] B. Fischl, M.I. Sereno, and A.M. Dale. Cortical surface-based analysis ii. inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999.
- [149] K. Rehm, K. Schaper, J. Anderson, R. Woods, S. Stoltzner, and D. Rottenberg. Putting our heads together: A consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. *NeuroImage*, 22(3):1262–1270, 2004.

- [150] S.W. Hartley, A.I. Scher, E.S.C. Korf, L.R. White, and L.J. Launer. Analysis and validation of automated skull stripping tools: A validation study based on 296 MR images from the honolulu asia aging study. *NeuroImage*, 30(4):1179–1186, 2006.
- [151] J.M. Lee, U. Yoon, S.H. Namb, J.H. Kima, I.Y. Kima, and S.I. Kim. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. *Computers in Biology and Medicine*, 33(4):495–507, 2003.
- [152] W.J. Niessen, K.L. Vincken, J. Weickert, B.M.T.H. Romeny, and M.A. Viergever. Multiscale segmentation of three-dimensional MR brain images. *International Journal of Computer Vision*, 31(2):185–202, 1999.
- [153] K. Boesen, K. Rehm, K. Schaper, S. Stoltzner, R. Woods, E. Lüders, and D. Rotenberg. Quantitative comparison of four brain extraction algorithms. *NeuroImage*, 22(3):1255–1261, 2004.
- [154] D.E. Rex, D.W. Shattuck, R.P. Woods, K.L. Narr, E. Luders, K. Rehm, S.E. Stoltzner, D.A. Rottenberg, and A.W. Toga. A meta-algorithm for brain extraction in MRI. *NeuroImage*, 23(2):625–637, 2004.
- [155] J. Hennig, M. Weigel, and K. Scheffler. Multiecho sequences with variable refocusing flip angles: Optimization of signal behavior using smooth transitions between pseudo steady states (TRAPS). *Magnetic Resonance in Medicine*, 49(3):527–535, 2003.
- [156] S. Rajagopalan, R.A. Karwoski, and R. Robb. Robust fast automatic skull stripping of MRI-T2 data. *SPIE Medical Imaging: Image Processing*, 5747:485–495, 2005.
- [157] A.M. Alyassin and H. Cline. Unsupervised brain segmentation using T2 window. *SPIE Medical Imaging*, 4684:1759–1769, 2002.
- [158] G. Helms, K. Kallenberg, and P. Dechent. Contrast-driven approach to intracranial segmentation using a combination of T2- and T1-weighted 3D MRI data sets. *Journal of Magnetic Resonance Imaging*, 24:790–795, 2006.
- [159] B. Fischer and J. Modersitzki. FLIRT: A flexible image registration toolbox. *Biomedical Image Registration, Lecture Notes in Computer Science*, 2717:261–270, 2003.

- [160] J.L.R. Andersson, M. Jenkinson, and S. Smith. Non-linear registration, aka Spatial normalisation. FMRIB Technical Report TR07JA2, FMRIB Centre, University of Oxford, 2007.
- [161] M. Jenkinson and S.M. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.
- [162] M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.
- [163] J. Modersitzki. FLIRT with rigidity: Image registration with a local non-rigidity penalty. *International Journal of Computer Vision*, 76(2):153–163, 2008.