Digital Media Technology Lab

Birmingham City University

# A Novel Statistical-based Approach for 3-D Surface Detection

Samuel Smith

A thesis submitted in partial fulfillment of the requirements of Birmingham City University for the degree of Doctor of Philosophy.
June 2021

## Abstract

The field of medical image analysis is concerned with the extraction of salient information from complex digital imagery. Developments in image acquisition tools has given rise to a number of 2-D and 3-D digital image domains that are capable of mapping the anatomy and internal structures of a patient in an non-invasive fashion. The data produced by these tools is inherently complex, and a number of image processing techniques are commonly applied to simplify the images in order to extract the information deemed most relevant. Image segmentation is one commonly applied process which partitions a digital image into multiple segments that correspond to various structures within the data. The goal of segmentation is to change the representation of an image into something that is easier to visualise and analyse in complex tasks (i.e. targeted clinical treatment planning). Often these segmentation tasks are performed manually by an expert clinician, however the task of drawing object contours is a time consuming process subject to human biases and interpretation. Automated and semi-automated segmentation is a complex, non-trivial process reliant on a number of pre-processing stages which first extract the spatial structural information contained within the image. For 2-D images, the structural information is contained within edge features and there are a number of edge detection algorithms in the literature which have been extensively appraised. For 3-D images this structural information is contained within the surface features, and while surface detection algorithms exist, their development is immature compared to edge detection and formal evaluation in the literature is largely absent. Furthermore, recent developments in statistical methods for 2-D edge feature extraction have showed promise in resolving 2-D structural information in medical data, however no work has yet explored these approaches in 3-D.

In this thesis two novel methods of statistical surface detection are presented, which contribute to the field by transferring approaches of 2-D statistical edge detection into 3-D. The proposed methods optimise the resolving power of the 2-D statistical methods while providing accurate surface detection in the $x$, $y$ and $z$ dimensions of images. The methods are presented with a range of parametric and non-parametric statistical tests which were extensively analysed using both qualitative and objective methods. In addition, the framework for evaluation is an additional novel contribution in this work, which considers individual aspects of surface detection performance, such as the effects of the statistical properties of the regions within the image, the impact of surface topology, and the response to multiple distinct regions present within the image. A comprehensive dataset of controlled interfaces is developed, and performance of the surface detection algorithms were judged using a novel fast implementation of F-measure analysis against

ground truth solutions which is new to this work. The surface detection methods are also analysed on real MRI data, and their performance was qualitatively assessed on the ability to detect brain tumour boundaries and structural pathologies in paediatric patient data. For a comparison against the state of art in surface detection, the methods evaluated in the thesis were compared against two existing baseline approaches, namely the 3-D Canny method, and 3-D Steerable filters.

The results of the evaluations reveal that the proposed 3-D statistical method for surface detection offers improved detection of surfaces on synthetic data with varying interface and topology considerations. Furthermore, the proposed methods improve detection of surfaces when the variability in image intensity high, such as within regions of texture, which is suitable for delineating regions of complex structure in MRI data. Additionally, the statistical methods were able to match the performance of the baseline methods under conditions considered optimal for the baseline approaches, such as the detection of surfaces with a strong intensity differential. Furthermore, the proposed methods of surface detection are shown to be suitable for real 3-D data which is anisotropic in resolution, namely on MRI imagery where the z-spacing within the dataset is often of poor resolution. Provided as a recommendation for further work, the best performing techniques are presented, notably these were the $\chi^2$ and Student $t$-test statistical methods. Characteristically these methods produced strong magnitude surfaces with good connectivity on real and synthetic data, with the $\chi^2$ test also achieving a good suppression of image noise. Therefore, illustrating the potential of this novel method of 3-D surface detection for medical image analysis applications.

## Acknowledgements

I would first like to thank my supervisor Ian, for his continued support, patience and for giving the me the confidence I lacked during various stages of the project, all of which have been instrumental in the successful outcome of this work. I would also like to express my gratitude to the Birmingham Children's Hospital and their Brain Tumour Research Group for the excellent work they do and their assistance with this project. In particular I would like to thank Jan for time consuming work of manually segmenting the MRI datasets, without which the project would not have been possible. I would also like to thank my friend Sean for providing a number of solutions to the many problems I faced along the way and inspiring me to approach problems in new ways. I would also like to thank my colleagues, friends and family, too many of which to name, that selflessly helped, guided and supported me throughout all of this, many thanks to you all.

# Contents

# List of Figures

# List of Tables

# Acronyms

| | |
|---|---|
| **CDM** | Closest Distance Match. |
| **CSA** | Cost Scaling Algorithm. |
| **CT** | Computed Tomography. |
| **EPS** | Efficient Pairing Strategy. |
| **FN** | False Negative. |
| **FP** | False Positive. |
| **HT** | Higher Threshold. |
| **LT** | Lower Threshold. |
| **MAD** | Mean Absolute Deviation. |
| **MR** | Maximum Response Method. |
| **MRI** | Magnetic Resonance Imaging. |
| **MSD** | Mean Square Deviation. |
| **NMS** | Non Maximal Suppression. |
| **PA** | Pilocytic Astrocytoma. |
| **PCM** | Pixel Correspondence Metric. |
| **PET** | Positron Emission Tomography. |
| **PFOM** | Pratt Figure Of Merit. |
| **SPECT** | Single Photon Emission Computed Tomography. |

**TN** True Negative.

**TP** True Positive.

**VM** Vector Magnitude Method.

# Chapter 1

# Introduction

Advances in computational power over recent decades have resulted in more applications for the acquisition, manipulation, and processing of digital image data. Commonly this data represents a two dimensional (2-D) array which consists of two finite dimensions (x,y) of pixel values. However, more recent developments in both consumer and commercial acquisition systems have resulted in an uprise in three-dimensional (3-D) digital image data which extends the number of finite dimensions of pixel information into three dimensions (x,y,z).

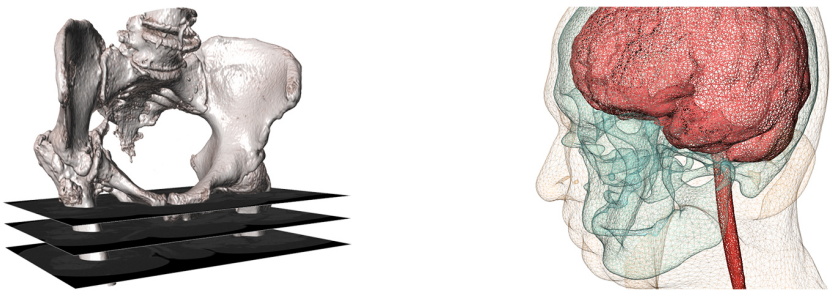


Figure 1.1: Example of a medical image software package, Synopsys's Simpleware™, with powerful 3-D visualisation and segmentation capabilities (Synopsys, 2021)

While many commercial software applications aim to exploit this 3-D data, commonly to improve visualisation and manipulation, for example Synopsys's Simpleware™ package and RadiAnt™ DICOM Viewer (Fig 1.1), the underlying digital processing of this information is commonly still applied in 2-D, thus not exploiting the information fully present in the third (or $z$) dimension. Thus, these applications while suitable for visualising an extrapolated form of 3-D visualisation, they do not truly represent the full structural information offered within the 3-D datasets. Furthermore, when visualising structures that exist in 3-D, processing on 2-D data alone may miss salient information

within the plane of processing and therefore structures of interest that exist within the $z$ plane of the data may be lost or misinterpreted with errors.

One feature of interest to many image analysis tasks such as segmentation, region profiling, and visualisation are surfaces. Surfaces being analogous to edges in 2-D data are highly sought after 3-D image features which separate two or more distinct regions of information. Intrinsic to the feature, surfaces provide information regarding the structure contained within 3-D image data. Surfaces are therefore vital in providing routes towards a clearer understanding of 3-D structures and offer potential methods for providing a truer and more accurate visualisation of often complex image data.

With current advances in image technology, surface information can support many applications in the medical visualisation, processing, and treatment planning domains. Often these advances are to automate or semi-automate manual tasks and improve efficiency, reliability, and accuracy of results. Commonly these methods rely on novel techniques and processes embedded into software packages that while partially automate processes, still require guidance by a human operator to achieve the desired result. Once such technique in medical analysis, visualisation and anatomical evaluations is contouring. Contouring is the highlighting and labelling of a boundary to a region of interest. This is a classical problem for 2-D image data and although some advances have been explored to semi automate this process, it is still largely a manual task. Thus, the de facto standard approach, whereby a human operator draws contours around required structures in 2-D, is time consuming and open to errors in interpretation. Furthermore, even though recent advances in 3-D acquisition and visualisation processes have resulted in improved clarity in complex 3-D image data, the contouring process for many medical applications is still applied in 2-D, therefore may not be fully representative of the true 3-D structures or reflecting the potential accuracy. One such example amongst many is in the contouring of Pilocytic Astrocytoma (PA) brain tumours in paediatric patients.

The 3-D structure of PA pathologies is complex and contouring is currently reliant on an interpretative, time consuming process performed by viewing the 2-D layers within the 3-D data in an axial view, and manually drawing a contour around the object of interest. Because of the rapid development of these pathologies in paediatric cases, this task often needs to be repeated throughout the diagnosis and treatment phases of a patients care multiple times to track the progression and to assess the impact of treatment. Further compounding these challenges are the limited availability of high resolution, isotropic, complete datasets. This is due to Magnetic resonance imaging (MRI) machines being a scarce resource, and image quality is often compromised by reducing the acquisition time to minimize patient discomfort. Furthermore, images are acquired using different MRI imaging modalities, which reveal different structures

within the data, while contrast agents can be added prior to acquisition of images to further increase the distinction between different image structures. The three primary modalities for magnetic resonance imaging of brain tumours are T1 weighted, T2 weighted, and T1-weighted with a contrast agent. Clinicians are then able to infer structures invisible in one domain, by their presence in a scan from an alternate modality. However, not all modalities are always available, and often images from the different modalities have different dimensional and resolution properties, thereby making image registration and contour assisted processing a non-trivial task for software applications.

Current machine learning methods offer some routes forward towards region segmentation, but the problems associated with limited datasets and lack of generalisability in the use case is apparent. Classical approaches, as opposed to more complex modern computer vision 'task-oriented' approaches such as those in the field of deep learning possess a number of useful advantages, chiefly amongst these are the fact Classical methods do not demand complex, lengthy training stages, which may also require extensive and sometimes unavailable datasets and powerful GPU processing. Classical approaches can avoid these difficult problems since they require no training, can be applied to all images without prior knowledge of the image content, and require comparatively less computational power.

While a number of classical approaches exist for 3-D surface segmentation in end applications like MRI contouring, their evaluation and formal exploration has not been conducted either objectively or against practical segmentation tasks. Furthermore, potential improvements 3-D surface segmentation can offer over 2-D counterpart techniques has not been appraised.

### 1.0.1 Motivation

The motivation behind this work is to explore methods for providing a reliable and fast method for 3-D surface detection which can be applied on common 3-D datasets. While 2-D edge detection methods for 2-D data have been extensively explored in the literature, 3-D surface detection methods lack the same level of maturity. Therefore, the potential advantages these methods can offer over their 2-D counterparts remains largely unexplored. Furthermore, the availability of 3-D imagery in the medical field has increased greatly in recent decades, and scanning techniques such as magnetic resonance imaging (MRI), computed tomography (CT), single photon emission computed tomography (SPECT), positron emission tomography (PET) and other 3D imaging modalities are now more prevalent (*MRI scanners in the UK 2000-2014 | Statistic* n.d.). However, the potential impact of 3-D surface detection on this data over 2-D edge detection has not been fully acknowledged. In recent times a large effort has been made

to utilise deep learning techniques to further assist in medical image analysis tasks (Ker et al., 2018) with 3-D data, however, a number of factors are currently limiting the potential of these approaches. Principally these include the sparsity and availability of sufficiently large, curated, and representative training data that includes expert labelling (Willemink et al., 2020). These limitations are compounded due to variations in image acquisition protocols, imperfections of scanning methods, patient cohorts and sample sizes, for example in paediatrics, where cases are of a lower frequency and patient data can be less generalisable.

As 3-D medical images contain complex structural information which can be difficult to visualise, typically classical image processing techniques are applied to simplify the visual content in order to assist practitioners in viewing the salient information. This can be as simple as viewing the image through a limited intensity (Brightness) range to enhance certain features, or more complex, such as algorithms which can segment or classify image structures. Powerful medical image analysis and visualisation software packages contain highly sought after image processing tools that can assist the medical practitioner with diagnosis and treatment planning tasks. These packages typically use a range of approaches, including both classical and machine learning based methods. Consequently, the performance of classical techniques is still relevant as they affect the diagnostic usefulness of these non-invasive imagining tools, this remains a long-standing computer vision problem having been expressed prior by (Zucker and Hummel, 1981). It has been firmly established that low level processes such as edge or surface detection can improve the performance of segmentation by removing the redundant information in the image (Pavlidis and Liow, 1990), therefore improvements to the underlying processes remain relevant.

Surface detection is a useful tool because it reduces the complexity of an image, while retaining information corresponding to the structure of the image content. In a medical context this structural information could the boundary of an organ, a skeletal object or a pathology such as a tumour. The extracted surface data is simple relative to the complexity of the unprocessed image data, affording other image processing techniques the ability to process an image which would otherwise be too complex. For example segmentation is a commonly applied computer vision process which can exploit surface features in order to divide an image into a collection of semantically meaningful, homogeneous, non overlapping regions. This is a crucial tool for medical image analysis tasks such as measuring and visualising anatomical structures, image guided interventions, surgical planning, analysing brain development and delineating lesions, tumours and other pathologies (Despotović et al., 2015).

The requirement of extracting structural features for 3-D segmentation has largely been left to 2-D edge detection methods (Lyra, 2012; Prakoonwit and Benjamin, 2012; Simmons et al., 2013), which do not produce optimal results in 3-D data. The work of Monga and Deriche (1989) presents the theoretical advantage in performance offered by 3-D methods over comparable 2-D algorithms in the detection of surfaces, inferring that fully 3-D methods should be employed over 2-D methods for the full performance potential of the detection process to be realised.

A notable problem for surface detection performance occurs when the differences between image regions are not clearly separable. For example, images obtained from 3-D scanning modalities often present with regions with non-homogeneous voxel intensity, noise, and anatomical structures which are sometimes more clearly defined by their texture profile than their intensity profile (Martin et al., 2004). Texture is therefore another important image feature (Petrou and García Sevilla, 2006a), one which can be used as a cue for medical image analysis tasks (Liyuan Li and Leung, 2002; Martin et al., 2004; Huan and Hou, 2008). However, it is more common for surface detection methods to detect the image structure by computing the intensity gradient function of an image. Where this is the case, texture based interfaces are not readily resolved by traditional gradient methods.

In 2-D data, statistical edge detection methods such as those evaluated in the work of Williams et al. (2014) were shown to be more conducive to resolving a range of interface types, including both intensity edges and texture based edges, yet evaluation of 3-D statistical surface detection methods remain lacking in the literature. Traditional methods of surface detection may therefore provide sub-optimal results when compared with statistical approaches, likely in areas of noise and where interfaces are defined by changes in a texture profile.

The motivation for this thesis was to develop and evaluate a statistical based surface detection method, which utilises both the theoretical advantages offered by 3-D data, described by Monga and Deriche (1989), in addition to the advantages offered by a statistical approach, expressed by Williams et al. (2014), in order to improve the current standard of surface detection.

## 1.1 Aims and Objectives of the Project

- To develop and evaluate the performance of a 3-D statistical filtering approach to detecting surfaces, to improve the current standard of surface detection for 3-D image data.

### 1.1.1 Objectives

1. Establish strengths and weaknesses of current state of the art methods of surface detection via literature, positioning the development of 3-D statistical methods.
2. Review, evaluate and develop robust quantitative performance measures for testing 3-D surface detection techniques.
3. Develop a protocol for creating synthetic data for objectively analysing surface detection methods.
4. Evaluate performance of novel statistical surface detection method against current state of the art systems using an objective analysis.
5. Identify trends between the results from synthetic data and that of tumour interfaces in real multi-modal MRI data.
6. Extrapolate findings to make recommendations for optimal application of techniques.

## 1.2 Structure of the Thesis

**Chapter 2** presents a review of the literature on existing surface detection methods. However, to fully understand the development of surface detection algorithms, 2-D edge detection methods first needed to be explored since the majority of 3-D surface detection algorithms are derived from 2-D edge detection techniques. This chapter explains functionality of the 2-D and 3-D techniques and demonstrates where these methods are effective and where they are not. The shortcomings of these techniques are highlighted to illustrate the historical need for improvements, and how these developments were driven for improved performance. Furthermore, Chapter 2 defines the properties of images both in 2-D and 3-D formats and defines what role those properties play in accurate detection. Chapter 2 also describes what constitutes an image region, defining both edge and surface features and their utility.

**Chapter 3** details the design and implementation of two novel approaches to statistical surface detection. Detailed are two unique configurations of the method, these variants are called the maximum response (MR) and vector magnitude (VM) methods. A computational cost analysis is undertaken for the newly introduced methods, which evaluate the completion times of each statistical test across a range of suitable neighbourhood scales. The analysis underpins the advantages offered by the Vector Magnitude method with respect to reducing the number of redundant calculations for a faster, scale invariant approach to statistical surface detection. The VM approach also newly introduces an option to reduce the surface map distortion which occurs

when applied to data which is not isotropic in its resolution. A further contribution presented in this chapter is a novel 3-D orientation filtering technique, which allows for 3-D non-maximal suppression to be applied when surface orientation is not known a-priori. The chapter discusses the effect of adjusting neighbourhood scale and choice of statistical test in a 3-D context which has yet to be explored outside of this work.

**Chapter 4** provides an overview and analysis of a number of different performance measure techniques which are used for the evaluation of edge and surface detection results. This work shows that there are a number of flaws with the accuracy of performance metrics that have been used prolifically in objective evaluations. Existing metrics which are accurate and robust against the issues discussed in this chapter are typically computationally expensive, rendering these techniques impractical when required for substantially large datasets. The problem of efficiency is further exacerbated where analysis of 3-D results are required, since the added dimension to the data effectively makes the analysis is more computationally complex while simultaneously increasing the size of the dataset images. This chapter introduces an improvement in accuracy of the closest distance match (CDM) of Bowyer et al. (2001), using a novel one to one correspondence matching procedure to produce an efficient paring strategy (EPS) (Smith. and Williams., 2020) which offers robust accurate performance analysis of edge and surface results compared to ground truth datasets.

**Chapter 5** provides a comprehensive analysis of the statistical surface detection models. The statistical methods are compared against the 3-D Canny operator which is a 3-D extension of the 2-D benchmark operator and prolifically used Canny (1986) edge detection technique. The statistical surface detection model is also compared against an oriented filter method, the 3-D steerable filter method of Aguet et al. (2005). The analysis uses a comprehensive dataset of synthetically created images where the interfaces are mathematically defined such that a ground truth reference solution can be used for comparison in objectively analysing the results. The criteria by which the methods are assessed are the resolving power of the detection algorithms against a single flat interface. A topological analysis where the surfaces are not uniform in structure. Finally on images which combine a wide variety of region types present in the same image volume.

**Chapter 6** introduces a real world application case study to verify whether the synthetic evaluation can be relied up on and has generalisability and transferability, here a qualitative approach is undertaken analysing surface detection filtering of real Multi-Model MRI data. The dataset is comprised of real MRI brain scans of paediatric patients with Pilocytic Astrocytoma (PA) pathologies present in the brain. The effectiveness of the different surface detection methods are analysed for their suitability of

locating the interfaces between healthy brain tissue and PA tumour pathologies across independent T1-weighted, T2-weighted and contrast enhanced T1-weighted MRI brain scans. As access to full 3-D isotropic scans for each modality is not always available in a clinical setting, these evaluations allow for independent assessment of the surface detection methods across individual modalities, and as such identify a preference for each modality. The statistical methods are again compared against the baseline 3-D Canny and Steerable methods.

**Chapter 7** explores the results, identifying the key characteristics of each filtering method, and discusses the correlation between the synthetic and real data results. Noting where the methods are effective and where they are not, identifying the problems and limitations. Finally discussing where improvements over the current standard in surface detection are achieved.

**Chapters 8** presents an overall conclusion of the findings in this body of work. Exploring the key contributions.

### 1.2.1 Contributions

- Development of two architectures for 3-D statistical methods of surface detection.
- Development of a novel 3-D orientation filtering method for 3-D Non-maximum Suppression.
- An evaluation of statistical surface detection methods against a baseline of existing gradient techniques.
- Development of a framework for evaluating surface detection methods using synthetic 3-D data.
- Development of an Efficient Pairing Strategy(EPS) for one to one correspondence matching between surface maps and reference images, which uniquely allows for computationally efficient and accurate performance metrics to be applied to large 3-D datasets.
- Qualitative analysis of statistical methods for tumour boundary delineation, in Multi-modal MRI data.
- Application of a confirmation and validation framework which assess correlation between synthetic and real image testing domains.
- Recommendations for the application of statistical surface detection

The following papers have been published as part of this work:

- **S. Smith** and I. Williams, "A Statistical Method for Improved 3D Surface Detection," in IEEE Signal Processing Letters, vol. 22, no. 8, pp. 1045-1049, Aug. 2015. (Chapter 3 and 5)
- **S. Smith** and I. Williams. 2015. A statistical method for surface detection. In Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM '15). Eurographics Association, Goslar, DEU, 217. (Chapter 3)
- **S. Smith** and I. Williams. 2020. Efficient One-to-One Pair Matching for 2-D and 3-D Edge Detection Evaluation. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, (Chapter 4)

## 1.3 Summary

Surface detection is a classical process which requires no prior knowledge of a 3-D image. These processes extract surface information from volumetric image data, which can be further utilised by medical image analysis tasks. Optimal performance for the detection of surfaces is desirable, however statistical surface detection methods which could potentially offer superior performance over traditional surface detection methods have not yet been evaluated. This thesis presents a model for statistical surface detection and an evaluation framework for robustly analysing surface detection methods.

# Chapter 2

# Image Filtering in 2-D and 3-D

## 2.1 Introduction

Vision is one of the major resources humans possess in order to perceive and understand their environment and computer vision is a field of study that aims to imitate aspects of the human vision system in order to electronically perceive images. Not all processes need a complex understanding of the image content, a number of computer vision processes include tasks such as noise filtering, image sharpening, contrast enhancement, thresholding, edge filtering and surface filtering. Typically, no information other than the magnitude (intensity) of the image pixels or voxels are required to complete these tasks. These are considered to be 'low-level' processing operations, and often the output can be used to aid more complex 'high-level' operations which require some understanding of the image content, such as segmentation, object detection and classification.

Edge filtering is a widely utilised 'low-level' image processing task applied to 2-D data, this process extracts edge features from an image which describe the interfaces between different image regions. Thus, edges provide information on the structure contained in 2-D images and this information has been utilised for a number of tasks such as finding objects (Biederman, 1987), separating the foreground from background (Gruenwedel et al., 2011), and segmenting regions of interest in an image (Jamil et al., 2011).

Computer vision tasks are not only applied to 2-D digital images, they can also be applied to image data with three spatial dimensions. Edge filtering in 3-D images is not the same as 2-D. If edges describe the interfaces between two or more distinct regions in 2-D data, it is a surface feature which describes an interface in 3-D data, and it is therefore surfaces, not edges which provide information about the 3-D structure contained within the data.

Many surface filtering algorithms are in fact extrapolations of 2-D edge filtering methods which have been converted to perform the task in 3-D and these include techniques such as the 3-D Sobel Operator, the 3-D Canny Surface detection algorithm derived from the Canny (1986) edge detection algorithm, and 3-D Steerable filters (Aguet et al., 2005) which were developed from 2-D Steerable filters (Freeman and Adelson, 1991).

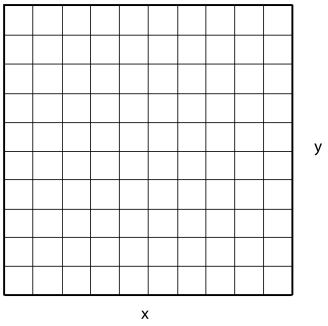
While a number of surface filtering techniques exist, the literature regarding edge filtering is more rich. Since many of the surface filtering algorithms are extensions of 2-D edge filtering methods, the problems and developments which have driven surface detection design are largely based on the issues which affect edge detection, therefore to understand the developments made in detecting surfaces, edge detection methods also require discussion.

Section 2.2 of this chapter describes the fundamentals of digital images, defining various features such as edges and surfaces. Section 2.3 then identifies the key problems associated with extracting accurate edge information, while section 2.4 and 2.5 tracks the major developments in 2-D edge detection technology, describing how the advancements overcome some of these limitations. Section 2.6 introduces a statistical approach to 2-D edge detection, discussing the various advantages and disadvantages offered by this approach. Section 2.7 then introduces surface detection for 3-D data, highlighting the parallels with the design challenges faced by 2-D operators, and finally providing a rationale for the development of 3-D statistical surface detection.

## 2.2 2-D Digital Image Fundamentals

A digital image is a grid of numbers which can be used by an electronic display to create a visual picture. More specifically a digital image is the visual manifestation of a digitally encoded representation of the visual characteristics of an object or scene (*FADGI* 2018). Thus, an image is a discrete multidimensional signal that conveys spatial information corresponding to specific measurements or values, such as photons collected by a camera sensor. Images predominately contain two spatial dimensions displayed on an $x$ and $y$-axis, two dimensions is typical since display technologies are usually a flat 2-D space, these image modalities include digital photography, ultrasound and x-ray photography amongst others.

A picture element (pixel) is defined as a value associated with the Cartesian coordinates ($x$, $y$) in a 2-D grid. Each pixel possesses a numeric or binary value specific to that spatial location. The pixels are positioned in a grid structure that preserves the spatial relationship of the acquired measurements or values (Fig. 2.1). For example, neighbouring light measurements from a camera sensor will be positioned as neighbouring pixels




(a) 2-D image Structure

Figure 2.1: Illustration of grid layout for 2-D images

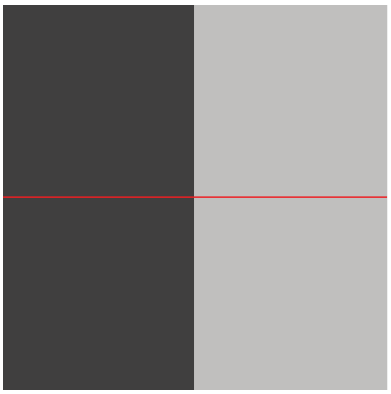
in a digital image. The number of pixels in the grid structure determines the image resolution.

### 2.2.1 Edge Feature

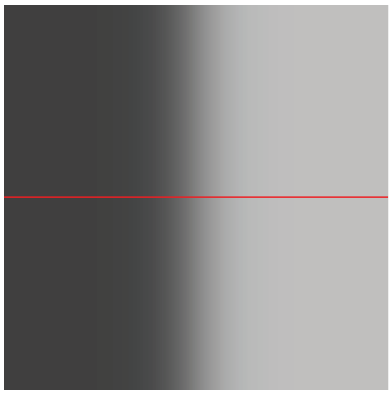
A number of different features or properties can be obtained by processing digital images, one highly sought after feature is an edge. An edge describes the interface which exists between different regions within the image. Visually, an abrupt change in pixel intensity can be recognised as a discontinuity in image brightness, these discontinuities are typically interpreted as edges by humans. When the image intensity gradient is steep, there is an abrupt change in intensity, and is recognised as an edge, whereas a more gradual change in intensity not recognised as having a defining edge, this is illustrated in Fig. 2.2.

'Edges' are always a line or curve of single pixel width and can be of arbitrary length subject to the restraints of the image resolution. The most prevalent edge detection techniques are those which target the detection of abrupt changes in luminosity.

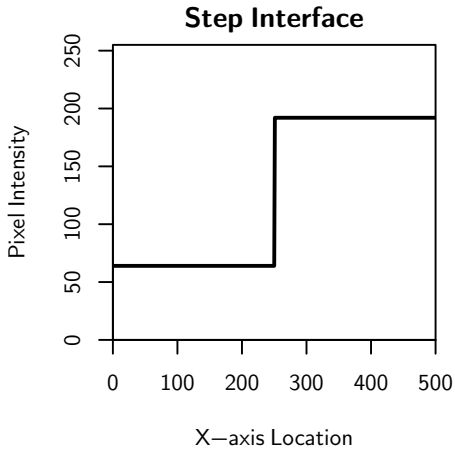
In order to identify what could constitute as an edge, the definition of an image region should first be established. A region describes a collection of pixels where image properties or features are homogeneous, such as the aforementioned properties of luminance, colour and texture. Fig 2.3 displays examples of these region profile types and the corresponding edge interface that exists between them. The edge therefore describes the spatial position in an image where these properties change.



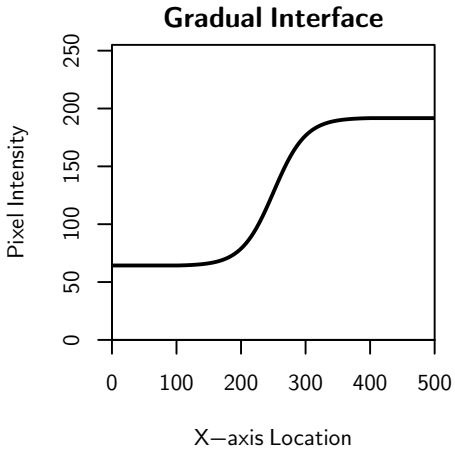
(a) Step change in intensity



(b) Gradual change in intensity

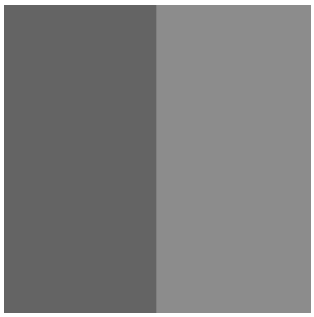



(c) Step Edge profile

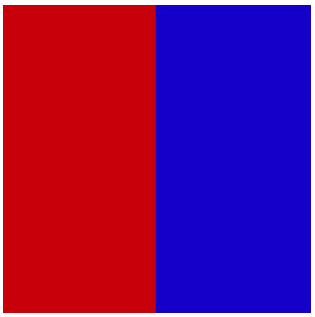



(d) Gradual Edge Profile

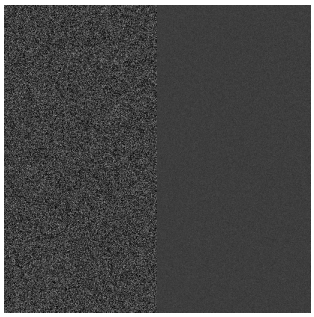
Figure 2.2: The intensity edge profiles of a step edge and gradual edge interface. The red line indicates location of pixel values evaluated. Edges formed from abrupt changes in intensity are easily defined, however when the change is gradual the position of the edge becomes less clear.



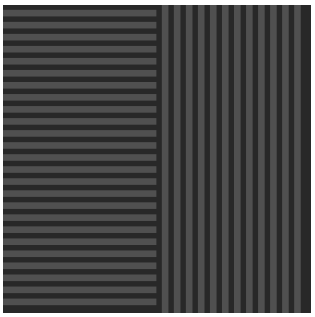
(a) Intensity



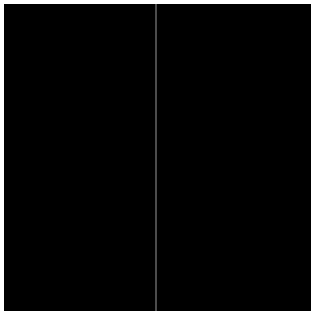
(b) RGB Colour



(c) Stochastic texture



(d) Regular texture



(e) Edge Location

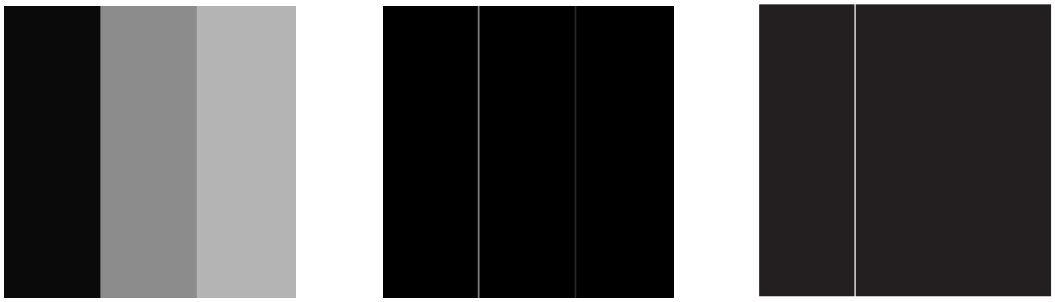
Figure 2.3: Different types of interfaces, each image contains 2 adjacent regions with properties defined by a) Intensity, b) Colour, c) Stochastic texture, d) Regular texture, e) Location of edge

## 2.3 Filtering Processes

In isolation, only a single feature describes a pixel, its intensity value at a single position in space. Operations which do not take into the context the surrounding pixels are known as point operators and these techniques perform vital tasks such as thresholding, contrast adjustment, scaling and histogram equalisation (Nixon and Aguado, 2012). Other filtering processes, such as image sharpening, smoothing and edge filtering require a different approach that accounts for the spatial relationship of the pixel intensities. Since edges are a structural feature assessing local changes, a point operator is unsuitable as the pixel neighbourhood must be evaluated in order to examine local features.

### 2.3.1 Edge Filtering Processes

Edge features are obtained using edge detection algorithms, which are a class of algorithms which resolve local changes in image properties. The resultant output of an edge detection algorithm is known as an edge map, an example of which is shown in Fig 2.4e. The edge map is a 2-D array which provides information about the position and magnitude of the edges within the image. The degree of change between region profiles is typically represented by the magnitude of the edge, this is illustrated in Fig 2.4. The magnitude of an edge is an important property of the edge feature since it is common practice to apply post processing techniques or higher level processes to edge map data, and the magnitude of edges can greatly effect the outcome and success of these processes.



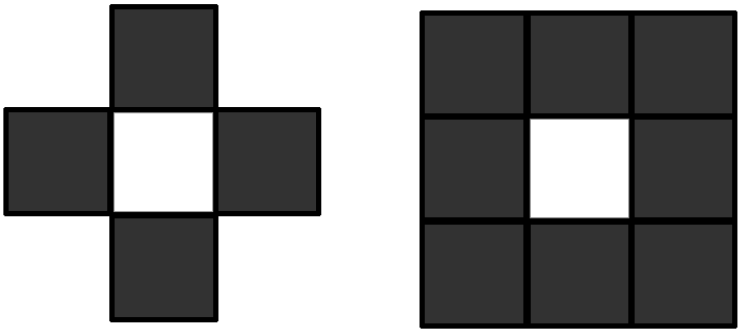
(a) Left 10, middle 140, right 180 (b) Edge Strength (c) Edge Thresholded

Figure 2.4: Gradient operators resolve edges by approximating the 2-D intensity gradient function of an image. The steepness of the gradient relates to the edge strength, a greater intensity differential will result in an edge with a larger magnitude. In this example the left edge is resolved with an intensity difference of 130, while the right edge is resolved with an intensity difference of 40. The result is a relatively "stronger" edge on the left, and a "weaker" edge on the right. The strength of edges is important when classification of edge and non edge points is taken into consideration, here the weaker edges are discarded during a thresholding process

## 2.4 2-D Neighbourhood Operators for Edge Detection

Neighbourhood operators process pixel values in a local neighbourhood that surrounds a point in an image. They can be classified according to type of domain, whether or not they are recursive and the type of neighbourhood (Haralick, 1981). Edges are a type of spatial feature, they provide information about the structure of an image, therefore in order to determine whether an edge exists or not, the spatial properties of surrounding pixels require examination. This can be achieved using a 2-D neighbourhood operator(Nixon and Aguado, 2012). Neighbourhood operators are known for various filtering capabilities, such as image smoothing, sharpening and edge filtering.

For 2-D neighbourhood operators, neighbourhood windows (masks) typically come in 2 spatial configurations of varying sizes, those with 4 pixel connectivity and those with 8 pixel connectivity. Fig 2.5 and Eq 2.1 identifies each of the spatial configurations for a $3\times3$ local neighbourhood window. 4-pixel connectivity selects the 4 adjacent pixels above, below and to each side of the pixel being evaluated, while 8-pixel connectivity includes all the surrounding pixels in the $3\times3$ space.



(a) 4 Pixel Window (b) 8 Pixel Window

Figure 2.5: Illustration of the two basic types of 2-D neighbourhood operators. Those with 4-connectivity are formed from pixels adjacent to the central pixel, and those with 8-connectivity, with both adjacent and orthogonal pixels.

$$h = \begin{bmatrix} h_{i-1,j-1} & h_{i,j-1} & h_{i+1,j-1} \\ h_{i-1,j} & h_{i,j} & h_{i+1,j} \\ h_{i-1,j+1} & h_{i,j+1} & h_{i+1,j+1} \end{bmatrix} \tag{2.1}$$

When processing 2-D images, $h$ is the index locations which define an 8-connectivity mask.

## 2.5 2-D Classical Methods

The principal development in edge detection involved the introduction of a first order derivative operator (Roberts, 1963). First order derivative operators attempt to approximate the image intensity gradient function. Since an abrupt change in relative pixel intensity between neighbouring pixels can be thought of as an edge, computing the image intensity gradient corresponds to resolving edge features in the image. This process produces a high value output when the rate of change of pixel intensity is high, and a low value when the rate of change is small (Fig. 2.2).

For the purpose of edge detection, it is typical for a convolution based mathematical framework to be used in order to approximate the image intensity gradient. Edge detection using convolution is non-recursive and requires two components, an image ($I$), and a filter kernel ($h$). The filter kernel is a 2-D array of coefficients mapped to a neighbourhood operator which has been specifically designed to output a high value when situated on an edge, and a low value when not at an edge location after the kernel is convolved with an image. The kernel traverses all legal pixel neighbourhood positions and in each location the resulting value is written back to the centre position of the

neighbourhood. The overall result is an edge map, which gives information to both the location and magnitude of edges within an image.

$$G(i,j) = \sum_{k,l} I(k,l)h(i+k, j+l)$$

$$G = I * h \tag{2.2}$$

Where $G(i,j)$ is the edge map, $I$ is the input image signal and $h$ is the edge filter kernel.

Various digital kernels (also known as a filter mask) have been proposed to approximate the image intensity gradient function. Computation of gradients involves a non-linear combination of approximations to the partial derivatives in orthogonal directions (Prewitt, 1970). The convolution process determines an output pixel value to be a weighted sum of input pixel values (Eq. 2.2), this is an arithmetic process, and it requires an arithmetic non-recursive neighbourhood operator, typically the 8-pixel connectivity variety is used, however these can be expanded to larger neighbourhood scales and the first edge operator used only 2×2 kernels (Roberts, 1963) .

### 2.5.1 Roberts Cross Operator

The first method and one of the simplest implementations of edge detection using a convolution framework was the introduction of the Robert's Cross edge operator (Roberts, 1963). This method uses two $2 \times 2$ directional filter kernels. Whereby the sum of the squares of the differences between diagonally adjacent pixels is calculated. To achieve this the image is convolved with two mirrored kernels which are the smallest possible difference filters to compute the gradients with a shared centre point (Eq. 2.3).

$$h_x : \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } h_y : \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{2.3}$$

If $G_x$ is an image formed from convolving $I$ with $h_x$, and $G_y$ is an image formed from convolving $I$ with $h_y$ (Fig. 2.6). These can be considered directional edge maps or vectors which provide information on the direction and magnitude of an image intensity gradient. The overall edge map $G_{(i,j)}$ is calculated as the square root of the sum of the squared vector values (Eq 2.4):

$$\nabla G_{(i,j)} = \sqrt{G_x + G_y} \tag{2.4}$$

Since direction of the image gradient is perpendicular to the direction of an edge interface, directional information can be used to assist with a number of post processing techniques aimed at improving edge classification. It is therefore common for the direction of the gradient to also be calculated during this operation. The direction of the gradient is defined in Eq 2.5:

$$\theta_{(x,y)} = \arctan\left(\frac{G_y}{G_x}\right) - \frac{3\pi}{4} \tag{2.5}$$

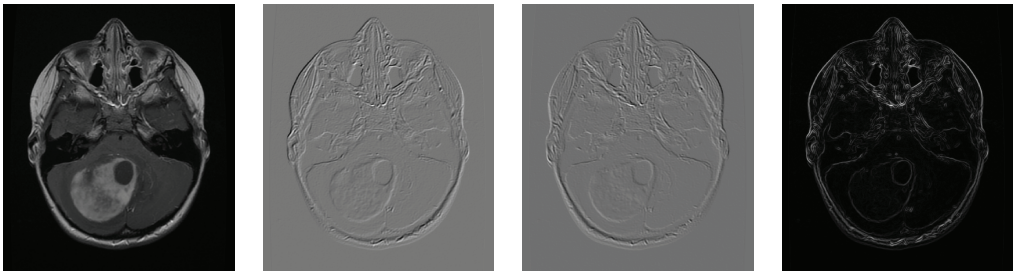
Due to the small kernel size, the Roberts cross method is highly sensitive to image noise. This is due to the fact that fluctuations in image intensity are assessed at a scale of 4 pixels (2×2) where small variations in pixel intensity influence the output result by a greater amount than if the pixels were assessed over a larger spatial region, where typically image properties are more uniform. Since image noise is often present in real images, the Roberts operator may be rendered unsuitable for many image processing tasks.

Since the introduction of the Roberts operator, it is more common for edge detection methods to use larger neighbourhood masks. Additionally, these operators are typically of odd integer size, providing a discrete central pixel in the kernel which corresponds to the neighbourhood in the image being convolved. This fundamentally differs with Roberts cross operator, since it is of even size (Eq. 2.3), therefore the Roberts result is an approximation of the gradient between two pixels instead of at a pixel.

### 2.5.2 Prewitt Operator

The sensitivity of the Roberts Cross operator to noise led to the development of the Prewitt edge detection filter (Prewitt, 1970). Prewitt found that a more precise estimate of the gradient is obtained by fitting a quadratic surface over a 3×3 neighbourhood by least squares, and then computing the gradient for the fitted surface. The filter kernels of the Prewitt operator are defined in Eq.2.6.

The procedure of the Prewitt edge filter is comparable to the Roberts operator, both methods employ 2-D convolution with filter kernels which approximate the image



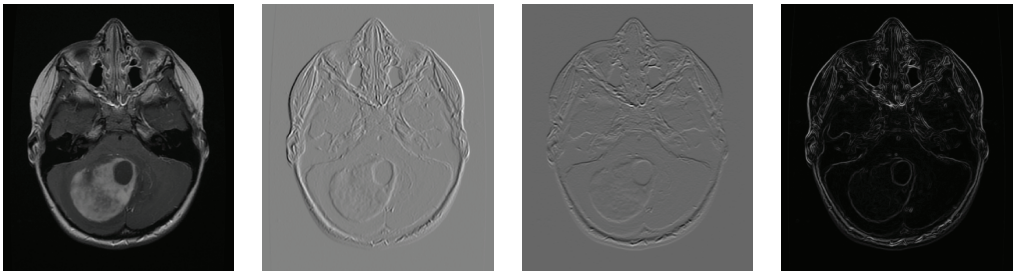
(a) Test image (b) $G_x$ (c) $G_y$ (d) $\nabla I_{(i,j)}$

Figure 2.6: Computation of the gradient using the Roberts operator involves the non-linear combination of the approximations $G_x$ and $G_y$ to the partial derivatives in orthogonal directions. $G_x$ is the approximated gradient in the $x$-dimension and $G_y$ is the approximated gradient in the $y$-direction, while the output is the combination of the two, which provides a 2-D gradient map, revealing image edges.

$$G_x : \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \quad G_y : \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \tag{2.6}$$

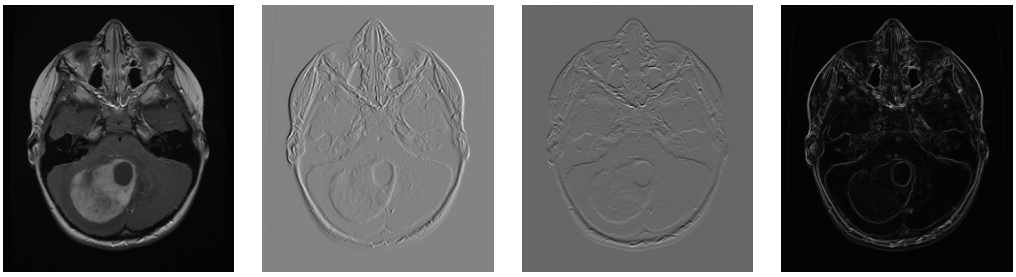
gradient. The Prewitt method differs from the Roberts cross operator in that it uses two 3x3 kernels (Eq. 2.6), one for approximating the gradient in the $x$ dimension and one for the $y$ dimension. These kernels are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid. The image is convolved with each of the filter kernels and the two filter responses are combined using Eq. 2.4 to form an edge map.

With 9 positions per filter kernel instead of 4, it is less sensitive to image noise when compared with the Roberts Cross operator. However, as a consequence this results in detected edges with greater variation in edge magnitude (Fig 2.4). The relative strengths of edges is an important consideration. There is typically a trade off in the performance of edge filters regarding sensitivity. The greater the sensitivity, the greater the ability of the filter to resolve weaker boundaries. However, typically there will be a higher propensity to produce spurious or unwanted edge points, often characterised as “noise”. In images where weaker boundaries are sought after, a more sensitive filter is required, if strong edges and spurious edge points are also present, it is a non trivial process to isolate the weaker edge information. Typical post processing edge classification techniques such as thresholding may deem weaker edges as unimportant, thus classify them as ‘non-edge’ points. However, weaker edges can be indicative of some underlying structure or detail that could be of importance, so should not always be discarded without consideration.



(a) Test image (b) $G_x$ (c) $G_y$ (d) Output

Figure 2.7: Computation of the gradient using Prewitt algorithm. $G_x$ is the approximated gradient in the $x$-dimension and $G_y$ is the approximated gradient in the $y$-direction, while the output is the combination of the two, which provides a 2-D gradient map revealing image edges.



(a) Test image (b) $G_x$ (c) $G_y$ (d) Output

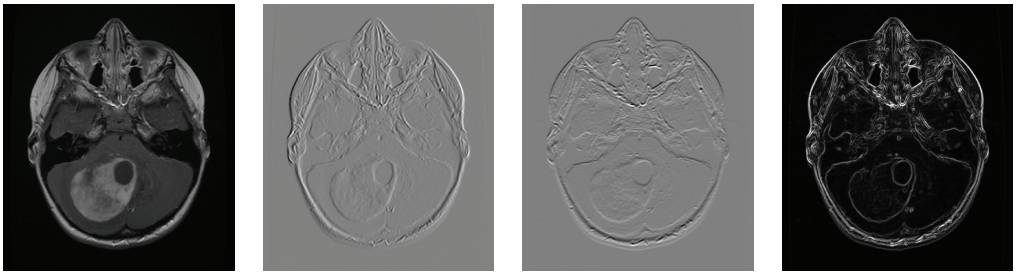
Figure 2.8: Computation of the gradient using Sobel-Feldman algorithm. $G_x$ is the approximated gradient in the $x$-dimension and $G_y$ is the approximated gradient in the $y$-direction, while the output is the combination of the two, which provides a 2-D gradient map revealing image edges.

### 2.5.3 Sobel-Feldman and Scharr Operators

The Sobel-Feldman operator (Sobel and Feldman, 1968) was developed with the intention of creating an efficient computable gradient estimate which would be more isotropic than the then popular "Roberts Cross" operator. More commonly known as a Sobel filter it was the most popular edge detection operator until the development of edge detection techniques with a theoretical basis, proving popular because overall it achieved better performance than other contemporaneous edge detection operators, such as the Prewitt operator, making it the preferential choice (Nixon and Aguado, 2012). The advantages are it offers a similar level of noise suppression as the Prewitt operator, but the added emphasis to the central row and columns provides more strongly defined edges. Again, this operator utilises isotropic $3 \times 3$ filter kernels (Eq. 2.7).

$$G_x : \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad G_y : \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{2.7}$$

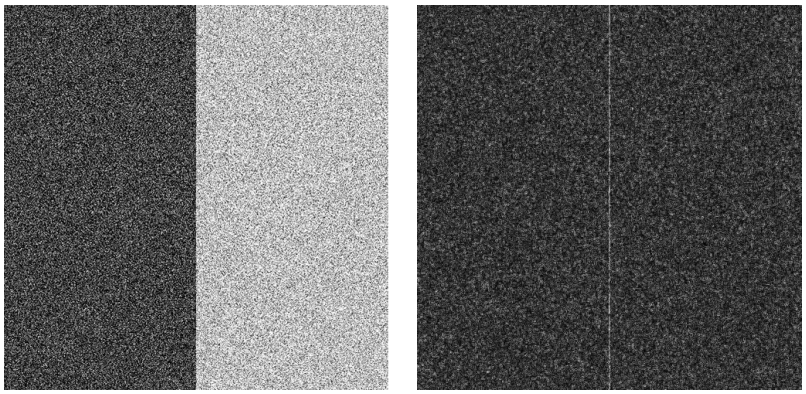
For the previously defined operators the approximation of the image gradient remains crude and inaccurate (Jähne et al., 1999). However due to their simplicity and speed, they remain prevalent in the field of image processing. There have been many attempts to improve upon the Sobel filter using a number of refinements. Scharr (2000) realised that the pure central difference operator from Sobel does not have perfect rotational symmetry, by optimising the kernels Scharr was able to minimise the weighted mean squared angular error. The result are filter kernels with even more emphasis to the central row and column, and an operator that more accurately resembles a first order derivative by 3 orders of magnitude (Eq. 2.8), providing edges of greater strength and noise rejection.



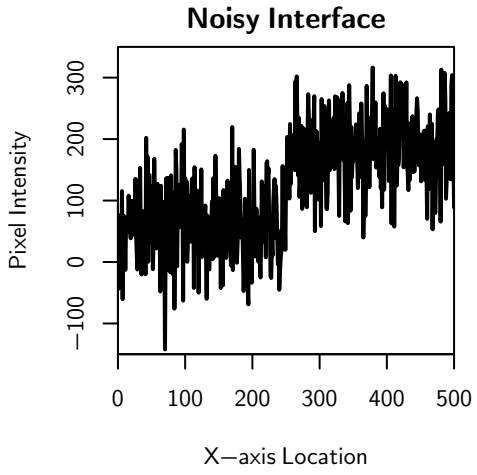
(a) Test image (b) $G_x$ (c) $G_y$ (d) Output

Figure 2.9: Computation of the gradient using Scharr algorithm. $G_x$ is the approximated gradient in the $x$-dimension and $G_y$ is the approximated gradient in the $y$-direction, while the output is the combination of the two, which provides a 2-D gradient map revealing image edges.

$$G_x : \begin{bmatrix} 3 & 0 & -3 \\ 10 & 0 & -10 \\ 3 & 0 & -3 \end{bmatrix} \quad G_y : \begin{bmatrix} 3 & 10 & 3 \\ 0 & 0 & 0 \\ -3 & -10- & -3 \end{bmatrix} \tag{2.8}$$



(a) Noisy Interface (b) Sobel filter response

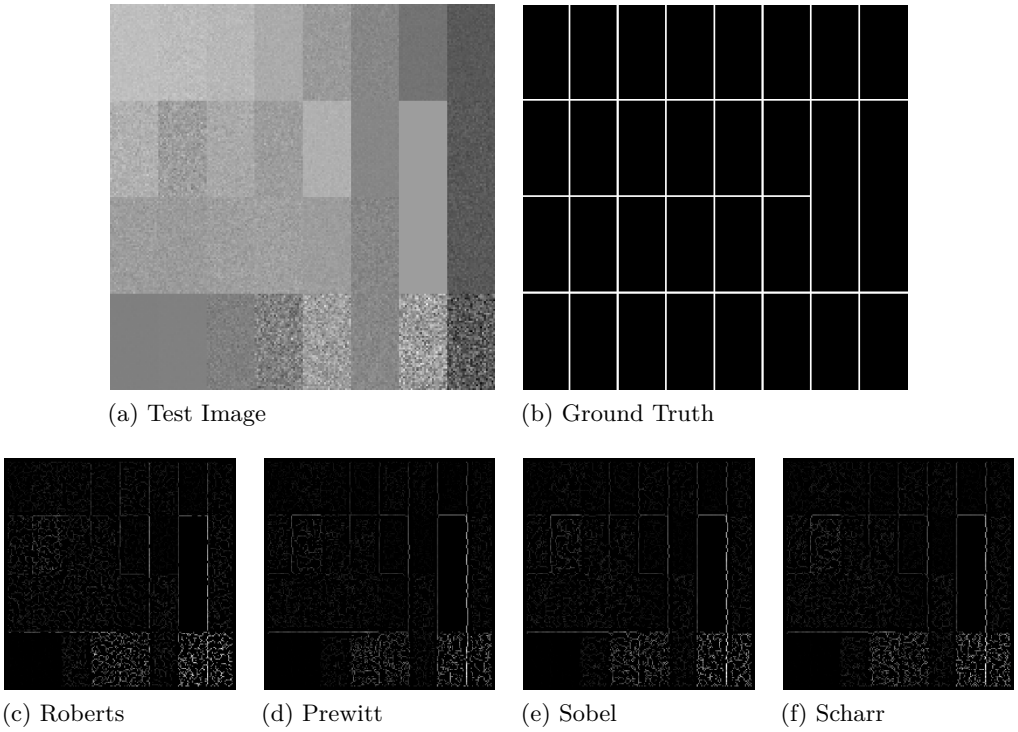



(c) Edge profile

Figure 2.10: Gradient method limitations, here it can be seen that the abrupt changes in intensities across the image space results in erroneous responses in addition to the central edge, which are not considered the edges of interest for this interface type.

### 2.5.4 Limitations

There exists a number of issues with first order derivative methods. Namely the effect of non-homogeneous regions within an image such as noise (Fig 2.10). Noise in an image resembles abrupt changes in pixel intensity, since gradient operators resolve edges using this principle, image noise results in spurious results in the edge map (Bovik and Munson, 1985; Bovik, 1987). If the steepness of the gradient of resolved edge is greater than the gradient of noise resolved in the image, the spurious points can be eliminated using a thresholding technique. However, if the noise response is equivalent in magnitude to that of the edges, removing the spurious responses and retaining the edge is a non-trivial task, since the simple thresholding technique is no longer suitable under these conditions.



(a) Test Image (b) Ground Truth

(c) Roberts (d) Prewitt (e) Sobel (f) Scharr

Figure 2.11: Presented here are some of the typical shortcomings of the aforementioned gradient based edge detection techniques. The results indicate a number of missed edge points, signified by discontinuities in the edges which should be continuous in this example, and the production of spurious responses, which occurred due to the high intensity variance properties in regions of the image.

Fig. 2.11 shows the responses of the gradient operators discussed thus far. Here the filters are presented with a test image which exposes the limitations that affect the performance of the edge detection filters. The image is a 2-D composition of several regions with distinct mean and standard deviation properties.

Fig 2.11 illustrates several missed or unconnected edge interfaces in the various filter responses. The high variance nature of some of the regions also simulates an image with noise leading to spurious responses in the edge map. The fact that there are several interfaces of varying edge strengths reveals another limitation of gradient operators. Since the output is an overall measurement of edge strength and not a relative one, this manifests itself as an embedded bias of the operator to emphasise interfaces such that a step change profile takes precedence. While this is sufficient for a number of tasks, when more poorly defined edges need to be detected, they are often lost during post processing such as thresholding methods, which classify edge maps into edge and non edge points. A visual inspection of these results show missing boundaries for all operators in addition to unconnected edge interfaces, notably from the Roberts operator as well as further spurious edge points (Fig. 2.11).

### 2.5.5 Canny Edge Detector

The Canny edge detector (Canny, 1986) is the perhaps the most rigorously defined operator and is the most widely used (Ding and Goshtasby, 2001). While incremental improvements were made by the likes of Prewitt, Sobel-Feldman and Scharr, derivative based operators generally do not perform well on images where noise is present (Suzuki et al., 2003), this problem is also presented in Fig 2.10. Key improvements in edge detection were achieved when Canny (1986) introduced an analytically optimal edge detector. Canny defined three criteria which an optimal edge detector must satisfy and these criteria were used to develop a total error cost function, Canny applied variational calculus to the cost function to determine the optimal linear operator for convolution with the image. The optimal solution was shown to be a very close approximation of the first derivative of a Gaussian. The method also introduces additional stages of processing to the edge detection process. First a pre-processing Gaussian smoothing stage is applied, followed by the edge gradient filtering stage. Non-maximum suppression in a direction perpendicular to the edge direction is then applied, the effect of this is to produce an edge of single pixel thickness. Finally a two level hysteresis thresholding stage is applied to remove weaker edges. Since its introduction, Canny's (1986) edge detection method has been extensively utilised. The precise criterion which Canny (1986) laid out for optimal edge detection are as follows:

1. Detection of the edge point.
2. Accurate positioning of the edge point
3. No duplicate or spurious edge points.

#### 2.5.5.1 First criterion

Detection of edge points: The amplitude signal to noise ratio (SNR) of the edge gradient has to be maximised in order to obtain a low probability of missed edge points, while simultaneously achieving a low probability of incorrectly marked edge points. These probabilities are monotonically decreasing functions of the output signal to noise ratio, thus maximising the SNR achieves this. This criterion is characterised by the sensitivity of the filter. Increased sensitivity allows for improved detection, however, it increases the propensity of falsely detected edge points, or spurious responses.

#### 2.5.5.2 Second criterion

Accurate positioning of edge points: The edge points should be as true to the centre of the edge interface as possible. Edges should be the local maxima of the gradient function of an edge. To achieve this, the Canny method of edge detection applies extra stages pre and post the edge filtering stage. Firstly a Gaussian convolution filter is used

to smooth the image(Eq. 2.9), removing unwanted noise. The standard deviation of the filter can be manipulated to remove more noise, at the expense of image detail, or contrarily remove less noise to spare image detail. Following the smoothing stage, an approximation of gradient of the intensity function is computed. This can be any 2-D gradient filter, such as the Sobel, Prewitt or Scharr Filters. Thirdly an edge thinning stage known as non maximum suppression is applied. This technique suppresses all gradient values to zero with the exception of the local maxima, which corresponds to the position with the most abrupt change in intensity. The result is an edge map image consisting of edges interfaces of single pixel thickness.

$$G(x, y) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.9}$$

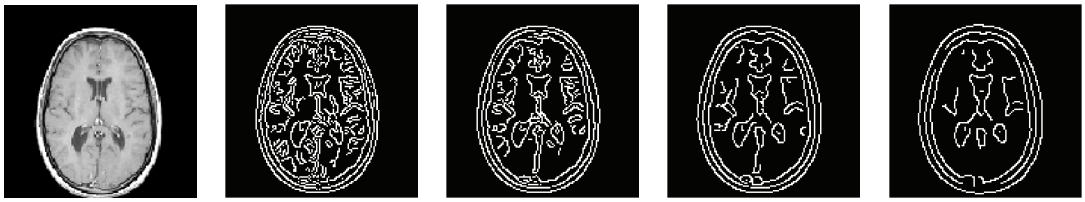
Where $x$ is the distance from the origin in the horizontal axis, $y$ is the distance from the origin in the vertical axis, and $\sigma$ is the standard deviation of the Gaussian distribution.

#### 2.5.5.3 Third criterion

Canny's third criteria states that an optimal detector does not produce duplicate edges and spurious responses. Gaussian smoothing assists in this by removing the higher frequency information likely to produce a spurious response or duplicated edge. The final stage of the Canny edge operator uses thresholding with hysteresis to classify edge and non edge points. Spurious edge points, tend to be unconnected, hysteresis thresholding prioritises connected edge points while removing weaker unconnected edge points.

Classifying an edge point when an abrupt change in luminosity occurs is straightforward, however with a more gradual shift in luminosity, the position of the interface between the light and dark regions is not as easily defined (Fig. 2.2b). By increasing the standard deviation and kernel size of the smoothing filter prior to computing the gradient, one can resolve edges that occur over a larger scale. This scale parameter also has the advantage of suppressing image noise and unwanted texture, however as can be seen in Fig. 2.12, this leads to other details which may be of importance being suppressed.

Due to the mathematical rigour in defining an optimal edge detection technique, the Canny method is widely adopted and is highly represented in the literature for a range of different applications (Ali and Clausi, 2001; Liu and Jezek, 2004; Chang et al., 2008; Punarselvam and Suresh, 2011; Ramya and Babu, 2015; Kaur and Kaur, 2016; Nikolic



(a) original (b) $\sigma 0.5$ (c) $\sigma 1$ (d) $\sigma 1.5$ (e) $\sigma 2$

Figure 2.12: Gaussian smoothing is the first stage of the Canny edge detection process, the standard deviation($\sigma$) of the Gaussian filter determines the amount of smoothing, by increasing $\sigma$ the amount of noise in the output can be reduced, however as a consequence finer image details are also lost in the edge map.

et al., 2016; Selvakumar and Ganesh, 2017; Tahmid and Hossain, 2017) and at the time of writing has 493 patent citations, making it the ideal reference technique for which all novel edge detection approaches should be evaluated against.

### 2.5.6 Limitations

While the Canny edge detection method offered strong improvements over the previous derivative based methods, there remains some limitations. Image noise and texture still negatively affects the performance of the Canny method (Wang Xiao and Xue Hui, 2010; Rong et al., 2014). This was also shown by Williams et al. (2014), here non linear methods of edge detection, such as statistical edge detection offered improvements over the Canny method on images containing region profiles with a high variance component, and on real histological images where image texture determines edge location.

The preprocessing Gaussian smoothing stage of the technique can remove small edge details, this effect is proportional to the standard deviation of the Gaussian kernel. In addition, applying a smoothing filter changes the location of an edge, thus can lead to excessive localisation error. When images contain noise, or non uniform regions, a higher degree of Gaussian filtering is typically required, compounding the issue (Smith and Brady, 1995). The result of Canny Edge detection is binary, however processes which utilise the results of edge detection such as segmentation, the magnitude of the edge is often an important component used in the process (Sharma and Aggarwal, 2010), but this component does not exist in a binary result. The method introduces parameters such as the Gaussian kernel size, and two hysteresis thresholding levels which need to be defined, this increases the degrees of freedom, and therefore the method requires either human input or additional processes for parameter selection to produce an optimal result. Due to the Gaussian smoothing stage blurring the image, the method often fails to accurate resolve corners and junctions, leaving open ended edges and missing junctions (Smith and Brady, 1995).

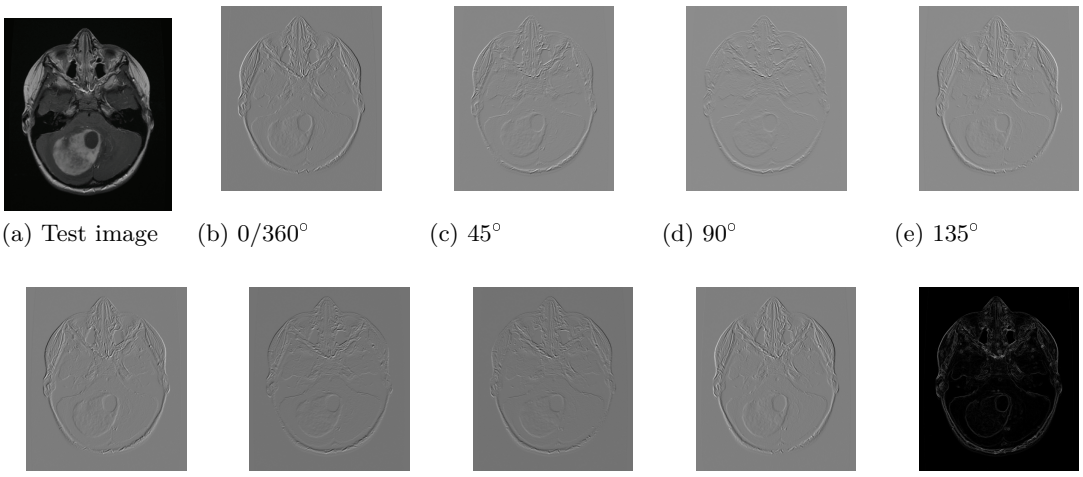
### 2.5.7 Oriented Filters for Edge Detection

Edge detection methods are not limited to derivative based operators. Kirsch (1971) developed a non linear edge detection method for the determination of constituent structures in biomedical images. The Kirsch Compass operator finds the maximum edge strength in a select few predetermined directions, at 45° intervals (N, NE, E, SE, S, SW, W, NW), the magnitude of the edge is calculated as the maximum magnitude difference across the different directions. While the edge direction is the 'compass' direction of the kernel which produced the greatest magnitude. This method was simplified by Robinson (1977) for improved efficiency, by using symmetrical filter kernels, the operation halves the number of directions required to determine the edge, since the opposite filters produce the same magnitude result but in the reverse direction. One of the issues with these methods is the poor connectivity of edges, Nevatia and Babu (1980) later adopted an oriented filtered approach similar to the compass methods of Kirsch (Kirsch, 1971) and Robinson (Robinson, 1977) proposing improvements. The Nevatia-Babu (Nevatia and Babu, 1980) operator functions by convolving a given image with masks corresponding to ideal step edges in a selected number of directions. However, here the number of unique filter orientations was increased from 4 to 6. Again the maximum magnitude output determines the edge strength and direction. However, for further improvement over the previous compass methods, the resulting output is thinned by identifying the local maxima, followed by a process of edge linking to form boundary segments using the edge positions and their orientations. Each boundary segment is approximated by a series of piecewise linear segments.

Developed by Freeman and Adelson (1991), Steerable filters further improved upon existing oriented filter methods by presenting an efficient architecture to synthesise filters of arbitrary orientations instead of a fixed number of predefined angles. This allows for a filter to be adaptively 'steered' to any orientation, achieving superior performance with more precise detection. The improvements achieved using this procedure are particularly noticeable in response to line features, where edge detection methods usually produce 2 parallel extrema displaced either side of the line in the output, as a opposed to just one for the line in the correct position, the Steerable method addresses this by being able to adaptively resolve edges and lines simultaneously in the same image, essentially reducing the presence of duplicated edges.

Steerable filters are a combination of derivatives of Gaussian filters, this permits the rapid computation of even and odd (symmetric and anti-symmetric) edge-like and corner-like features at all possible orientations. As a result one of the advantages

of steerable filters are that they are particularly good at accurately resolving curved interfaces.



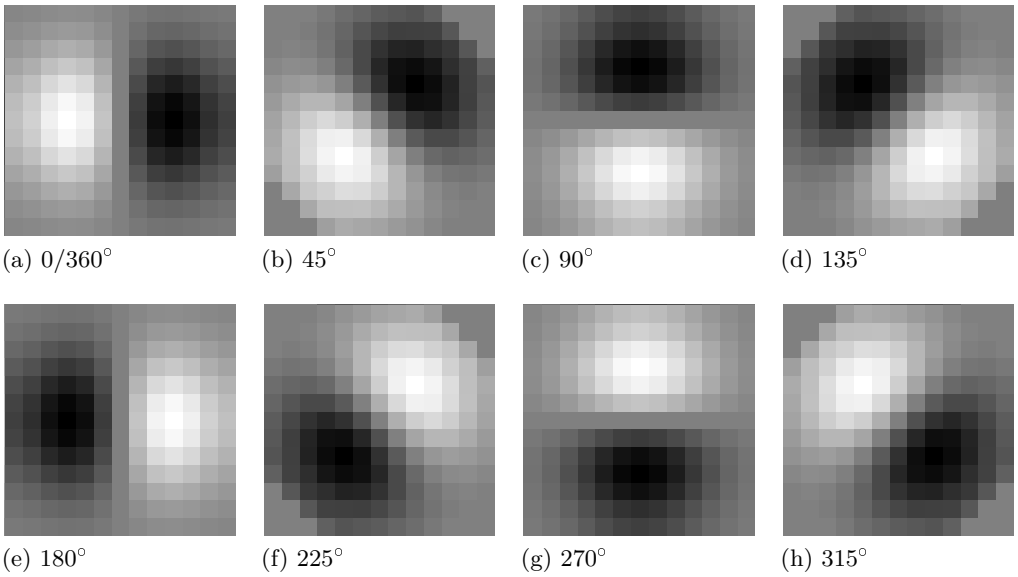
(a) Test image (b) 0/360° (c) 45° (d) 90° (e) 135°

(f) 180° (g) 225° (h) 270° (i) 315° (j) 2-D steerable response

Figure 2.13: Computation of 2-D steerable edge filter. With response of the input to rotated versions of the filter at different angles

Jacob and Unser (2004) further optimised steerable filters for edge detection, the updated design of the filters were based on "Canny-like criteria" which give an optimal response to edges, further suppressing noise and image texture. Jacob and Unser (2004) was able to demonstrate a higher Signal-to-noise ratio than the optimal Canny edge detector algorithm.

The filters are a 1st derivative of a Gaussian function and are rotated through different intervals, an example of the basis filters for edge detection in a steerable filter design are shown in Fig. 2.14. Work by Aguet et al. (2005) introduced the concept of optimising 3-D filters to respond to particular features within 3-D image sets building on the 2-D work of Jacob and Unser (2004), allowing steerable filters to be used for surface detection in 3-D modalities.



(a) 0/360° (b) 45° (c) 90° (d) 135°

(e) 180° (f) 225° (g) 270° (h) 315°

Figure 2.14: Steerable filters use oriented filters which are a 1st derivative of a Gaussian function and are rotated through different intervals, for example 8 different angles.

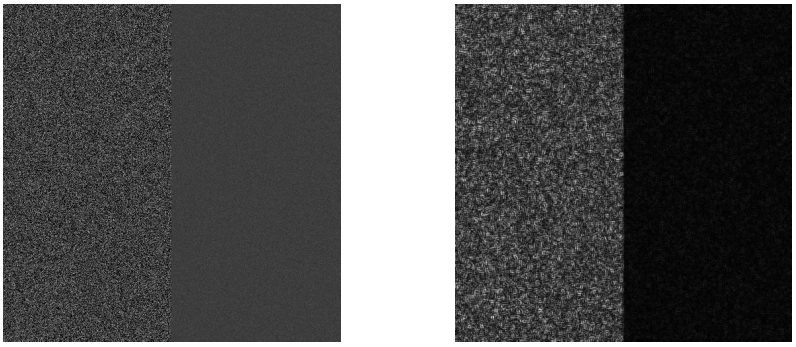
## 2.6 Statistical Approaches

Traditional methods commonly apply a first derivative computation of the image edge to locally assess changes in the intensity profiles of neighbouring pixels such as the aforementioned Roberts, Prewitt, Sobel, Scharr and Canny edge detection techniques. Since it is the intensity function being calculated by these methods, other region properties are not readily resolved using this approach, such as Texture. When textured interfaces are present in an image, an alternative approach which considers the statistical distribution of the data is better suited.

### 2.6.1 Texture Interfaces

Texture is one of the most prominent image attributes in image analysis (Jähne, 1997). Broadly speaking, texture is a local image characteristic that changes according to the scale at which it is observed, these texture regions can be characterised by a specific intensity distribution of pixel values, or a specific spatial positioning of correlated pixels (Fig 2.15a).

Formally defining texture is not straightforward, this is signified by the large number of differing definitions contained within the literature, such as those which have been catalogued by Coggins (1983). Haralick (1979) proposed a structural description of texture, such that texture can be considered an *"organised area phenomenon"* that can be broken down further into *'primitives'* with specific spatial distributions. While



(a) Stochastic texture interface (b) Filter response to stochastic

Figure 2.15: Traditional operators detect edges by measuring abrupt changes in intensity profiles. Spatial statistics over a wide area, such as texture and variance properties are not considered. This can result in edge maps with erroneous points.

Cross and Jain (1983) offered a stochastic approach, describing texture as a *"stochastic, possibly periodic, 2-D image field"*. However, one definition of texture proposed by Petrou and García Sevilla (2006b) considered texture to be *"The variation of data at scales smaller than the scale of interest"*. This definition is particularly useful in an edge detection context since the interfaces where the texture changes from one profile to another is the target for an edge detection algorithm, and not the details smaller than the scale of interest.

One approach to dealing with texture using traditional edge detection methods is to first apply a low pass filter such as an averaging or Gaussian filter, these remove the subtle changes in pixel intensity observed in a texture, however the act of applying filters can shift boundaries and therefore can reduce accuracy in the determining the position of an interface (Williams, 2008).

Real image data typically consists of regions with different texture profiles, images that contain more than one type of texture are known as non-stationary texture images, typically medical image datasets fall into this category (Petrou and García Sevilla, 2006c). Fig. 2.15 illustrates the typical responses to texture obtained when using an intensity gradient based technique. The result can often be edge maps with erroneous points, unconnected edges, and missed interfaces. (Khotanzad and Chen, 1989; Reyes-Aldasoro and Bhalerao, 2006; Petrou and García Sevilla, 2006a; Mirmehdi et al., 2008).

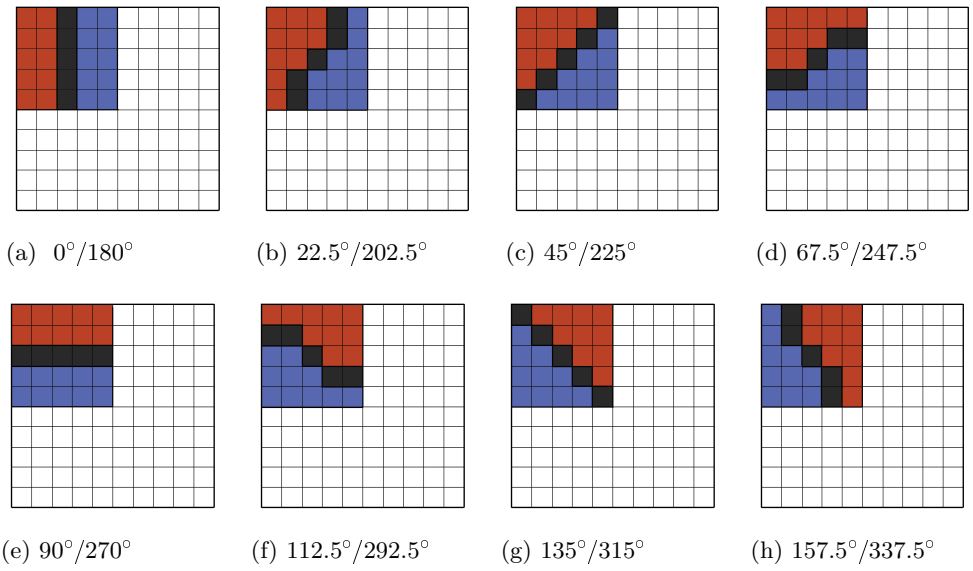
### 2.6.2 2-D Statistical Edge Detection

Statistical edge detection applies a different approach, instead of approximating the intensity function, statistical edge detection resolves interfaces between regions separated

by various statistical properties, it is therefore not limited to resolving only changes in intensity profiles, it can also resolve edges corresponding to changes in texture. This is useful, since commonly in real image data texture is often the main differentiator of image regions such as in CT (Liu et al., 2013), histology (Todman and Claridge, 1997), and MRI (Clarke and Velthuizen, 1995; Bomans et al., 1990). Where this is the case traditional gradient methods for edge detection can fail, resulting in missed edges and spurious responses. Objective performance of gradient and statistical techniques have illustrated how the statistical methods have an improved performance in images which exhibit noise, notably in the work of Kundu and Mitra (1987), Hou and Wei (2002), and Williams et al. (2014).

Statistical based analysis techniques similar to those surveyed by Williams et al. (2014), can be effectively used for edge detection, particularly when applied to images corrupted with noise. de Souza (1983) first described the comprehensive analysis of five statistical parametric and non-parametric tests for the detection of edges, particularly those of rib structures in x-ray imagery. Comparisons were presented using the parametric students t test, fisher test, likelihood ratio test, the tau ($\tau$) test and also the non-parametric chi square ($\chi^2$) test. Although this work was limited to a single dimension of the x-ray image data (1-D), the tests and methods are easily extended into 2-D and 3-D configurations. Further work reconfirmed the approach of using non-parametric statistical tests for edges as appropriate, notably in 2-D images from Bovik and Munson (1985). Bovik and Munson (1985) presented the successful use of non-parametric ranking based statistical edge detectors using the Wilcoxon median rank tests and also a novel implementation of the parametric difference of boxes (DoB) test for detecting boundaries in speckle images. Subsequent work by Bovik et al. (1986) enhanced this two sample statistical edge detection technique, presenting the ratio of averages (RoA) detector to suppress image sensor based speckle noise.

Similar analysis of non-parametric statistical tests is further described by Beauchemin et al. (1998), namely the Wilcoxon Mann-Whitney test ($u$-test), which was found to be effective in negating the low signal to noise ratio of synthetic aperture radar (SAR) images. Beauchemin's work showed a robust response from the $u$-test, illustrating the advantages of using ranking statistics owing to their reliability in the presence of extreme observations (Beauchemin et al., 1998). However, the author did note that the use of statistical rank based tests for edge detection will show a loss in the output efficiency, because only the rank of the pixel distributions are used. Further applications in 2-D edge detection using region statistics is given in the work of Guest (1994), this work presents both the Student's $T$-test and the Fisher test for edge detection in 2-D



(a) 0°/180° (b) 22.5°/202.5° (c) 45°/225° (d) 67.5°/247.5°

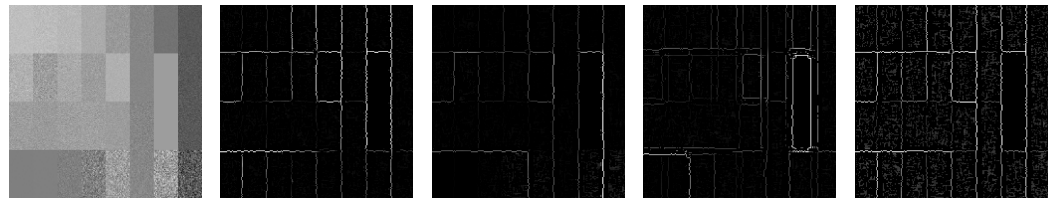
(e) 90°/270° (f) 112.5°/292.5° (g) 135°/315° (h) 157.5°/337.5°

Figure 2.16: Example of a 2-D dual region mask applied to the grid structure of an image in the different orientations. Region A is defined by the red pixels and region B is defined by the blue pixels. The shaded grey pixels constitute the dividing line of the mask

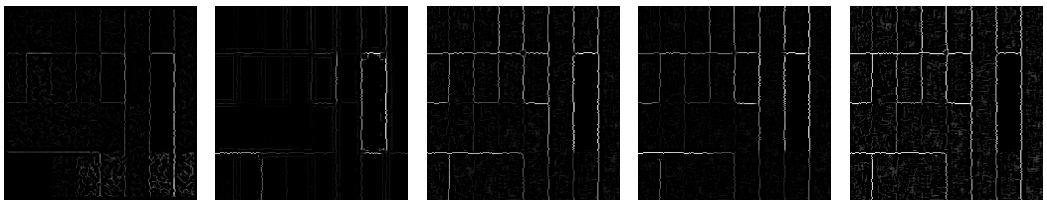
histological models. Fesharaki and Hellestrand (1994) used a similar technique, here implementing the Student's $t$- test only, to successfully detect edges in both noise-free and noise-corrupted images.

A comprehensive analysis by Williams et al. (2014) on the performance of edge detection methods in 2-D data, revealed that statistical oriented mask based methods for edge detection consistently outperform traditional gradient detection, namely where excessive texture is evident in the image or when the intensity profile of an edge is weaker, proving to be an optimal form of edge detection under both general and specific conditions (Williams et al., 2014). The method described by Williams (2008) and Williams et al. (2014) employs an oriented 2-D dual region statistical filter similar to the compass methods of Kirsch (1971) and of Nevatia and Babu (1980), however while those methods use convolution to resolve edges based upon determining the maximum intensity gradient in predetermined directions, statistical filters do not use convolution. Instead statistical filters use a neighbourhood window which is divided into two distinct opposing regions and applied in different orientations. Fig. 2.16 illustrates the various mask configurations for dividing a 2-D neighbourhood.

Pixel values from each region are collected to form samples to be evaluated by a two-sample statistical comparison test in order to apply some measure of dissimilarity between samples. The orientation of the mask which produces the largest statistical



(a) Test Image (b) $\chi^2$ test (c) Difference of Boxes (d) Fisher test (e) Kolmogorov-Smirnov test



(f) Canny (g) Likelihood test (h) RRO (i) Student's $t$-test (j) $u$-test

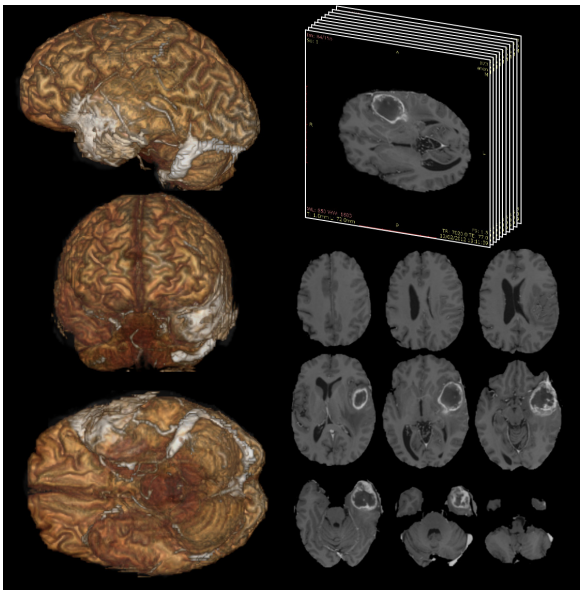
Figure 2.17: Comparison of 2-D statistical edge detection methods with the Canny edge detection algorithm. The Canny result reveals a number of missed edges in comparison to the non parametric statistical methods.

difference formulates the output measurement. The choice of statistical comparison test allows for the detection of edge features in addition to the intensity gradient, including higher order statistics which compare the sample distributions, such as variance and spatial properties such as texture. This provides greater flexibility to resolve different kinds of edge features (Williams et al., 2014).

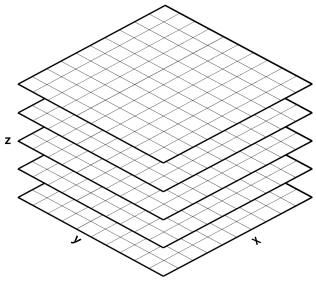
The outputs of various parametric and non-parametric 2-D statistical tests can be seen in Fig. 2.17, when visually inspected against the gradient methods presented in Fig. 2.11 improvements from some statistical methods are observed, notably greater noise suppression and fewer missed edge points.

## 2.7 3-D Digital Images

While 2-D images are the most common type of digital image, the prevalence of 3-D imaging modalities in the medical field is increasing as 3-D imaging technologies become more accessible. These include modalities such as Computed axial tomography (CAT), Single-photon emission computed tomography (SPECT) and Magnetic resonance imaging (MRI), all of which produce volumetric data with 3 spatial dimensions. MRI offers the ability to provide further image modalities which reveal different structural properties through manipulation of the pulse sequencing, the two most common modalities are T1-weighted and T2-weighted images. Additionally, a contrast agent is frequently utilised to visually improve the images by enhancing vascular and tissue phases. This makes MRI a very important tool for diagnosis and treatment planning for various



(a) 3-D Visualisation




(b) 3-D image Structure

Figure 2.18: 3-D Images can be displayed on a 2-D screen using software packages which apply, techniques such as pixel shading to render a 3-D image. The data that forms 3-D images resembles a lattice structure and is illustrated here.

tumour pathologies, and one of the most sought after imaging tools in medicine and is explored in further detail in Chapter 6.

While image displays are typically flat 2-D screens, 3-D visualisation software uses various algorithmic techniques such a light shading, alpha transparency maps and manipulation tools to allow for 3-D images to be visualised on a 2-D display and therefore utilised (Fig 2.18a).

Theoretically, while there is no strict constraint to the number of possible dimensions that can be used for a digital image, the trade off between utility and practicality offered by more dimensions typically limits the number of dimensions to 3.

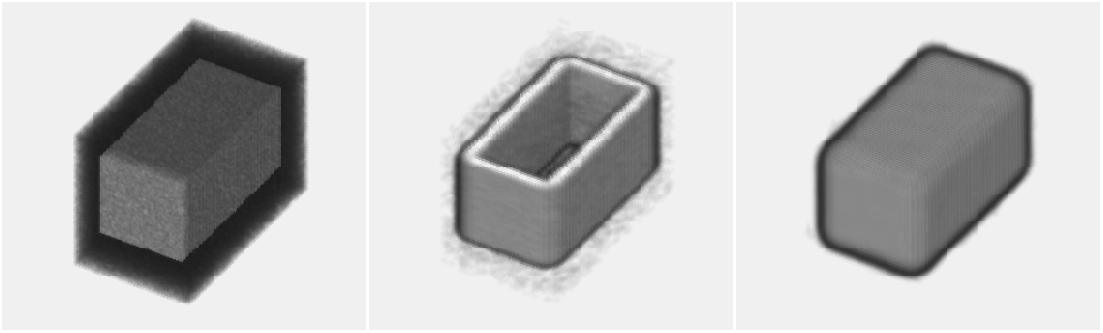
There are subtle differences between 2-D images and 3-D images. Instead of pixels positioned in a grid structure. Volumetric picture elements (Voxels) are positioned in a lattice structure that preserves the spatial relationship of the acquired measurements

or values (Fig 2.18b). Voxels are defined as triples of Cartesian coordinates ($x$, $y$, $z$) in a three-dimensional lattice structure (Morgenthaler and Rosenfeld, 1981). Similar to photons collected by a camera sensor to form a spatially representative 2-D digital image, voxel measurements can be acquired through a number of 3-D scanning techniques to form spatially representative 3-D images. However, not all scanning techniques provide isometric resolution in their lattice structure. Visualisation software usually compensates for this by making adjustments to the display parameters. However, for image processing applications that utilise spatial information, such as surface detection algorithms, the anisotropic resolution is often overlooked resulting in unreliable detection results (Brejl and Sonka, 2000).

## 2.8 Surface Filtering (Surface Detection)

A surface is a local spatial gradient in a sensory continuum, such as luminance, colour and texture in a three dimensional space. Surfaces define the interface between volumetric regions in 3-D image data, such as MRI, PET and SPECT. Like edges, surfaces also have a magnitude which describes the degree of change between regions. If 'Edges' are always a line or curve of single pixel width and can be of arbitrary length, a surface in image data can therefore be defined as a curved plane of single voxel thickness passing through a 3-D image which partitions two or more regions of voxels based on certain classification criteria (Liu, 1977; Zucker and Hummel, 1981). Surfaces are detected using surface detection algorithms, which are analogous to edge detection techniques, however they operate in a 3-D space and resolve local changes in properties in all three spatial dimensions. The result of surface detection is a surface map, which is a 3-D array describing the position and magnitude of detected surfaces contained within the image, and similar to edge detection further stages can be applied to classify surface points and non-surface points.

A common approach to detecting surface features is to treat the 3-D data as if it were a stack of 2-D image layers (Fig. 2.18b), processing each layer sequentially independent from the last with a 2-D edge detection algorithm, such as in the work of Kennedy et al. (1989), Tang et al. (2000), Braude (2005), Prakoonwit and Benjamin (2007), Lyra (2012), Prakoonwit and Benjamin (2012) and Simmons et al. (2013). However, this approach introduces some issues. The fundamental drawback of using a stacked 2-D approach, is that this does not exploit the information available in 3-D data sets to the extent 'true' 3-D processes are able to, due to the fact information across layers (along the $z$-dimension) is ignored. This can create circumstances where surfaces contained within an image are not located, often leading to unconnected surfaces. Fig.2.19 highlights this issue.

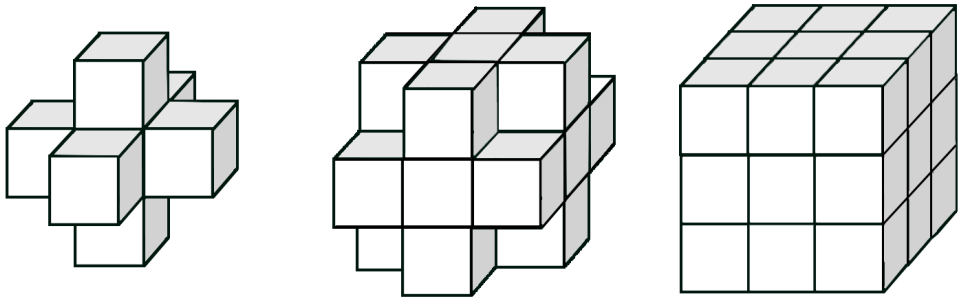


(a) 3-D interface (b) Common 3-D Approach (c) True 3-D Approach

Figure 2.19: Difference between the common approach of applying edge detection independently to each layer with that of a true 3-D approach. The former only resolves edges which exist in the plane of application, surfaces from higher planes are missed. This error is illustrated here with missing the top and bottom surfaces from the layered approach.

Processing the data in this configuration ignores the spatial relationship between values across image layers, thus by definition excludes interfaces between regions across image layers. For some tasks this may be adequate, however spatial information across the image layer can contain important structural information in MRI and and CT scans which could otherwise be missed. Bomans et al. (1990) showed how a 3-D implementation of the Marr-Hildreth edge operator can be used to resolve closed 3-D surface contours overcoming the incomplete surface issues which arise from using 2-D operators on 3-D data. While Monga et al. (1991) were investigating 3-D methods of recursive filters for edge detection on magnetic resonance images, 3 key observations were made when comparing a dedicated 3-D operator with a 2-D operator applied to image layers separately. Firstly the 3-D methods offered better immunity to noise when directly compared to a 2-D implementation of their recursive filters. Secondly, they noted that 3-D methods offered a better estimation of the gradient magnitude. Thirdly; computation of the 3-D edge gradient (A surface gradient) is achievable. These three factors combined to achieve better performance in 3-D.

The improved immunity to noise aspect from 3-D detection over 2-D was further exploited by Wang et al. (2009) which showed 3-D surface methods improve over 2-D methods by reducing the trade off which exists between the resolving power and the noise reduction aspects of the edge operator. It has been shown that a greater number of points in a neighbourhood can achieve greater noise suppression, but this comes at a cost of localisation accuracy (Papari and Petkov, 2011). For 2-D neighbourhoods, the number of pixels processed is proportional to the square of size of the neighbourhood mask, while for 3-D neighbourhoods, the number of voxels processed is proportional



(a) 6 Voxel Window (b) 18 Voxel Window (c) 26 Voxel Window

Figure 2.20: 3-D neighbourhood operator mask typically occur in 3 different formats. 26-connectivity is the most common surface operators, and in this configuration all neighbouring voxels are included in the mask.

to the size of the neighbourhood mask cubed. This results in more sample values processed in 3-D neighbourhoods of the same unit size as 2-D neighbourhoods. Thus, reducing the cost of the trade-off by offering improved noise suppression at greater accuracy. Results from Wang et al. (2009) show 3-D surface operators are superior to 2-D operators in terms of noise suppression and accuracy of surfaces. Therefore when 3-D data is available, a true 3-D approach which utilises a 3-D local neighbourhood should be used.

### 2.8.1 3-D Neighbourhood Operators for Surface Detection

3-D neighbourhood windows come in 3 spatial configurations of varying sizes, those with 6 voxel connectivity, 18 voxel connectivity and those with 26 voxel connectivity, these configurations are illustrated in Fig 2.20). By applying a 3-D neighbourhood operator, 3-D spatial properties of an image can be processed, allowing for a rich set of features to be extracted, whereby uniform areas of grey levels, regions of texture and surfaces can be distinguished.

As previously described, neighbourhood operators allow for filtering processes such as smoothing and feature extraction. For surface detection, the process is in principle the same as edge detection is for 2-D. The most common approaches to surface detection therefore have been adaptations of optimal 2-D edge detection methods into 3-D formats. Generally for surface detection the 26 pixel connectivity variety is used, for which the voxel index locations are presented in Eq. 2.4.

$$h = \begin{cases} h_{i-1,j-1,k-1} & h_{\ i,j-1,k-1} & h_{i+1,j-1,k-1} \\ h_{i-1,\ j,k-1} & h_{\ i,\ j,k-1} & h_{i+1,\ j,k-1} \\ h_{i-1,j+1,k-1} & h_{\ i,j+1,k-1} & h_{i+1,j+1,k-1} \end{cases}$$
$$\begin{vmatrix} h_{i-1,j-1,\ k} & h_{\ i,j-1,\ k} & h_{i+1,j-1,\ k} \\ h_{i-1,\ j,\ k} & h_{\ i,\ j,\ k} & h_{i+1,\ j,\ k} \\ h_{i-1,j+1,\ k} & h_{\ i,j+1,\ k} & h_{i+1,j+1,\ k} \end{vmatrix}$$
$$\left.\begin{matrix} h_{i-1,j-1,k+1} & h_{\ i,j-1,k+1} & h_{i+1,j-1,k+1} \\ h_{i-1,\ j,k+1} & h_{\ i,\ j,k+1} & h_{i+1,\ j,k+1} \\ h_{i-1,j+1,k+1} & h_{\ i,j+1,k+1} & h_{i+1,j+1,k+1} \end{matrix}\right\} \tag{2.10}$$

When processing 3-D images, $h$ is the index locations which define an 26-connectivity mask

## 2.9 3-D Classical Approaches

As with edges, surfaces are often interfaces between regions of high and low intensity. Liu (1977) was the first to exploit this commonality by extending the Roberts operator to work in resolving surfaces in three dimensions. Liu's work defined the surface to be the location of the gradient magnitude in multiple dimensions. In continuation of this work, Zucker and Hummel (1981) generalised a 2-D Hueckel operator for 3-D, offering improvements over Liu's generalised Roberts operator in areas of image noise and texture. In this process they consider a 3-D 'edge' (surface) as a plane passing through the centre of a unit volume, approximating the optimal operator with three basis functions:

$$\psi_1(x,y,z) = \frac{x}{\sqrt{x^2+y^2+z^2}} \tag{2.11}$$
$$\psi_2(x,y,z) = \frac{y}{\sqrt{x^2+y^2+z^2}} \tag{2.12}$$
$$\psi_3(x,y,z) = \frac{z}{\sqrt{x^2+y^2+z^2}} \tag{2.13}$$

$\psi_1$,$\psi_2$ and $\psi_3$ are the basis functions which correspond to the 3 Dimensions $X, Y$ and $Z$

$$h = \left\{ \begin{array}{ccc|ccc|ccc} \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} & 0 & 0 & 0 & -\frac{\sqrt{3}}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & 1 & \frac{\sqrt{2}}{2} & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & -1 & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} & 0 & 0 & 0 & -\frac{\sqrt{3}}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{3}}{3} \end{array} \right\} \tag{2.14}$$

3-D Zucker-Hummel 3×3×3 Filter kernel

Further to this work, Morgenthaler and Rosenfeld (1981) developed a multidimensional edge detection operator through a process of "hypersurface" fitting. Leading to the

development of a generalised Prewitt detector for three dimensional data sets. Although the surfaces resolved using this manner were characteristically more 'blurred' than the 2-D counterpart, successful detection of the surfaces provided justification for using generalized Prewitt operators to detect (hyper) surfaces of discontinuity in 3-D volumes, and further justifying 3-D neighbourhood operator design for surface detection. The 3-D Prewitt filter is presented in Eq. 2.15.

$$G_X = \left\{ \begin{array}{ccc|ccc|ccc} -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \right\}$$

$$G_Y = \left\{ \begin{array}{ccc|ccc|ccc} -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \end{array} \right\}$$

$$G_Z = \left\{ \begin{array}{ccc|ccc|ccc} -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \end{array} \right\}$$

3-D Prewitt 3×3×3 Filter kernel

$$\nabla I_{(x,y,z)} = \sqrt{{G_x}^2 + {G_y}^2 + {G_z}^2} \tag{2.15}$$

Later, Bhattacharya and Wild (1996) generalised the 2-D Sobel operator for use with 3-D modalities, Bhattecharya and Wild noted the versatility of a straightforward, 3-D operator for producing 3-D greyscale surface maps. Using volumetric datasets of biological data, they showed that the surface detection results were satisfactory, providing results comparable to the 2-D Sobel operator. A Generalised 3-D Sobel filter is presented in Eq. 2.16,

### 2.9.1 Limitations

Since 3-D gradient operators function using the same fundamental process as 2-D edge detection methods, the effect of non-homogeneous regions such as noise (Fig 2.10), still produces erroneous surface points. This is because both 2-D and 3-D methods alike interpret noise as local abrupt changes in image intensity and as such are consequently interpreted as surface points by the 3-D gradient methods. However 3-D techniques do offer some advantages in aspects of resolving power and noise suppression when compared to 2-D gradient methods of edge detection (Monga and Deriche, 1989).

$$G_X = \left\{\begin{array}{ccc|ccc|ccc} -2 & -4 & -2 & -4 & -8 & -4 & -2 & -4 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 4 & 2 & 4 & 8 & 4 & 2 & 4 & 2 \end{array}\right\}$$

$$G_Y = \left\{\begin{array}{ccc|ccc|ccc} -2 & 0 & 2 & -4 & 0 & 4 & -2 & 0 & 2 \\ -4 & 0 & 4 & -8 & 0 & 8 & -4 & 0 & 4 \\ -2 & 0 & 2 & -4 & 0 & 4 & -2 & 0 & 2 \end{array}\right\}$$

$$G_Z = \left\{\begin{array}{ccc|ccc|ccc} -2 & -4 & -2 & 0 & 0 & 0 & 2 & 4 & 2 \\ -4 & -8 & -4 & 0 & 0 & 0 & 4 & 8 & 4 \\ -2 & -4 & -2 & 0 & 0 & 0 & 2 & 4 & 2 \end{array}\right\}$$

$$\nabla I_{(x,y,z)} = \sqrt{{G_x}^2 + {G_y}^2 + {G_z}^2} \tag{2.16}$$

Generalised 3-D Sobel 3×3×3 Filter kernel

Additionally the performance of gradient methods in texture rich, or noisy environments continues to be an issue, even with the added noise suppression and resolving power that arises from 3-D operators. Fig. 2.21 illustrate 'over-detection' artefacts and missed surfaces from the outputs of 3-D gradient based surface detection methods.

### 2.9.2 3-D Canny Method

The Canny edge detector, which is widely regarded as being the optimal 2-D edge detection filter has also been adapted for multiple-dimensional data sets. Notably in the work of Monga et al. (1991) the use of a kernel operator which corresponds the first derivative of a 3-D Gaussian filter was proposed (2.17).

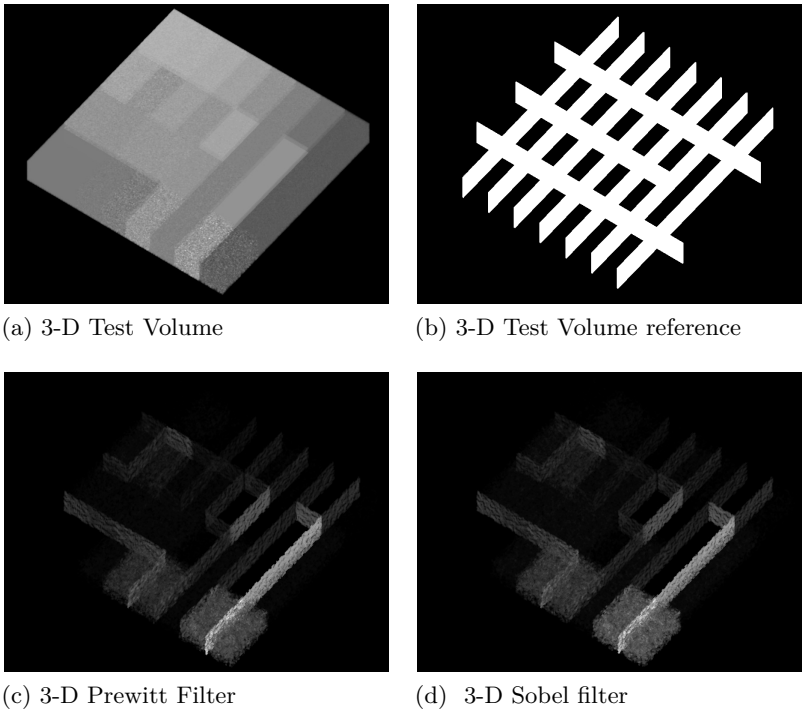
$$G_x(x,y,z) = \frac{-x}{\sigma^2} \times e^{\frac{-\left(x^2+y^2+z^2\right)}{2\sigma^2}} \tag{2.17}$$

$$G_y(x,y,z) = \frac{-y}{\sigma^2} \times e^{\frac{-\left(x^2+y^2+z^2\right)}{2\sigma^2}} \tag{2.18}$$

$$G_z(x,y,z) = \frac{-z}{\sigma^2} \times e^{\frac{-\left(x^2+y^2+z^2\right)}{2\sigma^2}} \tag{2.19}$$

$G_x$,$G_y$ and $G_z$ define the masks which correspond to the 3 Dimensions $X, Y$ and $Z$, while $\sigma$ refers to the standard deviation of the Gaussian distribution

Monga et al. (1991) concentrated on the application of recursive filtering for multi-dimensional edge detection and exposed the advantages that multi dimensional operators have over the more traditional two dimensional approaches when used with 3-D image volumes. The key benefits being improved noise immunity, improved estimation of the



(a) 3-D Test Volume (b) 3-D Test Volume reference

(c) 3-D Prewitt Filter (d) 3-D Sobel filter

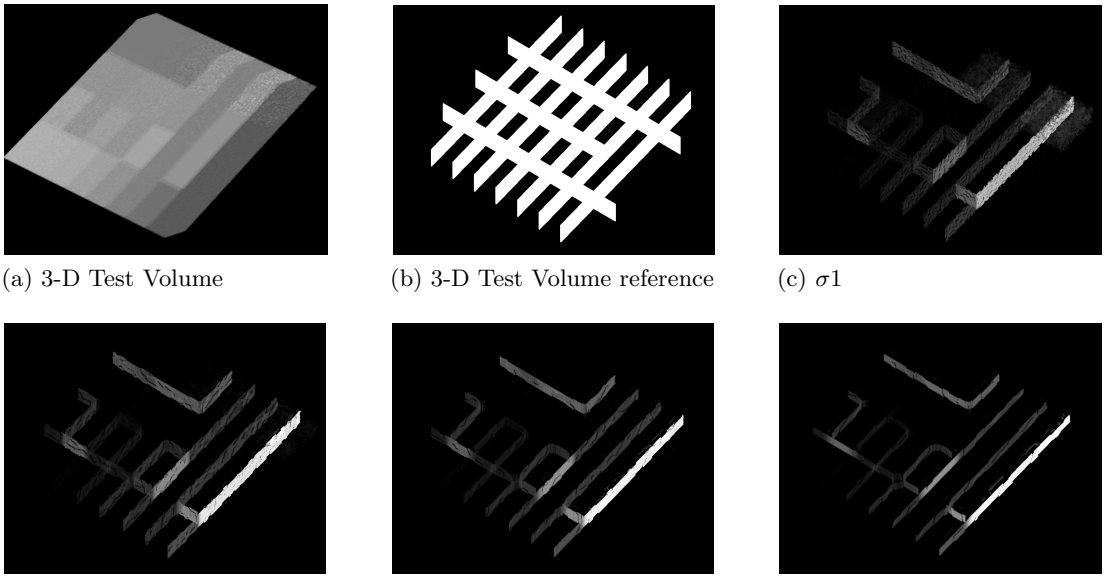
Figure 2.21: Example of intensity gradient based 3-D Prewitt and Sobel sobel methods on complex data. High intensity variance and different region profiles cause traditional methods to produce poor results, with missed surfaces and noisy responses.

edge (or gradient) magnitude and the computation of the accurate 3-D gradient, which itself improves the noise immunity by allowing smoothing in the $x$, $y$ and $z$ directions.

However, there is a compromise between noise immunity and edge location accuracy regarding the size of the convolution kernel applied. This effect is illustrated in Fig 2.22, as the size is increased there is greater noise reduction, but some surface detail is lost, and the position of the surface becomes less accurate. However the trade off in 3-D has been shown to be not as costly as in 2-D (Monga et al., 1991). More recently Bähnisch et al. (2009) further improved the adapted 3-D Canny method, omitting the traditional non-maximum suppression and hysteresis stages and replacing it with 3-D morphological operations for surface thinning and segmentation achieving better computational efficiency.

#### 2.9.2.1 Limitations

Image noise and texture still negatively affects the performance of the 3-D Canny method this can be seen in Fig 2.22c, here the regions with a strong variance component producing noise. Where regions are separated by texture differences, the surfaces are missed. In addition, the preprocessing Gaussian smoothing stage of the technique can remove small surface details, this affects performance on surface corner and junction



(a) 3-D Test Volume (b) 3-D Test Volume reference (c) $\sigma 1$

(d) $\sigma 2$ (e) $\sigma 3$ (f) $\sigma 4$

Figure 2.22: 3-D Canny Edge detection with different standard deviations of Gaussian smoothing. Increased $\sigma$ increases the suppression of noise and therefore improves signal to noise ratio, however this is at the cost of detection accuracy.

regions. When compared to other 3-D gradient methods there is more variability with parameter selection, increasing the degrees of freedom, which can make optimal detection more complex of a task. Typically, prior knowledge of image content is required for the user specified parameters in order to optimise performance.

### 2.9.3 3-D Statistical Approaches

Liu and Chen (2006) notes that medical imagery is inherently noisy and often in error, which is why edge detection based on texture characteristics can be comparatively effective on medical imagery when compared with intensity gradient based techniques (Thangam et al., 2009). Interfaces governed by differences in texture are not limited to images constrained to two dimensions, Maani et al. (2014) observed that 3-D texture features are also a useful tool to define region boundaries in 3-D modalities such as MRI. However, for 3-D data, the specific case of surface detection using statistical measures has previously not been extensively explored. While the use of 3-D statistical filters in the literature is sparse, a notable exception is the work of Petrou et al. (2006), here spherical dual region statistical filters were utilised to resolve non-linear *'invisible'* surfaces in MRI data. The statistical measures applied were limited to basic parametric statistics, assessing only differences in mean, variance and skewness. However it was shown that these invisible surfaces exist due to the differences in the skewness of the distributions of the different regions, and a 3-D dual region statistical filter can be

applied to resolve these boundaries. The authors recommended further work in this area, however there remains a lack of further work relating to 3-D statistical filters available in the literature.

Since 3-D techniques function using the same mathematical principles as their 2-D counterparts, they also share similar performance characteristics. This appears to be the case for gradient methods, Canny method, Steerable filters and other non-linear methods. Typically with added benefit of improved signal to noise ratio, and greater noise suppression in the 3-D variant. It is therefore expected that 3-D statistical measures, which use the same mathematical principles as 2-D statistical edge detection, would also provide the same performance benefits as 2-D statistical methods. Notably on textured interfaces and on surfaces in the presence of noisy regions, with the additional benefits of 3-D processing. The following chapters look to explore the gap in the literature and attempt to assess the characteristics of 3-D statistical filters and objectively quantify how they perform when compared against the existing 3-D technique benchmarks.

## 2.10 Summary

Edge and surface features are used in a number of different computer vision techniques such as segmentation, object recognition, image coding and robot vision. In the literature edge detection is underpinned by a comprehensive body of research. Traditional methods of edge detection use neighbourhood convolution operators which approximate the intensity gradient function, which is a measurement of the change in image brightness. A sharp shift in image brightness is an excellent visual cue for an edge, and therefore operators which measure the brightness shift are also very effective at detecting these types of edges. However, there are a large number of studies which indicate some the issues with this approach, these include problems such as the detection of edges described by other feature differences, such as texture, and the negative impact the presence of noise imposes on the result. The literature shows that statistical methods of edge detection provide an excellent alternative when these sub-optimal conditions exist.

The growing availability of 3-D imaging modalities provides an incentive to investigate surface detection methods in search of improved performance, but the availability of surface detection research is limited in comparison. However, the majority of traditional surface detection approaches are adaptations of their 2-D edge detection counterparts. Improvements to edge detection performance has increased incrementally, and many of the 3-D surface detection techniques mimic their 2-D counterparts in these incremental developments.

While there are extensive evaluations of edge detection methods, including 2-D statistical approaches, evaluation of 3-D surface detection approaches is much less established. Statistical edge detection has been shown to offer improved performance over traditional edge detection methods, but there are very few examples of statistical based approaches to surface detection in the literature and evaluation of statistical surface detection methods against traditional methods is absent.

The following chapters present a novel model for statistical surface detection, and the most comprehensive analysis to date of statistical surface detection is undertaken where these methods are evaluated along side the 3-D Canny and Steerable filter methods.

# Chapter 3

# A Statistical Surface Detection Model

The novel statistical surface detection methods detailed in this chapter were published in IEEE Signal Processing Letters (Smith and Williams, 2015) and presented at the Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM '15).

## 3.1 Introduction

This chapter presents two models for surface detection using statistical features, in order to determine where interfaces between different image regions exist. Different image regions, by definition contain at least one statistical property which differentiates a region from another, such as region brightness (intensity), texture features, or distribution properties. By using a statistical based surface detection method, properties of a region in addition to brightness can be exploited in order to locate the interface between regions. This chapter also discusses the various parameters available to statistical surface detection, such as the types of two-sample statistical tests which can be incorporated and the scale at which the tests are applied, indicating the likely effect the parameters have on the characteristics of the result. Finally this chapter details the post processing methods of non-maximum suppression and hysteresis thresholding, and their usage with the statistical surface detection methods.

## 3.2 Statistical Surface Detection

Dual region statistical filters were initially described by Guest (1994) and Bowring et al. (2004), and were comprehensively evaluated for 2-D images by Williams et al. (2014) and these approaches to statistical detection have been shown to offer improvements over derivative-based methods when texture is present within the image. In a general

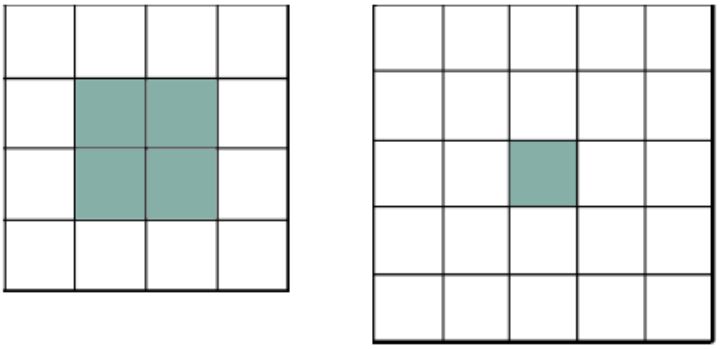


Figure 3.1: A 2-D example of an even integer sized neighbourhood, resulting in no individual central pixel, whereas with an odd integer size, a central pixel is available.

case, the neighbourhood mask can simply be defined as the local region surrounding a pixel/voxel in an image. This local neighbourhood can be of any size subject to the constraints of the image resolution, however odd, integer $x$,$y$,$z$ components to the neighbourhood size ensures that a central pixel or voxel can be established in the centre of the neighbourhood (Fig 3.1).

Similar to statistical edge detection, statistical surface detection also employs an orientated non-linear dual region filter. In order to compute the statistical difference using three dimensional data, introduced in this work is a 3-D neighbourhood mask (Fig. 3.2). Here the mask is extended from the work of Guest (1994) and Bowring et al. (2004) through the $z$-dimension to produce a mask capable of computing differences between regions A and B across three dimensions. Instead of a central dividing line commonplace in 2-D, the 3-D mask has a central dividing plane, and two volumetric sample regions (also shaded red and blue). In the 3-D configuration, there are two axis about which the mask can be orientated. Unlike the 2-D method, the scale of the mask does not determine the number of intervals which the mask is rotated through. Instead this mask is rotated through 13 uniquely orientated positions, derived from a 26-connectivity neighbourhood window (Fig.3.3).

If a single orientation is defined as the direction from the central neighbourhood voxel to a neighbouring voxel, due to the rotational symmetry, half of the positions would provide the same magnitude output, only in the opposite direction, and thus would be redundant. Therefore, these 3-D methods can make use of 13 different orientations which can resolve the 26 discrete surface directions at any given position within the image set. The positions are composed of orientations which form a dividing plane orthogonal to two voxels opposing the origin (centre mask voxel) of the mask, examples of $5\times5\times5$ implementations of the 3-D neighbourhood mask are shown in Figure 3.4.

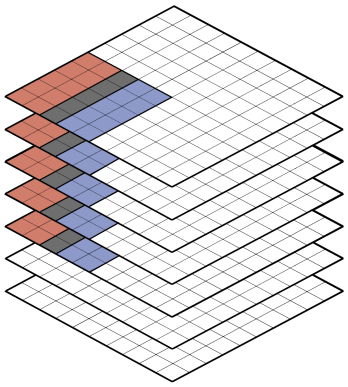


Figure 3.2: A 5×5×5 3-D dual region mask applied to the grid structure of an image. Region A is defined by the red voxels and region B is defined by the blue voxels. The shaded grey voxels constitute the dividing plane of the mask

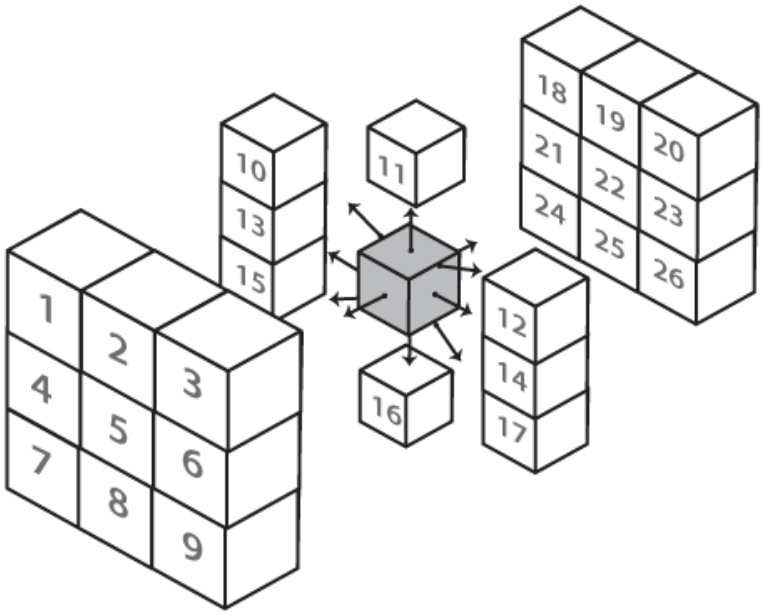



Figure 3.3: A 3-D 3×3×3 local neighbourhood, illustrating the 26 directions from the central voxel which can exist in the neighbourhood.

### 3.2.1 Surface Magnitude

The magnitude of a surface is calculated by comparing the two neighbourhood regions using a statistical method, here a two sample Kolmogorov-Smirnov (KS) statistical test has been applied to the neighbourhood regions in different positions and orientations in the image, but this could be interchanged with a different two-sample test. In this example the KS Statistic defines the magnitude of the surface map in those positions. The greater the statistical difference in region distribution properties, the higher the

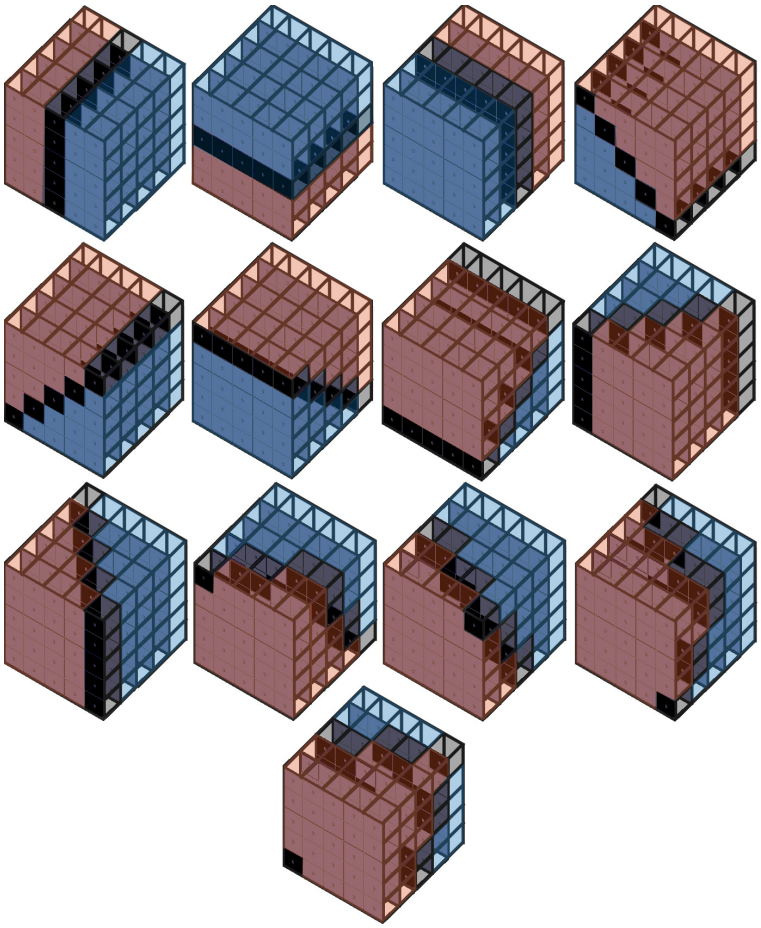
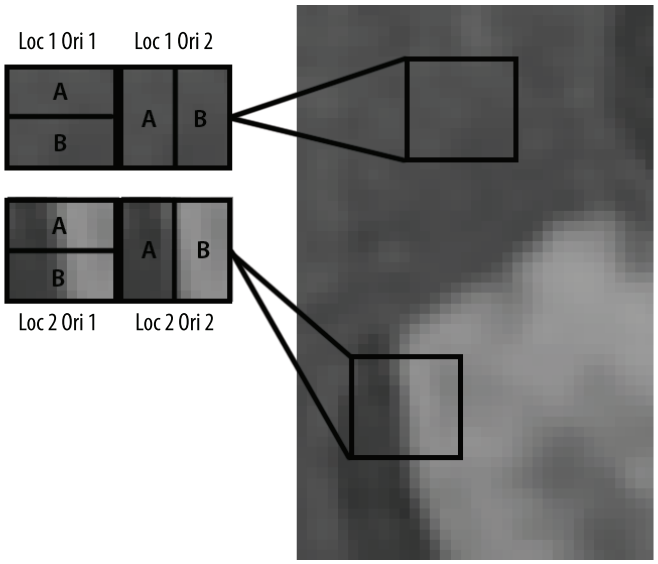
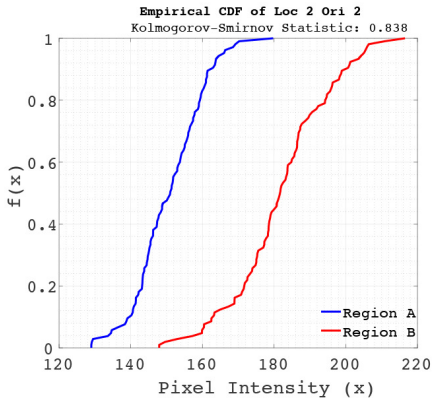


Figure 3.4: The 13 unique mask orientations defined from a 26-neighbourhood window applied to a 5×5×5 neighbourhood. Sample regions are colour coded as red and blue, while the central dividing plane is shaded black. Each sample region (noted by red and blue shading) are always of equal size

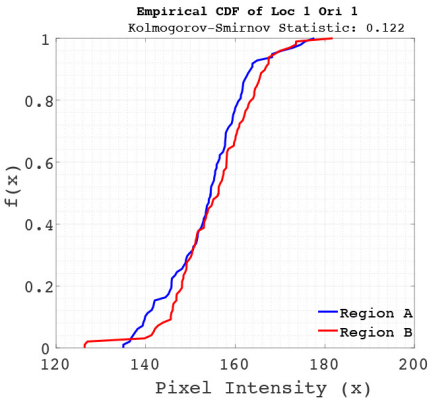
likelihood that the neighbourhood mask is situated on an surface and as such the magnitude of the statistical difference measure in such a position would be expected to be high (Fig. 3.5b). However, when the neighbourhood mask is not aligned on an surface, it is expected that the difference in the distributions would be small since both mask regions would lie in the same region (Fig. 3.5b,c,d).



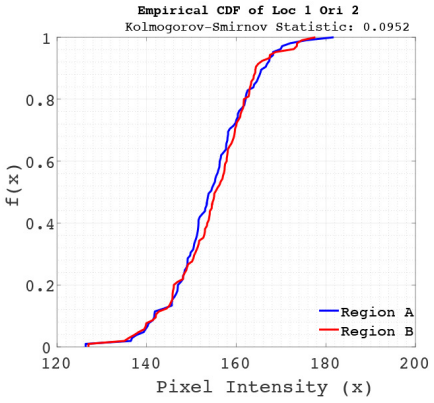

(a) Mask positioned in interface and non interface locations



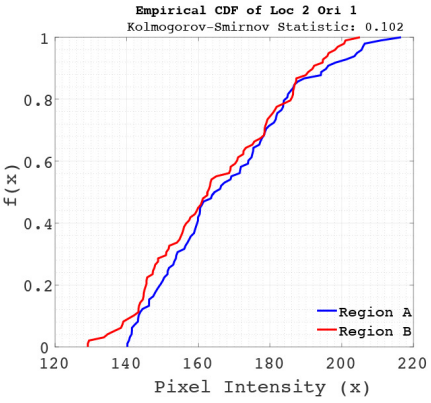

(b) Location 2 ori 2




(c) Location 1 ori 1




(d) Location 1 ori 2



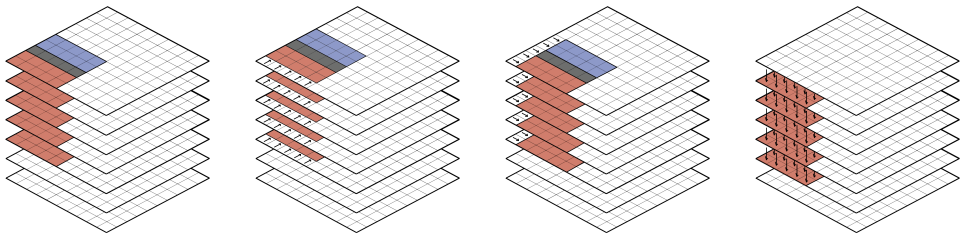

(e) Location 2 ori 1

Figure 3.5: Example of statistical detection principle using Kolmogorov-Smirnov test. The neighbourhood mask situated in both an interface and non-interface location, and resulting difference in the empirical distribution functions. A high $KS$ statistic indicates a strong candidate for an edge or surface. Notice position (b) where the mask is both correctly located at an interface in the correct orientation. The resulting output statistic is high magnitude, while other positions (c,d,e) produce a low magnitude result.

This contrast provides a way to discriminate between non-surface and surface pixels. When a dual region mask is situated on an interface between different image regions, and the central dividing line is aligned along the surface direction, the expected output is a high magnitude result. A dual region mask can be situated on an interface but not oriented along the surface direction. Under these conditions the expected result is of a low magnitude (Fig. 3.5e). This means that in order to resolve surfaces through 360° the dual region mask needs to be able to shift its orientation. In the 2-D case for statistical edge detection, this is achieved in a similar fashion to the Kirsch (1971) compass operator, whereby the filter orientation is shifted through different positions, such as in the case of the statistical methods evaluated by Williams et al. (2014). The mask rotates through a fixed number of discrete positions dependent upon the size of the mask, larger masks allow for a greater number of potential orientations, re-orientating a dual region 2-D mask can be seen in the earlier example in Figure 2.16. For 3-D data, the mask requires further orientations to satisfy the greater number of directions that can exist in a three-dimensional space, there are two different approaches to this which are further detailed in Section 3.3.

### 3.2.2 Global Processing of the Image

In order to produce a surface map of the entire image volume, every 'legal' voxel within the image is processed. In order for a voxel to be processed, the origin of the mask is situated in each voxel location, and rotated through all mask orientations in order to compute the statistical surface gradient of the image. The first location will be the first position in which the dual region neighbourhood mask can be fit into the image, this will be dependent upon the scale of the dual region mask, an example of a $5\times5\times5$ dual region mask initialised in the first legal position can be viewed in Figure 3.6a). After the mask has been orientated through all the designated orientations, and the value of the statistical difference has been evaluated, the dual region mask traverses the $x$,$y$ and $z$ dimensions of the image voxel by voxel (Fig. 3.6b-d) until all available positions have been evaluated. Not all locations in the image volume can be accurately evaluated, this is a typical problem associated with all neighbourhood operators. Since the origin of the mask is the central location of the neighbourhood, and needs to be situated on the voxel to evaluate that position, there are a number of voxel positions along the outer perimeter and layers where if the origin is situated on a voxel to be evaluated, the local neighbourhood region is incomplete, or padded with arbitrary values. Since these positions cannot be accurately processed, they are excluded, as a consequence the output surface map matrix will be smaller than the original image matrix in each dimension proportional to the scale of the mask.



(a) Initialise dual region mask position.

(b) Shift along $x$ dimension.

(c) Shift along $y$ dimension.

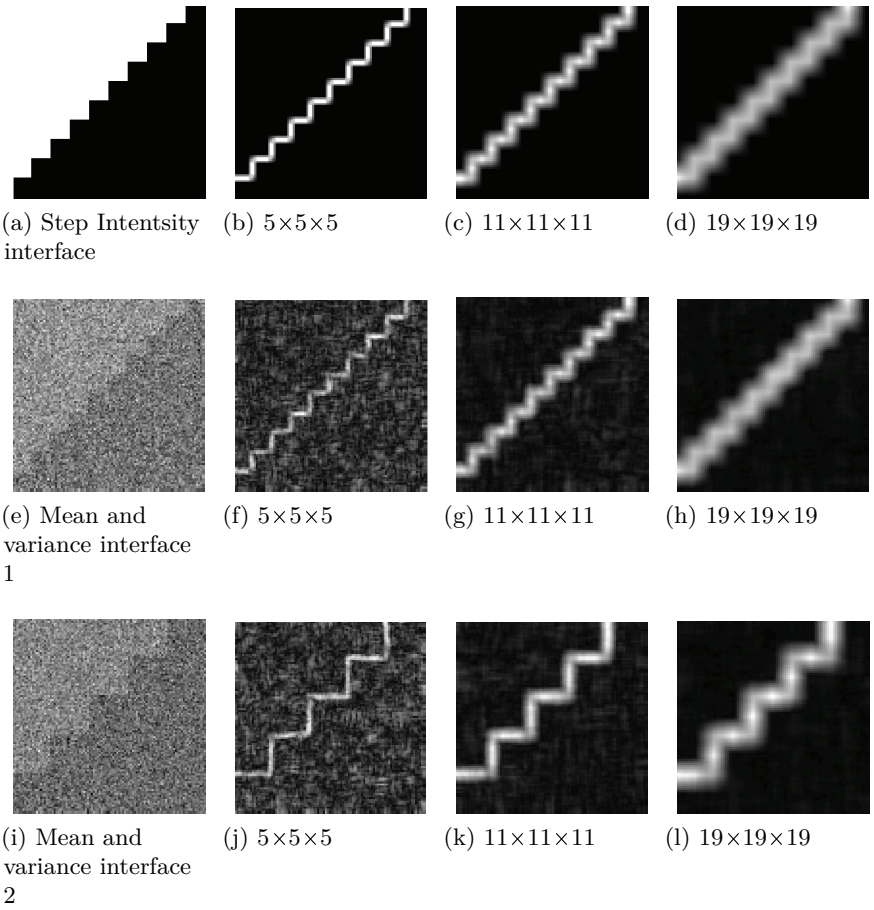
(d) Shift along $z$ dimension.

Figure 3.6: Statistical surface detection filtering procedure illustrating how the 3-D mask traverses the image volume.

### 3.2.3 Mask Scale

Scale refers to the parameter which controls the spatial resolution specification of the neighbourhood mask. As the scale parameter is increased, the dimensions of the neighbourhood mask are increased accordingly. For 2-D detection, a common approach to achieving greater noise suppression is to increase the scale of the mask, thus increasing the number of elements processed by the filter. However, increasing the scale of the mask generally comes at a cost in terms of accuracy, or localisation error (Ziou and Mohr, 1992). The advantage in terms of noise suppression achieved by surface detection filters over their edge detection counterparts are that surface detection filters resolve interfaces across 3-dimensions, thus require $n^3$ scale neighbourhood masks, while 2-D edge detection filter are of order $n^2$. Using three dimensions over two allows for a smaller scale parameter to be applied, while maintaining resolving power equivalent to larger 2-D mask sizes (Wang et al., 2009). For fixed resolution imagery, a larger mask is required in two dimensions to have the same statistical significance as a mask making use of three dimensions. For example a mask size of $5 \times 5 \times 5$ processes 100 voxels from a neighbourhood of 125, all of which are located no more than two voxels distance away from the centre position being computed. For a two dimensional neighbourhood mask to have similar statistical significance, it would need to be $11 \times 11$ pixels in size. This covers a significantly larger image region (Up to 5 pixels away from the pixel being processed). 3-D methods therefore allow for the use of a smaller scale parameter for a given amount of noise suppression which minimises the localisation error but also provides improved detection, this observation is a significant advantage and was first published by Monga et al. (1991). However, since 3-D methods process more values at an equivalent neighbourhood scale, 3-D methods are more computationally expensive than their 2-D edge detection counterparts, particularly in the case of rank based non-parametric tests which require sorting of the data. The scale parameter is a considerably powerful parameter (Bergholm, 1987), which needs to be taken into consideration

depending on the target edge and surfaces in the image. Fig 3.7 demonstrates the



(a) Step Intentsity interface (b) 5×5×5 (c) 11×11×11 (d) 19×19×19

(e) Mean and variance interface 1 (f) 5×5×5 (g) 11×11×11 (h) 19×19×19

(i) Mean and variance interface 2 (j) 5×5×5 (k) 11×11×11 (l) 19×19×19

Figure 3.7: 2-D layered view of surface detection at mask scales 5×5×5,11×11×11 and 19×19×19 on 3 test images. Surface detection at a fine resolution (small neighbourhood mask) typically yields noise while detection at a coarse resolution (large neighbourhood mask) while suppressing noise, typically distorts surface contours

trade off between noise suppression and detection accuracy with respect to the size of the mask. As the mask scale is increased from 5×5×5 through to 19×19×19, the level of noise suppression is increased, signified by fewer false positive (FP) edge points. However, the true position of the boundary becomes more unclear, signified by the increase in the degree of diffusion of the surface interface. One consequence of a larger scale parameter is therefore the loss of detail finer than that of the scale of the mask.

Scale is a common parameter for many edge and surface detection methods, similarly gradient based methods typically also control the scale by adjusting the size of the filter kernel as is the case with Prewitt and Sobel operators in the same fashion as the statistical surface model presented here. However, it should be noted that there alternate techniques for addressing the scale parameter trade off between noise suppression and

accuracy. For instance the Canny operator uses Gaussian filtering as a pre-processing step, here the scale parameter is determined by changing the standard deviation ($\sigma$) of the Gaussian filter (Canny, 1986). A large standard deviation will apply a greater degree of image smoothing, this suppresses weaker interfaces, image noise, and resolves surfaces which occur over a larger region (Ziou and Mohr, 1992). A small standard deviation however results in a much more sensitive detector, able to resolve more image detail, but with less noise suppression.

Work by Williams (2008) and Williams et al. (2014) has shown the impact scale has on the performance of 2-D statistical edge detection methods. It was shown that non parametric tests achieve optimum results when using large mask size, relative to the scale which achieves optimal results for the parametric tests, for example the 2-D $\chi^2$ and Kolmogorov-Smirnov tests performed optimally at scales above $21 \times 21$. At scales smaller than this, the sensitivity of the filters resulted in noisy results. This is because non parametric tests that rank the sample data, effectively normalise the output locally at the scale of the neighbourhood mask. As a result this makes non-parametric tests sensitive. Increasing the scale of the neighbourhood mask can offset the sensitivity to achieve better results.

Applying a larger scale parameter potentially removes finer details in 3-D data, as well as increasing the uncertainty in surface location. But by using a 3-D mask, more sample values are processed when compared with 2-D and this suppresses the noise. This has the noticeable effect in the quality of the 3-D results in comparison to 2-D, this effect is explored later in the Chapter 7 discussion.

### 3.2.4 Statistical Test Selection

In the work of Williams (2008) and Williams et al. (2014) different statistical measures for dissimilarity were evaluated in the context of 2-D statistical edge detection, but not 3-D. The statistical measures employed in that work demonstrated various advantages and disadvantages across different interface types, all methods were shown to be effective under particular conditions, offering improvements over 2-D gradient methods, however no single statistical measure produced the optimal result under all conditions. In order to see if the advantages offered by the statistical edge detection are transferable to 3-D surface detection. The same statistical measures were employed as those evaluated by Williams et al. (2014) in order to observe whether the characteristics of statistical edge detection are transferable to surface detection. The addition of the Robust Rank Order statistical test was also included, this method was not evaluated by Williams et al. (2014). However, it is included here as it was shown in the work of Lim (2006) to

be advantageous for edge detection over both the Students $t$-test and Mann-Whitney Wilcoxon $u$-test in noisy imagery. The list of statistical measures assessed in this work consists of 4 parametric (mean and variance based) and 4 non-parametric (distribution or ranking based) measures, but the list is not exhaustive and scope remains for other statistical comparison measures to be explored.

#### 3.2.4.1 Difference of Boxes Test ($DoB$)

An adaptation of the Difference of Boxes test initially defined by Rosenfeld and Kak (1982) and Bovik et al. (1986) was included in this work. The difference of boxes method is typically a difference measure between the application of two different scaled mean or averaging filters, which offers an approximation of a Laplacian of Gaussian filter. Essentially a convolution based operator used to compute the second spatial derivative of an image to resolve edges. Here the adaptation was further simplified by Williams (2008) to compute a comparison of the mean between the two mask regions. The $DoB$ statistic is given using the equation 3.1.

$$D = |\bar{x}_A - \bar{x}_B| \tag{3.1}$$

Where $\bar{x}_A$ is the mean value of region A and $\bar{x}_B$ is the mean value of region B.

As this method measures the difference in the average intensity between regions, it is particularly suited for resolving step changes in image intensity, and is therefore comparable to standard gradient based operators. In 2-D edge detection, the Difference of Boxes method provided good detection of interfaces which features strong mean and variance components and provided a response comparable to the derivative based Canny Edge detector. However the $DoB$ method was not ideal where a significant difference in mean values between regions is not present, in line with the performance associated with derivative based detectors. In 2-D this method did produce results with closed contours, and at larger neighbourhood mask scales, offered good noise suppression qualities. However its application into 3-D remains unexplored.

#### 3.2.4.2 Student's $t$-test ($t$-test)

The next mean based parametric test used is the Students' $t$-test which compares the location parameter of two independent data samples. The $t$-test is used to test the null hypothesis that two independent sample distributions have the same or similar mean values. Typically the $t$-test is applied to sample distributions with the same population size and similar variance. However, it has been shown by Bovik et al. (1986) to be effective even when the sample variance greatly differs. The test statistic for a two-sample Student's $t$-test for two equally sized regions, and is given here as Eq. 3.2.

$$T = |\bar{x}_A - \bar{x}_B| \sqrt{\frac{N-1}{s_A + s_B}} \tag{3.2}$$

Where $\bar{x}_A$ and $\bar{x}_B$ are the mean of mask region $A$ and $B$, respectively. $N$ is the number of pixels in one region, with $s_A$ and $s_B$ being the variances of the two mask regions $A$ and $B$.

In the work of Williams et al. (2014) 2-D edge detection methods employing the Student's $t$-test successfully resolved interfaces where a prominent difference in the mean intensity of regions existed. Furthermore, when there was a strong difference in variance between regions in addition to a mean difference, the Student's $t$-test method also proved to be successful. Similarly to derivative based methods, where interfaces are not associated with a strong change in the mean intensity of a region, the Student's $t$-test was shown to not be the optimal method. Williams et al. (2014) also showed that the Student's $t$-test produced edges with good connectivity, and offered good suppression of noise at larger mask scales.

#### 3.2.4.3 Fisher Test ($F$-test)

If the means of the two mask regions does not differ significantly Williams et al. (2014) demonstrated that a more suitable approach would be to analyse the variance of the two regions. The Fisher test (F-Test) is the first variance based test to be used for this situation. The F-Test is applied to check the hypothesis that two datasets will have the same or a similar variance. The two-sample F-test is defined as the ratio of the variances of region $A$ and $B$ using equation 3.3:

$$F = \max\left(\frac{s_A}{s_B}, \frac{s_B}{s_A}\right) \tag{3.3}$$

Where $s_A$ and $s_B$ are the variances of region $A$ and $B$, respectively.

The work of Williams et al. (2014) indicates that the F-test provided successful detection of variance interfaces, and interfaces with strong mean and variance difference components. However, when a strong variance component was not present, the F-test is less successful. One of the issues found with the F-test is that correctly detected edges can be shifted, introducing localisation error. This shift was shown to proportional to the scale of the neighbourhood region of the F test and can therefore be suppressed with post processing. While these advantages and problems have been observed in 2-D edge detection, it remains unknown whether or not these characteristics are transferable for surface detection, and therefore analysis is required.

#### 3.2.4.4 Log-Likelihood Ratio Test ($L$)

A further test of variance is the Likelihood Ratio test. The test used here is based on the work of de Souza (1983) and is defined in equation 3.4. In 2-D edge detection it was shown that the maximum likelihood method offered successful detection of both mean interfaces and combined mean and variance interfaces, however it failed with on interfaces with no significant mean difference component, and was outperformed by the Fisher test, as well as non parametric tests such as the Kolmogorov-Smirnov and $\chi^2$ tests in 2-D. One notable feature was that when the method was applied to 2-D histological tissue images (which are typically highly textured) the Likelihood test offered a strong suppression of unwanted texture even with small mask scales while preserving connected contours of required segments. Analysis of the Likelihood-ratio test for 3-D data sets is also previously unexplored and requiring analysis.

$$L = -N \cdot \log_e \lambda \tag{3.4}$$

Where $N$ is the number of pixel points covered by both regions of the mask used in the test, and $\lambda$ is defined in Eq 3.5

$$\lambda = \frac{4v_A^2 v_B^2}{v_{A\cup B}^4} \tag{3.5}$$

Where $N$ is the number of pixel points covered by both regions of the mask used in the test. $v_A^2$ is the variance of region A, $v_B^2$ is the variance of region B and are defined in Eq 3.6

With $v_A^2$ defined as: Set $B$ of the mask and the combined sets of $v_B^2$ and $v_{A\cup B}^4$ are then

$$v_A^2 = \sum_{x\in A} x^2 - \frac{1}{|A|}\left(\sum_{x\in A} x\right)^2 \tag{3.6}$$

where $x$ is the pixel intensity value

defined analogously.

#### 3.2.4.5 Kolmogorov-Smirnov Test ($KS$)

The two sample Kolmogorov-Smirnov (or $KS$-test) provides a statistic which quantifies a distance between the empirical distribution functions of two samples. The Empirical Cumulative Distribution Function (ECDF) of sorted ascending data points is defined in equation 3.7. As this statistical method compares differences in the ECDF of the neighbourhood region data, and not specifically the distribution values, the KS score will always produce a score within the range of 0 - 1, where 0 indicates no difference

between mask regions and 1, indicates a maximum difference between regions. This is irrespective of the position of the neighbourhood within the image, therefore this has the effect of normalising the output measurement within its local neighbourhood. The effect of this on the on the result globally is that local differences between regions are emphasised, this makes the filter very sensitive to small differences in image features and the result of that is that "weak" edges are locally normalised and therefore resolved to a similar level to what would be typically considered "strong" interfaces. In the work of Williams et al. (2014) this was confirmed by results which show the $KS$-test is very sensitive at smaller mask sizes to small changes in image features, while this enables weak surfaces to be resolved, this also made the method sensitive to image noise. However it was shown that the $KS$-test offers improved edge detection at larger mask scales when compared against traditional gradient methods of edge detection. The use of a 3-D neighbourhood, as opposed to a 2-D neighbourhood, for a given mask scale results in 3-D sample regions which contain more sample values than a 2-D region of a given scale, it is therefore expected that a 3-D adaptation using the $KS$-test should improve upon its 2-D counterpart.

$$F_A(i) = \frac{n(i)}{N} \tag{3.7}$$

Where $F_A$ is the ECDF for data set A, $n(i)$ is the number of data points less than the $i^{th}$ data point in ranked set A and $N$ is the number of overall points contained in data set A. The two-sample KS test checks for the maximum difference between the ECDFs for both data sets. From this it returns the value of $D$ given in Eq 3.8

$$D = \max_{i \in \{1,\ldots,N\}} |F_A(i) - F_B(i)| \tag{3.8}$$

Where $F_A$ is the ECDF for data set A, and $F_B$ is the ECDF for data set B.

#### 3.2.4.6 The $\chi^2$-test ($\chi^2$)

The $\chi^2$-test ($\chi^2$) is a rank based test which checks for the independence of the two sorted data sets. This calculation is a comparison measure that takes the relative difference in points at the same rank position for two binned data sets. Here the two binned datasets are simply taken as region $A$ and region $B$ therefore giving two-sample Chi Square test in equation 3.9. Two variants of the $\chi^2$-test were available for this study. The standard local-$\chi^2$-test ($l\chi^2$-test) and the alternate global-$\chi^2$-test($g\chi^2$-test). The fundamental difference between these two configurations is the method used to group the data into bins. The first method, the local-$\chi^2$ test, is the same process as described by (Williams et al., 2014). Here pixel or voxel values are binned in local

neighbourhoods at the neighbourhood mask level. For 2-D edge detection, the local $\chi^2$-test variant is better suited to larger mask sizes due to the local normalisation effects implicit in the fact that the data is organised in isolation. These effects were shown to be similar to those exhibited by the KS and $u$-test methods, but this has yet to be established for surface detection. The local $\chi^2$-test filter response exhibits a high degree of sensitivity and is therefore likely to be better suited to targeting weaker surface interfaces. The alternate method introduced here is the global-$\chi^2$ method, which is the configuration used in Chapters 5 and 6. In this configuration, the $\chi^2$ method does not process the neighbourhood mask in isolation. Instead the values are processed and organised in the context of the entire image volume, this is accomplished by initially assigning all the voxels to predefined bins. In this configuration the method is computationally less expensive, less sensitive to weak boundaries, and is hypothesised to be better suited to resolving strong boundary interfaces, while suppressing image noise. Unlike other traditional methods, this configuration tends to create a response with a consistent strong magnitude which is not dispersed over a large area. Thereby minimising fragmentation and missing boundaries. Evaluation of the local $\chi^2$-test method undertaken by Williams et al indicates that the $\chi^2$-test offers successful detection of mean, variance and combined mean and variance interfaces. In addition the $\chi^2$-test offers improved edge detection at larger mask scales than traditional methods of edge detection. However, similar to the KS-test, it can fail in excessively noisy environments with small mask scales due to over-detection. The evaluation of the $\chi^2$-test in 3-D data is novel and all results were obtained using 4 data bins.

$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i} \tag{3.9}$$

Here $R_i$ is the number of values in *bin* $i$ of region $A$, and $S_i$ is the number of values in *bin* $i$ of region $B$.

#### 3.2.4.7 Wilcoxon Mann-Whitney ($u$-test)

The non-parametric Wilcoxon Mann–Whitney test is commonly used for detecting differences in central tendency between two samples Again this rank based test checks the hypothesis that two data sets under evaluation are taken from the same distribution. Here the Mann-Whitney test statistical value U is defined in equation 3.10. 2-D evaluation of the method indicates that the $u$-test also offers a local normalisation of the edge detection results due to the ranking nature of the test, and therefore provides a strong response for all edge types within the image (i.e. weak and strong edges). However, in 2-D imagery, the $u$-test failed with zero mean interfaces and was

outperformed by the Fisher, KS and $\chi^2$ tests. In the work of Williams et al. (2014), the $u$-test was found to be the most successful statistical test for mean based interfaces. It returned strongly connected contours and where mean interfaces are assumed it outperformed all statistical tests and traditional derivative filters, but analysis for surface detection is absent, and it is not yet known if the performance in 3-D data is as effective.

$$U = \min(R_A, R_B) \tag{3.10}$$

$$R_A = \sum_{x \in A} \left( \sum_{y \in B; y < x} 1 \right) \tag{3.11}$$

Where $R_A$, and $R_B$ correspond to a ranking score calculated for both data sets (or mask regions) A and B. The equation 3.11 refers to the rank calculation ($R_A$) of data set A with $R_B$ being defined analogously. Each rank ($R_A$ and $R_B$) is calculated by summing a rank score for the observed differences of current data points between both sets.

#### 3.2.4.8 Robust Rank Order Test (*RRO*)

The Robust Rank order (*RRO*) statistic was developed by Fligner and Policello (1981) as an alternative to the Wilcoxon Mann–Whitney $u$-test, in order to create a statistic which is more applicable across different distributions and sample sizes. Lim (2006) describes a statistical edge detection method which incorporates the Robust Rank-order statistic of Fligner and Policello (1981), in order to resolve edges in noisy images. In the work of Lim (2006) it was shown that the *RRO* edge method outperformed the $u$-test, Student's $t$-test method and the Canny Edge detection method on both synthetic and real images when in the presence of impulse and Gaussian noise, and was therefore included for 3-D evaluation. Given that medical images are often noisy (Liu and Chen, 2006) and the *RRO* measure performs well in noisy imagery, the *RRO* test statistic was included for this reason. While this method was analysed by Lim (2006) for noisy 2-D imagery, it was not subject to the comprehensive testing of the other 2-D methods explored in the work of Williams et al. (2014). Evaluation of this statistical approach in 3-D surface detection is completely novel.

## 3.3 3-D Statistical Model Variants

There are two variations in which the statistical masks can be applied. One approach finds the maximum statistical difference through a range of different neighbourhood

orientations, the other approach applies the statistical measure in 3 orthogonal planes, and the L2-Norm is calculated to find the statistical difference gradient. The first variant is named the '*Maximum Response*' method (MR-method) and is a direct extension of 2-D statistical filtering evaluated by Williams et al. (2014) into a 3-D framework. The second approach is novel in context of statistical surface detection, but shares similarities with traditional gradient approaches in that it measures a statistical feature gradient instead of an intensity gradient.

### 3.3.1 Maximum Response

The MR approach rotates the neighbourhood mask through 13 unique orientations, Fig 3.4 illustrates how the neighbourhood can be positioned, such that the neighbourhood is oriented to find differences in the direction of all the voxels in a 26-connectivity neighbourhood. From these positions the statistical measure is applied and the differences are recorded. The orientation which produces the maximum dissimilarity between regions provides the statistical difference magnitude for that location in the image volume. The stages of the approach are presented in Fig. 3.8. In addition, the mask orientation for that location can also be stored, this allows for the surface direction to be calculated. This data can be used to assist further post processing techniques where the direction of a surface needs to be known a priori such as with non-maximum suppression.

This method is not particularly efficient, since calculations need to be made in orientations in which the statistical difference is small in order to find the position which is the most statistically different. In addition, as the neighbourhood mask scale is increased, the number of possible orientations can be further increased to account for the smaller increments in rotation available with a larger mask size. However, in this work a constrained case is used to limit the number of orientations to 13 irrespective of mask size for practicality. In addition, when data is anisotropic in its resolution, weighting the mask positions to compensate adds further complexity to the process. This MR approach is included here as a basis for comparison to 2-D statistical edge detection, which also utilises a maximum dissimilarity approach.

### 3.3.2 Vector Magnitude

To avoid the aforementioned problems of the MR method, a different approach to the detection method is required. The second variant presented here, named the '*Vector Magnitude*' method (VM-method) is a novel approach to both statistical edge detection and statistical surface detection. In this configuration the 3-D algorithm applies the dual region filter in 3 orthogonal orientations (Fig. 3.9) in order to form a vector which contains 3 directional magnitude components ( $T_x$,$T_y$,$T_z$ ) of dissimilarity (Eq. 3.12).

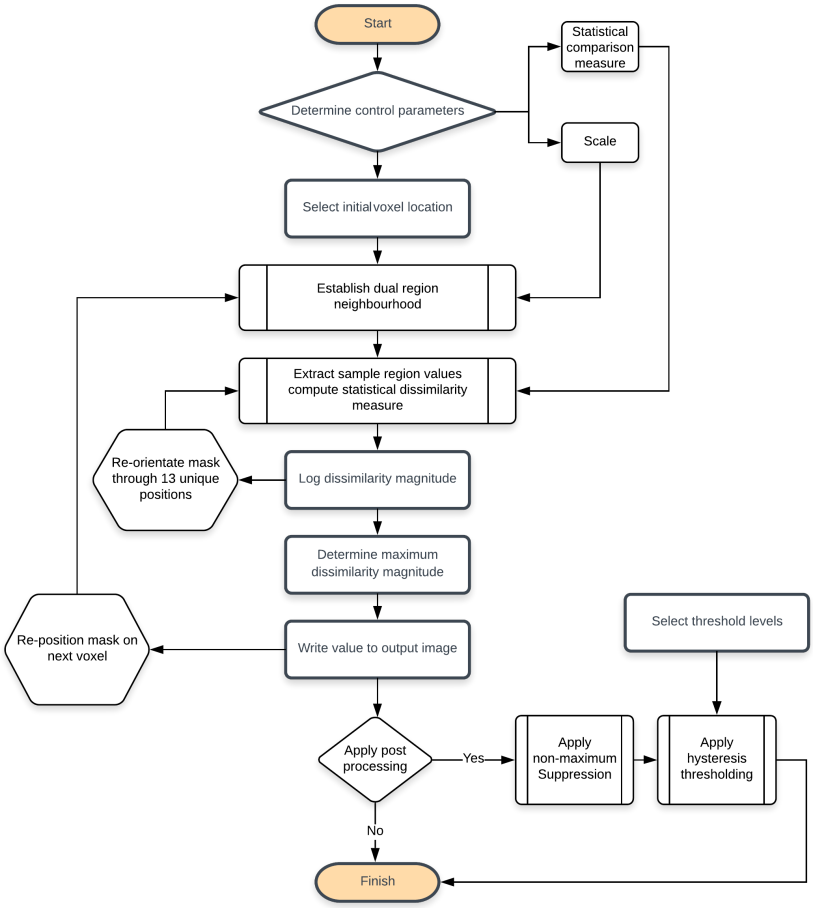


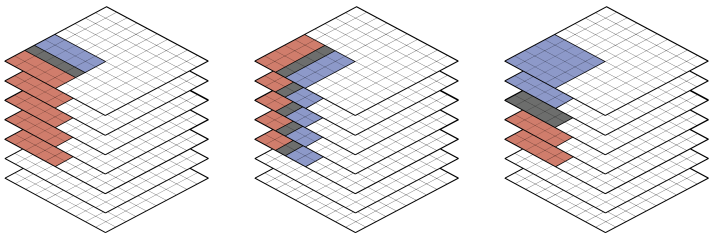

Figure 3.8: Workflow of the Maximum Response procedure.

These statistical dissimilarity measures conform along the $x$,$y$ and $z$ planes of the image. From this vector, a Euclidean Norm (L2-Norm) can be applied in order to compute the statistical feature gradient (SFG or $\|x\|$) using Eq. 3.13. The procedure for the VM method can be observed in Fig. 3.10.

$$V =_a T_x\hat{i} +_b T_y\hat{j} +_a T_z\hat{k} \tag{3.12}$$

$$\|x\| = \sqrt{({}_aT_x)^2 + ({}_bT_y)^2 + ({}_cT_z)^2} \tag{3.13}$$

The primary advantage of this method is that it is more efficient for a number of reasons. Firstly it is computationally less costly due to the fact that the statistical difference calculation is only required for 3 mask positions, in contrast to a minimum of 13 positions for the MR approach. The MR approach also retains only a single difference measure from the 13 orientations which are calculated, and therefore the



(a) $x$ (b) $y$ (c) $z$

Figure 3.9: The 3 mask configurations for the vector magnitude method. Each mask enables a directional statistical differential to be determined, which are then combined using a L2-norm to compute the output.

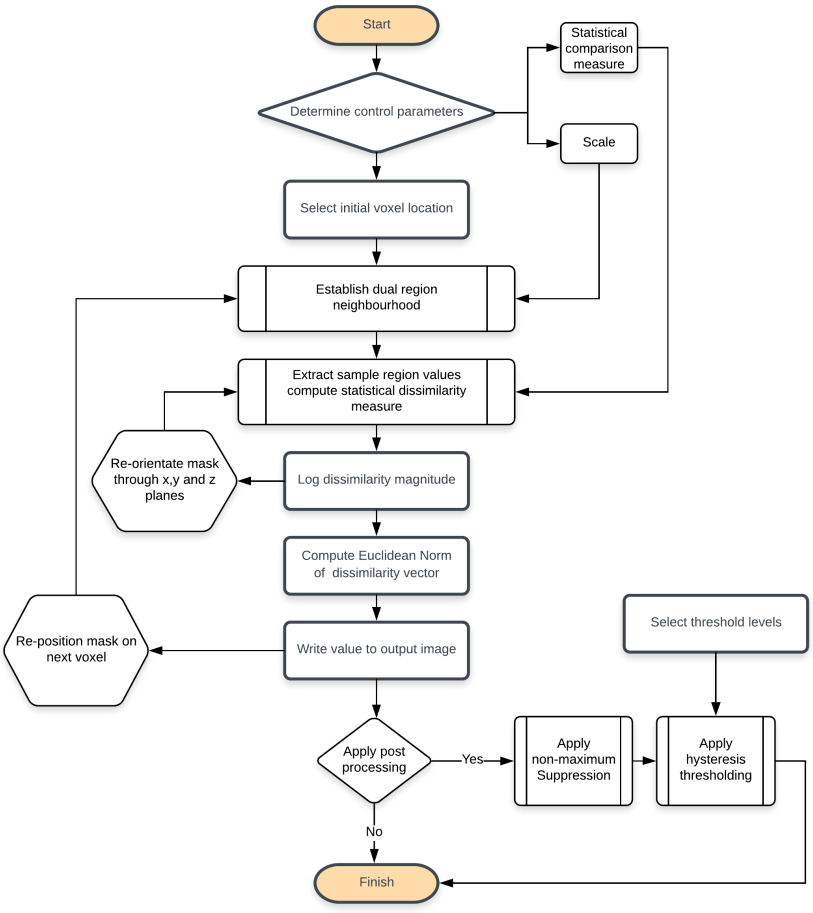



Figure 3.10: Workflow of the Vector Magnitude procedure.

redundancy of the MR approach is much greater. The VM approach is scale invariant, as the mask size is increased, further mask orientations are not required, as accurate measurement of the direction of the gradient can still be calculated from just 3 mask positions, whereas the MR approach requires further mask orientations to maintain the same single voxel resolution in calculation of the directional component.

### 3.3.3 Comparison of Variants

To illustrate the difference in computational efficiency between the two approaches, the computation time using different statistical measures is presented in Table 3.1. Even when using a constrained MR method, limited to 13 orientations for each neighbourhood scale, the computation time for the VM approach is significantly faster by at least a factor of 3, and for some tests by as much as an order of magnitude.

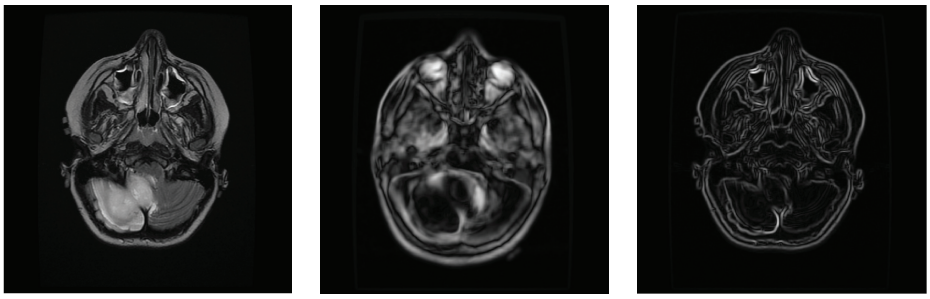
| Statistical method | MR time (s) | | | | VM time (s) | | | |
|---|---|---|---|---|---|---|---|---|
| | 5×5×5 | 7×7×7 | 9×9×9 | 11×11×11 | 5×5×5 | 7×7×7 | 9×9×9 | 11×11×11 |
| $DoB$ | 0.07 | 0.10 | 0.13 | 0.21 | 0.01 | 0.02 | 0.03 | 0.04 |
| $t$-test | 0.10 | 0.15 | 0.24 | 0.37 | 0.02 | 0.04 | 0.05 | 0.09 |
| $F$-test | 0.11 | 0.17 | 0.25 | 0.38 | 0.02 | 0.04 | 0.06 | 0.08 |
| $L$ | 0.14 | 0.25 | 0.39 | 0.63 | 0.03 | 0.06 | 0.09 | 0.15 |
| $\chi^2$t | 6.03 | 13.74 | 25.58 | 38.65 | 1.35 | 3.25 | 5.85 | 8.92 |
| $KS$ | 9.49 | 24.97 | 46.73 | 74.80 | 3.35 | 8.96 | 17.92 | 29.58 |
| $U$-test | 9.49 | 24.96 | 46.72 | 74.80 | 2.16 | 5.43 | 10.35 | 16.72 |
| $RRO$ | 56.19 | 267.65 | 919.03 | 2443.08 | 9.49 | 44.29 | 148.82 | 398.05 |

Table 3.1: Computation time of different statistical methods, with the different model variants. Average time was calculated using 10 synthetically created image volumes of normally distributed random noise (entropy = 4.16) of size 120×120×25. Calculations were performed using an Intel(R)Core(TM) i7-10750 2.59 GHz processor with 16GB of RAM available, using MATLAB v2020a

#### 3.3.3.1 Anisotropic Distortion of the Surface Map

Another issue which the VM approach can help alleviate, is the distortion which occurs when the resolution of data is not isotropic. Often with 3-D medical data, such as MR images, physicians view 2-D axial layers of the scan, rather than 3-D representations, as a result of this the MRI machine is often calibrated such that the $x$ and $y$-dimension resolution is greater than that of the resolution in the z dimension for improved 2-D resolution and shorter scan times(Brejl and Sonka, 2000; Du et al., 2020). This means that each voxel within the scan will occupy more space in the $z$-plane than the $x$ and $y$ planes. This means that surfaces orientated with a strong $z$-component dissimilarity will appear more abrupt, thus produce stronger interfaces. Thus 2 issues can arise, ambiguity in the accuracy of a surfaces' location, and a bias toward producing stronger interfaces with a large $z$-component. The effects of anisotropic image resolution on surface detection can be pronounced, Fig. $3.11_b$ demonstrates how z-plane surfaces are emphasised when the resolution in the $x$ and $y$ dimensions exceed the $z$ dimension by a ratio of 8:1.

The L2-Norm calculation of the Vector Magnitude method allows each individual directional component to be weighted using a resolution coefficient $(a, b, c)$ to account for anisotropic resolution EQ. 3.12. While this does not solve the issue of localisation, the weighting helps mitigate the biases associated with anisotropic resolution, enabling surfaces to be resolved without the strong $z$-component surfaces emphasised to a greater degree. This mitigation can be observed in the example case presented in Fig. 3.11c.



(a) Anisotropic MRI Volume

(b) No dimensional weighting

(c) With dimensional weighting

Figure 3.11: Surface detection with and without dimensional weighting applied to MRI volume with anisotropic resolution ($x, y$ - $z$ 8:1). The distortion occurs as strong responses for surfaces with a high magnitude $z$ component

This approach of weighting the directions is novel in both statistical edge detection and statistical surface detection. Brejl and Sonka (2000) applied a similar modification to the Canny surface detector, in their analysis of detectors in anisotropic data it was shown to significantly improve over the standard Canny surface method.

#### 3.3.3.2 Parametric test efficiency

Dual region statistical measures are fundamentally more computationally expensive when compared to derivative based neighbourhood operators. However, parametric tests can be applied in an efficient fashion also using convolution. However, non-parametric tests cannot utilise convolution in the calculation process due to the ranking nature of the techniques which are reliant on the use of sorting algorithms. In addition, as the mask size increases, the cost of using a sorting algorithm to rank the data also increases. Thus non-parametric techniques are more computationally costly than parametric tests as represented in Table 3.1.

To utilise convolution for fast efficient processing of parametric tests, consider the following vector magnitude example using a difference of boxes test:

1. Select a scale parameter and generate 6 directional convolution kernels as defined in Fig 3.9. Eg. Mask scale of 5.

$$h_x = \left\{ \begin{array}{ccccc|ccccc|ccccc|ccccc|ccccc} 1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0\\ 1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0\\ 1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0\\ 1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0\\ 1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0 \end{array} \right\}$$

$$h_{x'} = \left\{ \begin{array}{ccccc|ccccc|ccccc|ccccc|ccccc} 0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1\\ 0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1\\ 0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1\\ 0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1\\ 0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1&0&0&0&1&1 \end{array} \right\}$$

$$h_y = \left\{ \begin{array}{ccccc|ccccc|ccccc|ccccc|ccccc} 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \end{array} \right\}$$

$$h_{y'} = \left\{ \begin{array}{ccccc|ccccc|ccccc|ccccc|ccccc} 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \end{array} \right\}$$

$$h_z = \left\{ \begin{array}{ccccc|ccccc|ccccc|ccccc|ccccc} 1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \end{array} \right\}$$

$$h_{z'} = \left\{ \begin{array}{ccccc|ccccc|ccccc|ccccc|ccccc} 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1 \end{array} \right\}$$

2. Scale the kernel coefficients such that kernels become mask region averaging filters, where ms = mask scale.

$$ms' = (ms^3 - ms^2)^{-1}$$

$$H = h \times ms'$$

3. Convolve the image ($I$) with each directional pair of filter kernels to create 6 output images. For example the $x$ component is as follows:

$$G_x = I * H_x$$

$$G_{x'} = I * H_{x'}$$

4. Since the difference of boxes is the difference between the two averages, determine the absolute difference to establish the vector component for each direction

$$G_X = |G_x - G_{x'}|$$

$$G_Y = |G_y - G_{y'}|$$

$$G_Z = |G_z - G_{z'}|$$

5. Finally compute the vector norm of the 3 directional component images.

$$G = \sqrt{{G_X}^2 + {G_Y}^2 + {G_Z}^2}$$

This type of procedure can also be applied to the other parametric test methods (Difference of boxes, Fisher test, Student's $t$-test, Likelihood) resulting in efficient application of these statistical measures. However, the non parametric methods cannot be employed using this technique, thus are more computationally expensive.

## 3.4 Post Processing Stages

### 3.4.1 3-D Non-Maximum Suppression (NMS)

The surface map is considered the explicit separation measurement of region profiles performed by the surface detection algorithm. It is the direct output of the dual region statistical filtering. A surface map will highlight areas where interfaces between differing region profiles exist. These interfaces can be instantaneous step changes between profiles, or a more spread out gradual change. However, depending on the scale of the neighbourhood mask, a surface interface may be resolved over several voxel locations as the mask traverses the image. This will manifest itself in the surface map as a diffused boundary rather than a single voxel thickness interface. This diffusion or smoothing scales as a function of neighbourhood mask size, this effect is illustrated in Fig 3.7. The larger the mask, the greater the uncertainty of the location of a surface. A typical requirement for many computer vision tasks is to precisely measure the location of a surface to single voxel accuracy. Thereby there is a need to determine the actual position of the interface. In 2-D imagery, a method known as non maximum suppression, a process to thin edges in an edge map to single pixel width was first achieved by Canny (1986) for his stand out edge detection technique. Figure 3.12 shows the location of an edge in the edge map. The single pixel interface location of the edge is not precisely clear, as the magnitude of the edge is spread out along the gradient direction. After the application of the NMS algorithm, all points adjacent to the edge direction which are not the maximum intensity value are suppressed and given the value 0. Thereby retaining the surface where the gradient is at its steepest and removing it elsewhere.

While this process is primarily a 2-D technique, in this work the algorithm was modified to work in a 3-D environment for surfaces (Fig 3.12c). Instead of a diffused edge being thinned to a single pixel line. 3-D NMS converts a diffused surface to a single voxel thickness line.

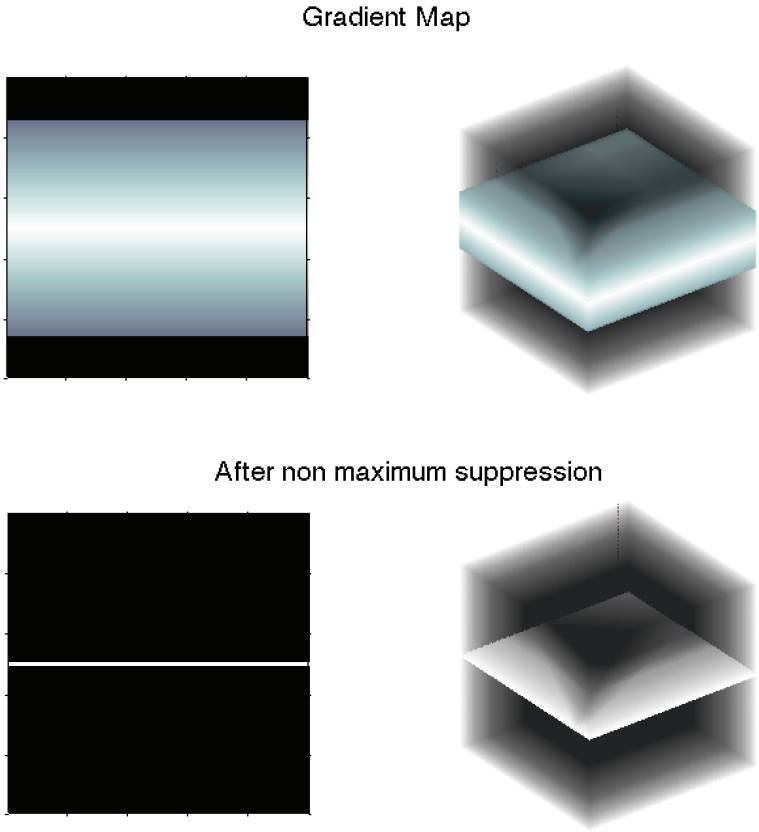



Figure 3.12: Non maximum suppression applied to an edge and surface map. Edge and surface points which are not at the gradient peak are suppressed to zero. Resulting in an edge and surface result that is unit pixel and voxel thickness respectively at the position where the gradient magnitude is at its maximum

Non Maximum Suppression (NMS) is an intermediate step used in many computer vision algorithms. Notably the Canny (1986) edge detection method. In terms of an edge detection application, this method is applied post edge filter to the edge map image. In reality an interface has no thickness at it is the location of where two or more regions meet. However, an edge map does not directly show where the edge exists, but rather a map of the rate of change of intensity, or texture, or region properties. The effect of applying the NMS algorithm to the edge map is to thin edges to single pixel thickness. This then allows for direct comparison between a ground truth solution image, and the result image.

In this work a 3-D non Maximum Suppression algorithm is presented which converts surfaces in a surface map to single voxel thickness by utilising two distinct 3-D operations.

1. Orientation filtering
2. Suppression of non maxima points

#### 3.4.1.1 Orientation filtering

In order to suppress surface voxels which are not the local maxima, the direction of the statistical gradient first needs to be obtained. The maximum response method will determine the gradient direction to be perpendicular to the mask orientation which produced the maximum statistical difference. Because of the discrete arrangement of voxels and limited orientations in the constrained MR approach, the gradient direction will always lie in one of 13 specific directions. For the MR method this can be obtained from the mask orientation resulting in the maximum statistical difference.

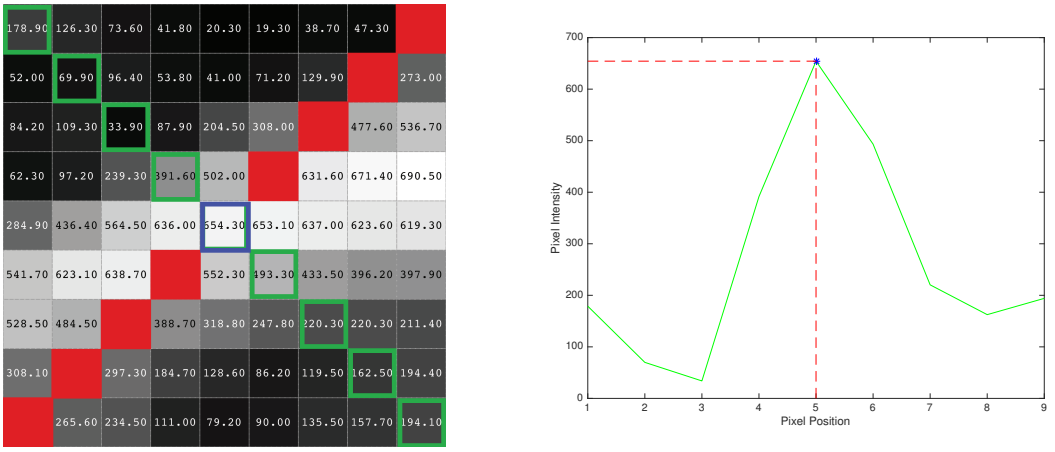
When utilising the Vector Magnitude approach, in order to maintain consistency in a discrete system, the gradient direction requires truncation to one of the fixed orientations defined earlier by the 26-connectivity neighbourhood for the 3-D region. To determine the truncated gradient direction of the surface plane, an efficient orientation filtering process using convolution can be applied. To achieve this a 3-D neighbourhood convolution operator is applied to the surface map image. The filter kernel is divided into 3 regions, a central divisive plane and 2 outer regions (An example of a $3\times3\times3$ of the filter kernel for a single orientation can be viewed in eq. 3.14). In order to confine the number of orientations to 13 (those derived from a 26-connectivity neighbourhood) The orientation filter neighbourhood is orientated through the same 13 orientations as the maximum response method detailed prior. When the filter kernel is convolved with the surface map, the output of the filter for a given location will be of a high intensity when the central plane of the filter kernel is positioned and oriented along a surface interface in the surface map. The direction which produces the maximum intensity determines the surface direction, and each filter kernel orientation is assigned a label corresponding to a direction perpendicular to the surface plane, providing the gradient direction. This process is applied to all the voxels in the image to produce a gradient direction map, which is an image which describes the truncated gradient direction for every position in the image.

$$h = \left\{ \begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{array} \right\} \tag{3.14}$$

One of 13 oriented 3 x 3 x 3 filter kernels for orientation filtering. All 13 masks resemble MR masks, with both A and B regions represented with 0s and the central dividing plane with 1s.

#### 3.4.1.2 Suppression of Non-Maximum Points

Using the labelled orientation data from the above process, voxels in the surface map along the gradient direction can be identified by selecting those voxels adjacent to the surface plane which intersect the central voxel for given distance (dependent upon the scale of the neighbourhood mask operator). The intensity of the voxels along the gradient direction are compared in order to identify the local maxima. If the central voxel in question is the peak value this would imply that the maximum amount of dissimilarity is located in this position, and therefore this is the actual location of the surface, thus this voxel value is retained. If the voxel in question is not the peak value, this means it is not at the exact location of the interface, thus this voxel should be excluded or suppressed, thus the value will be converted to zero. Figure 3.13a shows 2-D implementation for edge detection. The red shaded squares show the approximated edge direction, while the green squares signify pixels along the gradient direction. Values in the green squares are compared, if the the central blue square is the maxima (Fig. 3.13b), this pixel is retained, if it is not, its value is set to zero.




(a) NMS Process (b) Local Maxima

Figure 3.13: Example of 2-D NMS procedure a) Edge direction approximated by red pixels. Gradient direction is perpendicular to that of the edge (green). The pixel being examined (blue) is compared with pixels along gradient direction (green) to see if it is the local maximum. If it is that value is retained, if it is not, it is suppressed. b) Values along the gradient direction are plotted, the evaluated pixel in position 5 is the maximum (654), therefore this value is retained. The process is repeated for all positions in the image.

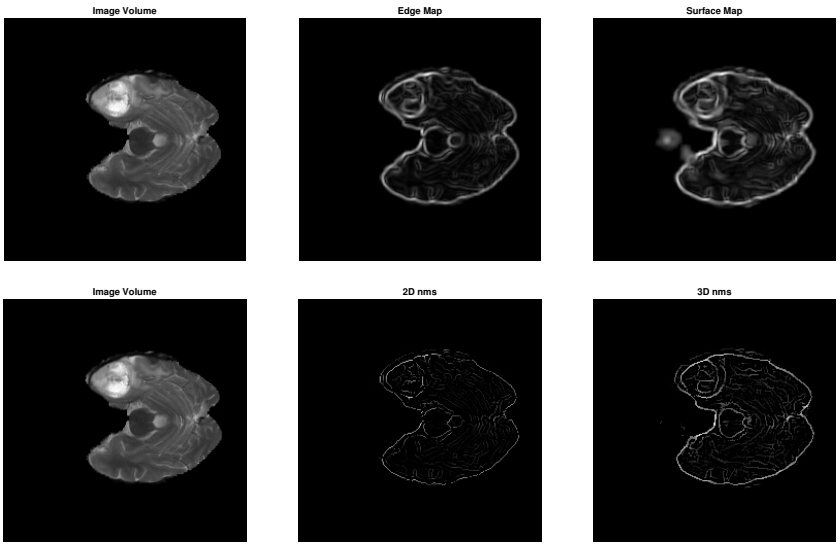



Figure 3.14: A comparison of 2-D and 3-D NMS. 2-D NMS does not determine maxima across $z$-dimension. Therefore surfaces which are maximal across layers are not returned resulting in a result which more closely represents a layered edge detection approach rather than a true 3-D result.

The results of 2-D and 3-D NMS can be viewed in Figure 3.14. Here the 2-D NMS results show that diffused edges have been thinned to single pixel thickness. This is essential for objective analysis when comparing to a ground truth interface which is also single pixel thickness. The primary difference between the the 2-D results and 3-D results in Figure 3.14 is that unlike the edge, the surface is not suppressed into a 1-D line, but instead a 2-D plane of single voxel thickness in a 3-D environment.

### 3.4.2 3-D Hysteresis Thresholding

After the application of a NMS stage, the resultant output is a 3-D image volume which possesses the local surface maxima information of the original image. For the purpose of objective performance evaluation, it is important to classify the NMS result into surface points and non surface points in order for a one to one correspondence performance metric to be applied, where the resultant image can be objectively compared against an ideal result (ground truth solution). There are a number of available options for classification. The most straightforward method is a point operation known as thresholding. Thresholding will classify a voxel as a surface point or non-surface point, by comparing the magnitude of a voxel in the output against a threshold value. Magnitudes above the threshold are classified as surface points, while magnitudes below are deemed non-surface points eq 3.15. Threshold operations could also include a range of magnitudes in which a voxel could be classed as a surface or non surfaces, with voxels that fall outside the upper and lower thresholds being classified as non surface points. These types of thresholding methods are considered to be uniform thresholding methods (Nixon and Aguado, 2002).

$$g(x,y,z) = \begin{cases} +1 & \text{if } f(x,y,z) > T, \\ 0, & \text{if } f(x,y,z) \leqslant T. \end{cases} \tag{3.15}$$

Where $g(x,y,z)$ is the 3-D binary image classified into surface points (1) and non-surface points (0), and $f(x,y,x)$ is the NMS result, $T$ is the threshold intensity level specified by a user.

More advanced methods of thresholding known as 'optimal thresholding' are not point operators, but group operators, and take into consideration the range of magnitudes of voxels, or the histogram distribution in order to set an optimal threshold level, one of the most common optimal thresholding methods is the Otsu's method (Otsu, 1979) A method to select a threshold automatically from a grey-level histogram by maximising the discriminant measure of separability of the resultant classes in grey levels. One of the characteristics of local surface maxima information is that the surface magnitude values are not consistent and are dependent on the dissimilarity of regions where the surface

interface lies. Nixon and Aguado (2002) indicates that for edge data, the variance in edge strength can lead to unconnected edges after a uniform method of thresholding is applied. Since surface detection and 3-D non maximum suppression methods follow the same principles for resolving interfaces, uniform methods of thresholding are also not optimal for surface detection. An adaptive method of thresholding can overcome the broken edge problem associated with uniform thresholding. One such method, and one notably exploited by Canny (1986) is a method described as Hysteresis thresholding.

Hysteresis thresholding is a post processing segmentation technique that applies recursion instead of being a simple point operator. Hysteresis thresholding assigns a positive or null label to each image voxel. Where surface segmentation is concerned, voxel labels are defined as surface and non surface points, and are represented by a logical True or False (1 or 0) within the digital domain. The property of connectivity for surfaces can be exploited using this thresholding method. Surfaces are not typically composed of individual pixels or voxels, but rather a continuous connected plane of voxels. Hysteresis thresholding requires two threshold values, an upper threshold limit (UT) and a lower threshold limit (LT). These can be independently defined to maximise performance. However in this work the lower threshold is set at 40% of the upper threshold to share consistency with common implementations of the Canny Edge detector which uses this ratio. When a surface point from the non maximum suppression out exceeds the magnitude of the predefined upper threshold level, this voxel is assigned the label 'True', thus classing the voxel as a surface point. Voxels with a magnitude value smaller than that of the lower threshold value are assigned the value 'False' thereby classing that voxel as a non surface point. This produces a third category of voxels, those that lie between the upper and lower ranges defined by the threshold. These voxels are potential surface candidates, the criteria for which defines their classification is the dependent upon whether or not these points share a connected path to surface point above the upper threshold limit. Through a process of recursion, voxels with a path connected to an upper thresh surface point are classified as a surface, those that do not are classified as non surface points.

The benefit of hysteresis thresholding over standard thresholding for surface segmentation is that since we know that surfaces are composed of connected pixels and voxels, weaker surface points resolved by the filter can be retained, while image noise can be greatly suppressed, preserving connectivity(Nixon and Aguado, 2002). Figure 3.15 presents examples of hysteresis thresholding set at different threshold levels, with the lower threshold set to be 40% of the upper threshold. With hysteresis thresholding there is a trade-off as $T$ is manipulated. Higher thresholds preserve only the strongest detected surfaces, this reduces the quantity of clutter in the results, however this comes

at a cost of detection, typically surface points will be missing and the image will present with connectivity issues. Selecting a low threshold should preserve most of the surface points, however, noise and clutter are likely to be erroneously classified as surface points leading to a sensitive result characterised by 'over-detection' features.

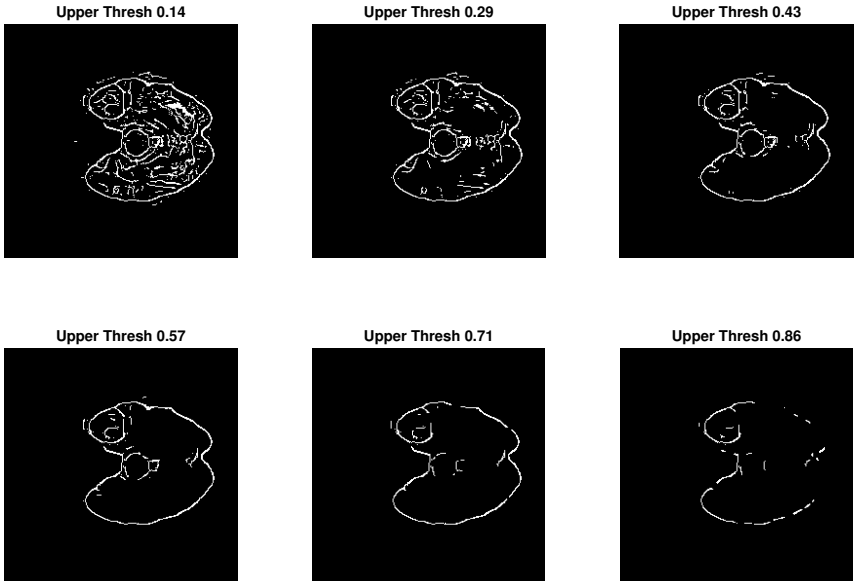



Figure 3.15: Example of hysteresis thresholding with varying threshold levels. As the threshold level is increased (lower thresh set at 40%), noise is reduced but at the cost of detection.

## 3.5 Summary

Presented in this chapter are a number of novel contributions. Firstly, two novel approaches of surface detection are introduced which exploit the statistical properties of different 3-D image regions in order to resolve the surface interfaces. The former is a 3-D adaptation of the statistical edge detection methods comprehensively evaluated by Williams et al. (2014). The latter method employs a method of computing a statistical feature gradient, which is a superior technique in terms of computational efficiency, due to reduced redundancy in the number of calculations required. Furthermore, it utilises a scale invariant approach such that the computational complexity increases proportionately when the size of the neighbourhood mask is increased, this is in contrast to the 3-D maximum response and the 2-D method on which it is based, which are required to process more orientations when the scale is increased.

For some common types of 3-D image formats, the resolution of the data is anisotropic (Du et al., 2020). This can be a problem for surface detection methods, in which the surface map is distorted and biased towards surfaces which exist in particular directional planes (Brejl and Sonka, 2000). A further contribution is the ability to apply coefficients which weight the directional components of the L2-statistical gradient vector, mitigating some of the distortion effects which arise in surface maps when the resolution of the image is not isotropic.

While the statistical tests, and the impact of neighbourhood scale employed within the statistical detections method have already been evaluated by Williams et al. (2014) for 2-D edge detection, they have been reintroduced here in a novel 3-D surface detection context, including a 3-D implementation of the Robust Rank Order test which has yet to be comprehensively evaluated in either 2-D or 3-D data.

In addition to the statistical detection methods, contributions were also made in the form of a novel efficient 3-D orientation filtering method, the purpose of which is to determine a surface gradient direction. This was created to be used in conjunction with a new 3-D non-maximum suppression algorithm for when surface orientation is not known a-priori.

This chapter also provides a computational cost analysis of the newly introduced Maximum Response and Vector Magnitude statistical detection techniques, across different scales and statistical tests. Including discussion of how convolution can be utilised for fast application of parametric statistical tests for surface detection. The next chapter will explore methods for how to accurately assess the performance of surface detection techniques.

# Chapter 4

# Objective Methods for Surface Detection Performance Evaluation

The primary findings of this chapter were published at the Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Smith. and Williams., 2020)

## 4.1 Introduction

In Chapter 3 a novel method of statistical surface detection has been introduced, however any new advancements in surface detectors irrespective of the approach (morphological, topological, model based or machine learning) should be objectively evaluated to determine the valued improvement over prior methods and thus determine the potential contribution. Yet objective analysis of statistical surface detection techniques remains mostly unexplored. A comparison between statistical surface detection and existing classical methods is necessary in order to establish the performance limitations of each surface detection algorithm under the various conditions which may be encountered.

Performance evaluation for edge detection has been a long-standing research topic, for example as early as 1975, Fram and Deutsch (1975) introduced a requirement for finding reliable metrics that correspond to parameters which accurately affect the edge detection's performance. However, this requirement is not always adhered to (Jain and Binford, 1991; Papari and Petkov, 2011).

Performance measures for edge or surface detection can be broadly broken down into two categories, qualitative and objective. Qualitative measures are analytical and typically consist of single or multiple human assessors grading the edge or surface detection results subjectively against specific criteria. Objective performance measurements differ in the fact that the analysis is not graded subjectively, but instead graded by a defined mathematical calculation applied to the detection result. This allows for the systematic comparison of different algorithms in a repeatable, quantifiable manner often to find the optimal algorithm for a specific task (Lopez-Molina et al., 2013).

This chapter explores the methodologies available to evaluate the performance of surface detection algorithms. Detailed are the limitations of available objective techniques, which at present leaves 3-D performance evaluation largely unexplored.

A contribution to the body of knowledge in performance detection is also presented in this chapter in the form of a novel approach for fast and accurate correspondence matching. Accurate matching is necessary for the correct application of a number of performance metrics, however current matching algorithms are very computationally costly to when applied to 3-D data, the newly introduced Efficient Paring Strategy (EPS) allows for accurate performance metrics to be applied to large 3-D datasets and to be evaluated in a practical time frame.

## 4.2 General performance evaluation approaches

Qualitative methods of analysis typically consist of a visual inspection of results. However the fundamental issue with qualitative approaches is the inconsistency introduced by subjective Human visual analysis. This limits the ability to make an accurate comparison with existing research since the grading criteria is subjective (Papari and Petkov, 2011). It is common practice in edge and surface detection research, as well as most computer vision research to present the visual results with some degree of subjective assessment that contextualises the results. This allows the reader to draw their own conclusions(van Vliet et al., 1989). It should be noted that efforts have been made to improve upon and standardise qualitative approaches, for example hybrid methods also exist, which combine elements of both subjective and objective analysis such as in the work of Heath et al. (1996). However, in order to make accurate consistent comparisons, reduce biases and eliminate as much subjectivity as possible, objective performance measures which provide a quantifiable metric score are required (Papari and Petkov, 2011).

### 4.2.1 Quantitative analysis strategies

In general, quantitative, or objective performance measures can be further classified into two types, reference based and non reference based (Nercessian et al., 2009). Non reference based techniques grade the edge detection result based on known favourable edge response properties such as connectivity or uniformity, while reference based methods compare the edge or surface result against a reference (ground truth) solution. A reference solution depicts the ideal response for a given image, and the objective performance measure applies some method of calculation which measures the similarity of the result against the reference.

Non reference based performance measures have the advantage of avoiding the need for ground truth data, this is beneficial since acquisition of reference solutions for real imagery is non-trivial. The primary barrier to the acquisition of accurate reference solutions is that real images are complex and feature rich (Milan et al., 1993). This makes the edges and surfaces in real or natural images highly subjective to an observer due to the many possible interpretations of what constitutes a surface interface. In addition, there can be ambiguity in the precise location of an interface where a change in region properties are more gradual and the change between regions is less defined (Nercessian et al., 2009). Furthermore, manual creation of reference data for real images is typically performed using a 2-D display of the image data, with a single individual layer view at a time, therefore any surfaces that exist between the layers are not be observed by the person manually labelling the data and therefore can be missed.

### 4.2.2 Non reference based assessment.

Non reference based assessment methods are applied directly to the output of the edge or surface filter, therefore only require a test images and the detection result with no reference image. The test image is typically a real world or natural image, and can be from a wide range of imaging modalities, such as a digital photograph, or an MR image. The second image used is the edge or surface map response from a detection algorithm. Non reference based techniques are able to assess aspects of edge quality, through a process of analysing the structure of the output image properties irrespective of the accuracy of the edges.

For example, the connectivity and uniformity of the edges can be assessed such as in the work of Zhu (1996), additionally edge coherence can be used as a measure of edge evaluation (Kitchen and Rosenfeld, 1981). Haralick (1994) uses a covariance propagation technique to characterise the performance of an algorithm, this technique was expanded upon to assess uncertainty in 2-D corner points (Yi et al., 1994) and further used by Ramesh and Haralick (1994) to evaluate the performance of gradient-

based edge detection methods. While Cho et al. (1995) used a method of bootstrapping to evaluate edge detection.

Non reference methods provide a measure of edge or surface quality, which can be useful for determining how fit for purpose the outputs are for higher level operations, such as segmentation, since certain properties such as unconnected edges or duplicated edges can impede the performance of the higher level operations. However, the fundamental issue with non-reference based methods is that they do not measure the accuracy of a detected surface or indicate how effective an algorithm is at resolving interfaces between regions in the image.

Non reference based methods also suffer many biases and are restricted in their use due to the difficulty in assigning metric scores to a subjective element such as edge quality (Nercessian et al., 2009). Real images are complex and feature rich, thus the edges and surfaces in real or natural images are highly subjective to an observer. This makes it non-trivial to effectively incorporate the original test image in order to assist in the evaluation.

### 4.2.3 Reference based assessment.

Reference based measures objectively compare the result image against a corresponding ground truth (ideal) solution. This can be achieved through a variety of metrics and performance evaluation methods which are widely represented in the literature in a range of contexts (Savitzky and Golay, 1964; Abdou and Pratt, 1979; Cyganski et al., 1995; Borra and Sarkar, 1997; Forbes and Draper, 2000; Bowyer et al., 2001; Prieto and Allen, 2003; Martin et al., 2004).

Crucially, these kinds of processes allow for an accuracy measurement of the detected surfaces, since every point in the surface detection result can be directly compared against every point in the reference solution. Allowing for the number of correctly detected surface points (True Positives, TPs), correctly determined non surface points (True Negatives, TNs), missed surfaces (False Negatives, FNs), incorrect surfaces (False positives, FPs), to be counted. The main advantage of reference approaches is that they can be used to determine edge detector performance within a controlled environment where edge pixel locations are known, and without subjective biases (Nercessian et al., 2009).

These methods require two types of images for the assessment.

- Result Image
- Ground Truth Image

The typical procedure for reference based analysis is as follows, firstly, a test image is required. This can be a synthetically created image, or a real image. If synthetic, the image is designed to possess features closely resembling interfaces in a real images that match the context in which the detection method would be typically used. Applied to the test image is the detection method being evaluated, producing the 'result image'. In its raw output form the result image is usually some variation of an edge or surface map. Additional processing steps can be applied to classify edge and non edge points, such as thresholding. Thus, the result image is simply the output from an edge or surface filter. The result image is then assessed for accuracy with the ground truth image using one or more of several objective performance algorithms such as the Pratt figure of merit (PFOM)(Abdou and Pratt, 1979), Probabilistic Rand Index(PRI) of Savitzky and Golay (1964), Pixel Correspondence Metric (PCM) of Prieto and Allen (2003), Receiver Operating Characteristic (ROC) Curves of Bowyer et al. (2001), the Precision (P), Recall (R) and F-measures of Martin et al. (2004), and also the Variation of Information (VI) measure by Meilă (2005). Each of these methods produce an objective metric which grades how accurately the result image correlates to the the ground truth image.

### 4.2.4 Limitations

The ground truth data used in objective frameworks with 'real-world' test images are manually segmented boundaries of objects. This presents difficulties since the acquisition of ground truth solutions for 'natural' or 'real-world' images is, as previously stated, a non-trivial problem (Nercessian et al., 2009). Real images are complex, they possess many details, artefacts and features that lead to inconsistent, time consuming manual contouring. Attempts have been made to create robust real image datasets with ground truth data to be used for performance analysis. For example the University of South Florida dataset in the work of Heath et al. (1996) and Bowyer et al. (2001), also the Berkeley segmentation dataset applied in the work of Martin et al. (2004). These methods can offer a good objective evaluation framework for 2-D boundary detection and object classification methods, however their application is limited to 2-D and the subjective nature of how a boundary is defined in these types of dataset is not fully mitigated.

An alternative approach is to use synthetically generated image volumes, whereby the edge or surface location is predefined, then a simple image is constructed with carefully controlled parameters to define the region profiles characteristics. This approach is not without its problems, Forbes and Draper (2000) shows that synthetic data for testing

purposes often does not correlate with real world performance, citing the manipulation of the test image properties and parameter selection of the filters in order to manipulate favourable results. Bowyer et al. (2001) further supports this and cites the reasoning for poor correlation is due to synthetic data which is too simplistic, not accounting for interface topology, noting the lack of interface curvature, interface strengths, and more importantly the presence of multiple different interfaces of different strengths present in the same image.

With edge and surface segmentation, there is typically a trade off in sensitivity and noise suppression in respect to parameter selection. Filters can be made more sensitive, allowing for weaker interfaces to be resolved, however this can also lead to over detection of edge and surface points, alternatively, parameters can be selected which reduces over-detection artefacts, either through increased pre-processing smoothing, or through a larger scale parameter, in turn this makes weaker edges and surfaces more challenging to resolve. For single interfaces, this trade-off can be mitigated through precise parameter selection and a post processing thresholding procedure which differentiates between the spurious responses and interface response. The introduction of multiple interfaces in a synthetic image volume results in a more challenging thresholding scenario, since some edge or surface interfaces will be weaker than some of the spurious voxels, thus the spurious responses can no longer be removed through thresholding without removing some of the detected weaker edges or surfaces.

The reduced complexity in synthetic image detail enables tests to be performed on particular aspects of a filter in isolation. The effects of interface type can be examined by resolving interfaces between differing statistical region profiles. The effects of interface topology can be examined by changing the structure of the interface while controlling for interface type and strength. The creation of synthetic data with multiple interfaces also assesses the robustness of an edge or surface detection algorithm with respect to resolving power (the ability to resolve interfaces while minimising over-detection).

## 4.3 Current Methods

Current reference methods are not limited to the calculation of TPs, FPs, TNs and FNs in order to produce a performance metric. A surface detected in an incorrect position would typically be a more useful result than a surface missed entirely. Fig 4.1 presents an example of a displaced edge in 2-D data, and the effect a displacement of 1 pixel has on the metric score using a Pratt Figure of Merit analysis (Abdou and Pratt, 1979). Here the displaced boundary still gets a good score (0.9). Metrics which do not consider displacement, such as a *Dice score* or *Jaccard index* (Bertels et al.,

2019), would observe all positive cases as spurious responses (False Positives or FPs), and incorrectly determine a number of False Negatives (FNs) leading to a potentially unreliable metric score of 0.

(a) Reference (b) Displaced boundary

Figure 4.1: Pratt figure of Merit. In this 11×11 example, a score of 0.9 is achieved for a surface with a displacement of n=1 (one position to the right).

Metrics which do not consider displacement are better suited to evaluating region segmentation tasks, since the omission of boundary displacement data does not significantly effect the performance metric score in this context. A pixel shift of $n = 1$ of a large region does not result in a significant change in the number of correctly determined TPs and FNs. However for an edge or surface, a single pixel or voxel shift is enough to miss the entire interface. Edge and surface images typically contain at least an order of magnitude more true negative points compared with true positive points, and therefore a metric which accounts for displacement is more appropriate. This is because a minor displacement of an edge or surface in a detection result is often sufficiently fit for purpose. This is a significant advantage of distance based metrics over other reference based performance measures in the context of edge and surface evaluation, as distance metrics provide a careful balanced appraisal of the detected interface (van Vliet et al., 1989). There are a number of distance based metrics in the literature which employ the sample principle of quantifying performance using a rudimentary measurement of distance ($d_i$) including the commonly applied metrics defined in equations 4.1, 4.2 and 4.3.

$$MSD = \frac{1}{I_D} \sum_{i=1}^{I_D} (d_i)^2 \tag{4.1}$$

Mean square deviation (MSD)

$$MAD = \frac{1}{I_D} \sum_{i=1}^{I_D} |d_i|^2 \tag{4.2}$$

Mean absolute deviation (MAD)

$$PFOM = \frac{1}{\max\{I_D, I_i\}} \sum_{i=1}^{I_D} \frac{1}{1 + \alpha(d_i)^2} \tag{4.3}$$

Pratt Figure of Merit (PFOM)

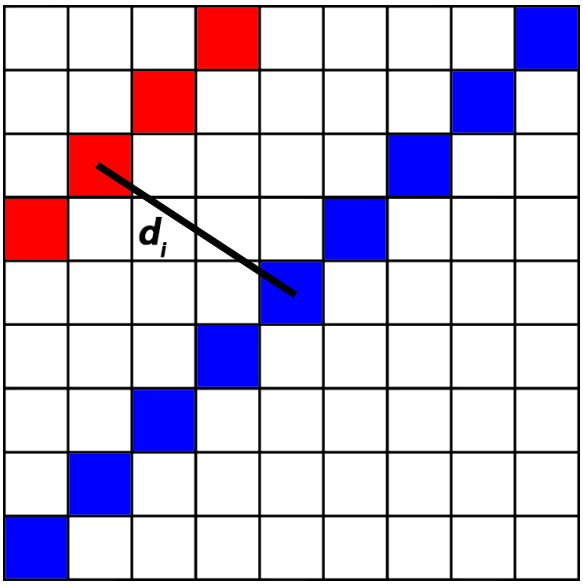


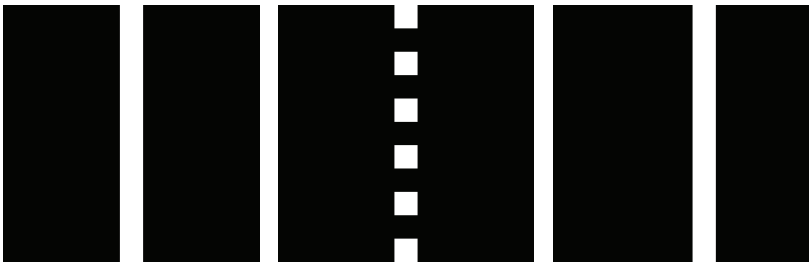

Figure 4.2: $d_i$ is an array of values corresponding to the euclidean distances between detected surface pixels (red) and reference surface pixels (blue). It represents the amount of displacement between the detected edge, and the reference edge. This examples shows one of many ways in which $d_i$ is determined, here by the Euclidean distance. However determining which pixels should be matched to form the displacement measurement correctly requires a correspondence matching process.

Where $I_D$ is the number of detected surface points, $I_I$, the number of ideal surface points (ground truth), $\alpha$ a scaling constant to ensure a score between 0 and 1, and $d_i$ the surface deviation for the i'th detected surface pixel and is measured as the Euclidean distance between the corresponding pixels.

The Pratt figure of merit is a prominent metric for 2-D edge detector evaluation (Abdou and Pratt, 1979; Peli and Malah, 1982; van Vliet et al., 1989). This is achieved by utilising the distance measurement between the position of a detected surface point and the corresponding surface point in the reference image as illustrated in Fig 4.2.

The main issue with this approach is determining which two positive points correspond with each other between the reference and result image. The Pratt Figure of Merit simplifies this by matching the closest positive pixel in the ideal and test images. This presents a problem since the method does not factor in whether the pixel evaluated for $d_i$ has previously already been matched. As the metric also factors the disparity

between the total number of detected edge points in the image and the total number in the reference, reusing previously matched edge and surface points does not always produce a reliable score. Ideally a displaced edge is preferential over a fragmented edge, however this is not reflected by a Pratt Metric score. Fig 4.3 presents a case where a fragmented edge in 2-D data, provides a superior metric score than a displaced edge.



(a) Ground Truth (b) Fragmented edge 0.9545 (c) Displaced edge 0.9000

Figure 4.3: Pratt figure of Merit, in this 11×11 example, the PFOM does not penalise fragmented edges, leading to preferential metric scores for fragmented interfaces over displaced interfaces

Approaches which do not factor in previously assigned pixels are referred to as one to many correspondence techniques and include the metrics defined in equations 4.1, 4.2 and 4.3. This problem is further exacerbated when the number of ideal points in the reference and the number of detected points are not equal. This is because it allows for multiple-to-one and one-to-multiple correspondence between the reference image and the detection result. Multiple to one correspondence matching can lead to inconsistencies in determining how $d_i$ is defined. This is due to the fact that the previously matched pixels are required multiple times in the assessment.

As a consequence, when the data is noisy, and more spurious responses exist in the edge or surface image, or cases where interfaces are fragmented, the instability of these types of metrics can lead to inflated performance scores which do not accurately reflect with the visual assessment of the result, likely rendering the result invalid. Fig. 4.4 shows how the discrepancy between the number of points in the reference and result image can lead to inconstant measurements of $d_i$.

To further complicate matters regions can often transition from one region to another over a gradient, where it is not clear where in the transition the exact location of the interface exits. This 'spreading-out' or 'smearing' of the interface can be problematic for surface detection algorithms, with unconnected, displaced or duplicated boundaries being produced. It can also be problematic in defining where the interface actually

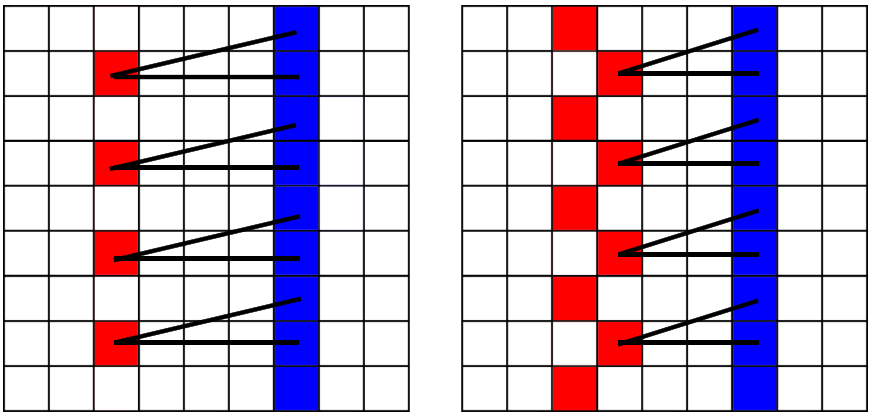


Figure 4.4: Examples of the inconsistencies in defining $d_i$. In the first instance, the detected points (red) are being used multiple times in the performance evaluation for determining $d_i$ from the reference interface (blue), inadequately penalising the fragmented surface. In the second example, $d_i$ is established by pairing the closest detection to the reference, however this ignores many of the surface points

lies in a ground truth image. Therefore, when assessing the performance of a surface detection algorithm an allowance for displacement should be available to account for these localisation errors, since detected, connected boundaries even with a displacement are of value.

### 4.3.1 Accounting for Localisation Error

Bowyer et al. (2001) recognised that image context is important when considering the noise or spurious content in the result. False positive detection results in contextually unimportant regions, such as those sufficiently distant from an ideal edge, can be considered spurious. However, some detected points could be in close proximity to an ideal edge or surface, and could be considered a correct detection, but with some degree of localisation error.

Compensation or tolerance for displacement of boundaries is a consideration that cannot be ignored since digital images are comprised of discrete data, thus the location of an edge or surface point in a filter response is constrained by the pixel or voxel resolution of the image (Fig 4.5). However, since an edge or surface is the interface between regions, and different regions occupy adjacent voxels, this means the true position of a surface cannot be accurately represented by a discrete pixel in an edge map or voxel in a surface map. A surface detection algorithm must position a surface in accordance with the discrete framework of the image and this introduces location error within a margin of at least 1 voxel. Most objective measures do not account for a pixel or voxel shift between the detected interface and the ideal interface (Fig 4.5) or do not account for

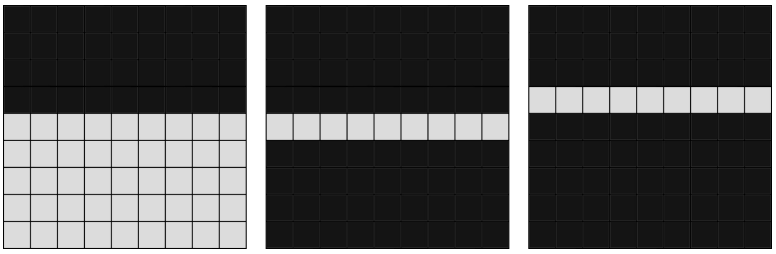


Figure 4.5: Displacement of edge. True edge fits between the 2 rows of pixels, however the edge must conform to the pixel grid, therefore when maintaining single pixel edge thickness, this results in 2 potentially correct edge locations, which needs to be accounted for during performance evaluation.

the fragmentation of the detected interfaces which can result in incorrect assumptions about the quality of the detection **Williams2008**.

To account for localisation error introduced by the discrete nature of images, a tolerance of at least one unit (pixel or voxel) should be required as minimum. However, this window or tolerance zone (defined as $T_{match}$) can be increased to include a larger area (Fig 4.6) to account for detection methods which introduce a shift in the surface response. If a detected surface falls within this tolerance zone, the detection point should be classified as correct. The fundamental problem with this approach is that it increases the number of ideal points in the reference image, and then a one to many correspondence imbalance can occur, where there are more true positive points in the result image than the actual ideal points in the reference image.

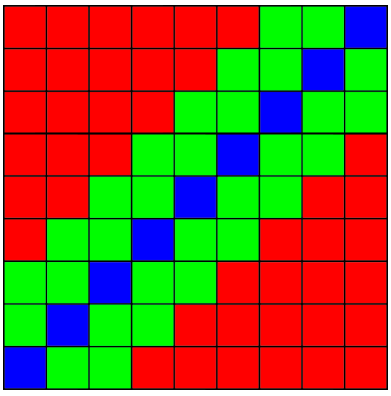


Figure 4.6: Example of $T_{match} = 2$. Blue pixels indicate ideal response. Detection points found in red zone are automatically determined as spurious, points detected in $T_{match}$ zone (green zone) are assessed for an assignment match

In order to avoid the instability in measuring $d_i$, a method of forcing one to one correspondence between ground truth and result image is required (Liu and Haralick, 2000). Whereby each detected point in the result image corresponds to a detection point in the reference image. It is important to compute the correspondence of the images in

order to penalise multiple detections (Martin et al., 2004), since single detection is one of the three important criteria of edge detection laid out by Canny (1986) for assessing optimal detection.

Establishing one to one correspondence matching between reference ideal points and detected points are not trivial problems (Liu and Haralick, 2000; Liu and Haralick, 2002). Some detection points that fall in the tolerance zone and not in the ideal position should be recategorised as True Positives. Some consideration of how this classification is made is required, as there are a number of inefficient or sub-optimal ways to pair the detection points.

Forbes and Draper (2000) paired farthest distance pixels in the tolerance zone with an ideal edge pixel. Bowyer et al. (2001) instead opted for the closest distance matched (CDM) pair, in both cases, the detection points in $T_{match}$ region are ranked in order of proximity to the correct detection location. Assignments are made based on based upon the detections closest to the reference, once a match has been made, the pixel or voxel cannot correspond with another point in the reference.

However, multiple missing points and multiple displaced points in the result image can lead to a confusion of which correspondences should be made. Using the CDM approach of Bowyer et al. (2001), correspondences are prioritised simply by the order in which they are processed, but this does not always offer the most consistent approach.

### 4.3.2 Classical Assignment Problem.

Liu and Haralick (2000) addressed the inconsistencies with the CDM approach by developing a strategy for creating optimal one to one correspondence, implementing a process which treats $d_i$ as a cost function which should be minimised. This was achieved by framing the task as the classical mathematical assignment problem which has already been optimally solved.

The assignment problem is a fundamental combinatorial optimization problem which is well established in the field of mathematics, in which: *'The problem instance has a number of agents and a number of tasks. Any agent can be assigned to perform any task, incurring some cost that may vary depending on the agent-task assignment. It is required to perform as many tasks as possible by assigning at most one agent to each task and at most one task to each agent, in such a way that the total cost of the assignment is minimized.'*. The optimal solution to this problem can be solved using the *Kuhn-Hungarian* algorithm (Kuhn, 1955).

In re-framing this problem for one to one correspondence matching for edge and surface detection evaluation, an agent can be assumed to be a missing surface point, while a

task can be a nearby spurious point in the test image. The cost function is the Euclidean distance between the points. By framing it in this fashion the problem can be solved using the Hungarian algorithm. It achieves this by finding minimum-weight matchings in bipartite graphs. However, the main barrier of using the Hungarian algorithm for providing one to one correspondence matching is that it is very computationally expensive with a time complexity of $(O(n^4))$, Munkres (1957) later reduced this to $O(|V|^3)$. The current most efficient method for solving the assignment problem is the *cost scaling algorithm* (CSA) implementation of Goldberg and Kennedy (1995) at $O(|V|^2)$ but these approaches remain very computationally expensive process when all pixels and voxels are considered for a correspondence match.

Liu and Haralick (2002) using the $T_{match}$ principle were able to constrain the number of potential matches to a limited range and reduce the complexity of the problem, producing an optimal method suitable for performance evaluation of small 2-D datasets using the Hungarian algorithm that can be completed in a practical time-frame. However for 3-D data and large evaluation datasets, a more practical approach is required.

Further attempts have been made to constrain the problem and simplify the task such as in the work of Martin (2003) using an optimised BiPartite graph method, and Prieto and Allen (2003) similarly used weighted matching in Bipartite graphs to create the 2-D pixel correspondence metric (PCM) for one to one correspondence matching. However these methods still remain computationally expensive when compared with one to many correspondence metrics, and when dealing with large evaluation datasets or 3-D data, analysis can be impractical.

Liu and Haralick (2002) made recommendations for enforcing a maximal one-to-one correspondence between ground truth and result images as a general principle to follow in evaluating edge detection performance. However, while the optimal strategy for solving the assignment problem can be achieved using the aforementioned processes, strategies to establish $d_i$ is feasible for low resolution 2-D images or for small data sets only. When considering the 3-D images of surface detection, the increased complexity arising from more potential positions for a match only adds to the computational cost. This complexity is compounded by the fact 3-D images typically contain a greater amount of voxels than a 2-D image contains pixels, making this strategy at current computational speeds resource intensive and often impractical for establishing $d_i$.

Additionally, a reliable comprehensive analysis require the performance metrics to be applied to many thousands of results, for example datasets such as BRATS (Menze et al., 2015) are large, and many thousands of results can be generated. In order to undertake an evaluation in a reasonable time-frame and maintain the most representative objective

accuracy, an efficient method of solving the assignment problem for this case is required.

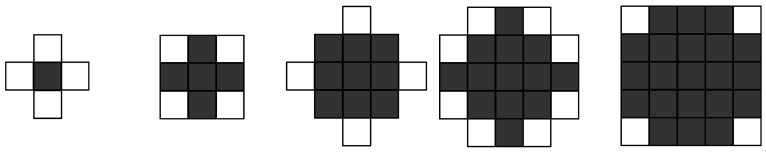
## 4.4 Efficient Paring Strategy (EPS)

Presented here is a novel surface pairing strategy based on the work of Liu and Haralick (2000) and Liu and Haralick (2002). The strategy was developed to assign displaced detected surface points (False Positives) in the result image to missing surface point locations (False Negatives) in order to achieve optimal one to one correspondence matching efficiently.

The EPS procedure aims to closely replicate the accuracy of the Hungarian algorithm (Munkres, 1957), creating a fast approximate solution to the assignment problem using a novel inflationary zone method within a defined $T_{match}$ region. This method is an adaptation of the Closest Distance Match (CDM) method of Bowyer et al. (2001). The adaptation allows for a more consistent metric, which more closely replicates the correspondence matching of the optimal Hungarian method utilised by Liu and Haralick (2002). However, unlike methods which try to solve the assignment problem, this technique is suitably fast for analysis of 3-D image volumes or fast computation of multiple 2-D images in large data sets.

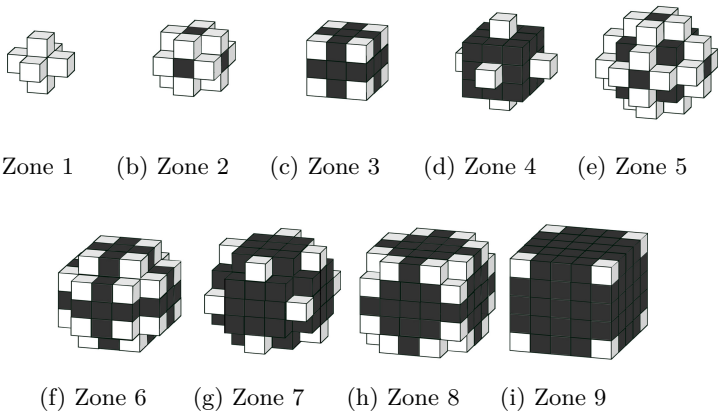
### 4.4.1 Technique

The method utilises a concept of zones within a local neighbourhood window. Zones are defined to be regions within a local $T_{match}$ neighbourhood window, which occupy the same Euclidean distance from the central pixel. Zones are ranked in levels from closest to farthest from the central pixel. 2-D examples are presented in Fig 4.7, while 3-D examples are presented in Fig 4.8.



(a) Zone 1 (b) Zone 2 (c) Zone 3 (d) Zone 4 (e) Zone 5

Figure 4.7: Visual representation of 2-D Zones. Each zone, levels 1-5 in white, possesses a cost function equal to the Euclidean distance to the centre of the missing edge point. Locations from prior zones are shaded dark grey.



(a) Zone 1 (b) Zone 2 (c) Zone 3 (d) Zone 4 (e) Zone 5

(f) Zone 6 (g) Zone 7 (h) Zone 8 (i) Zone 9

Figure 4.8: Visual representation of 3-D Zones. Each zone, levels 1-9 in white, possesses a cost function equal to the Euclidean distance to the centre of the missing edge point. Locations from prior zones are shaded dark grey.

In order to compute one-to-one correspondence, and avoid creating multiple partners for each candidate, each match needs to be computed concurrently. For every false negative (FN) in the result edge and surface image, a set of 2-D or 3-D zones derived from a $T_{match}$ scaled neighbourhood is established, centred on the FN location.

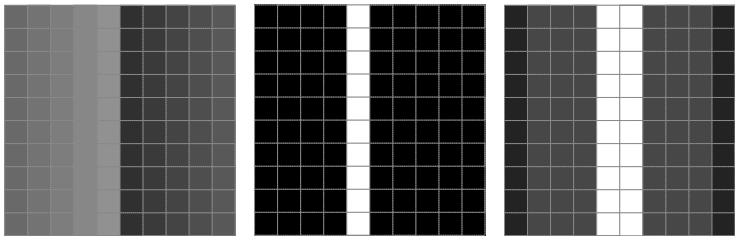
In each zone the number of potential voxel candidates for a match are counted. Then in order to minimise the assignment cost, the closest match is preferential, thus pairings made between voxels in the ideal and surface detected image are made in the lowest level zone first. However, as some FNs share the same candidate match, FNs with the fewest candidates are assigned a match first. This technique produces a closest distance match correspondence, however, unlike the CDM method, the matches are optimised such that the maximum number of correspondences are produced.

Once all candidates from the first zone are exhausted, if any FNs remain, the procedure repeats with the next zone and so on until each zone in the $T_{match}$ neighbourhood has been checked in its entirety. Any missed responses in the result without a pairing are classified as actual False Negatives, while any FPs in the result without a pairing are determined to be correctly labelled false positives. From here a number of performance methods can be applied. Either Precision recall based such as in the work of Bowyer Bowyer et al., 2001 but also a distance metric score can be applied such as in the work of Prieto Prieto and Allen, 2003, .

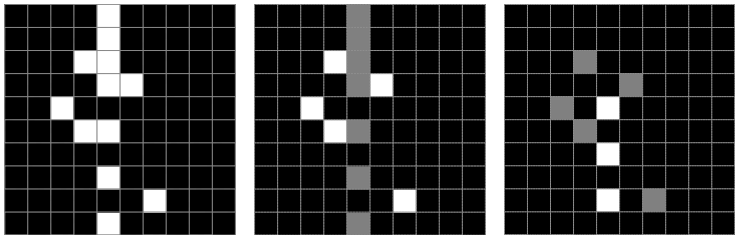
### 4.4.2 EPS Examples

A detailed step by step walk-through of the this procedure is given in the following section. Presented for clarity is a 2-D example case for correspondence matching of FN

and FP responses within a designated $T_{match}$ neighbourhood. For 3-D examples, cubic $T_{match}$ regions and inflationary zones can be used (Fig. 4.10f-n).



(a) Example Image (b) Reference Image (c) Edge Map

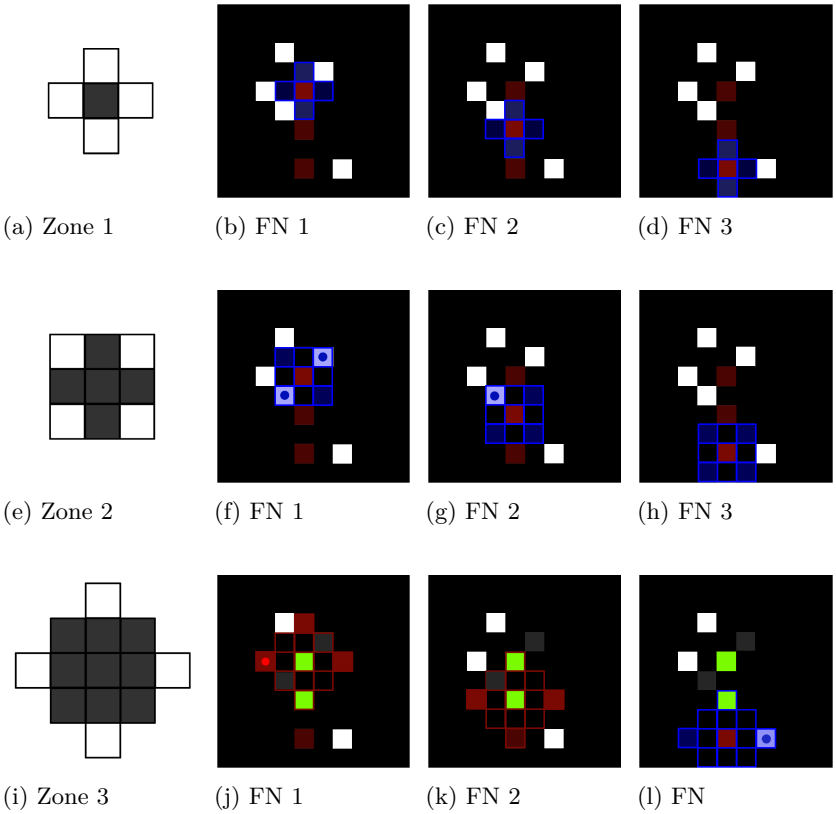


(d) Binary map (e) Count and remove TPs (f) Locate FNs and unmatched edges

Figure 4.9: Illustration of matching process. a) 2-D Example image. b) Reference image. c) Edge map filter response. d) Sub optimal binary map from post processing (e) An image showing the position of TPs (Labelled grey) which need to be counted and then removed from the image. (f) The location of the missing edge points (FNs) then need to be identified (white), in addition to the unmatched edge points (grey) .

1. First apply an edge detection algorithm to an image(Fig 4.9a) in order to obtain an edge map (Fig 4.9c) For 3-D data this would be a surface detection algorithm.
2. The EPS requires both the reference image(Fig 4.9b) and filter response (Fig 4.9c) to be binary data, therefore post the processing methods NMS and hysteresis thresholding need to be applied to label edge and non-edge points, thereby creating a binary edge or surface map (Fig 4.9d)
3. The result and reference images must then be compared using a Boolean "ADD" operation in order label all the true positives (TP) in the result image. These are counted and the total is stored (Fig 4.9e).
4. The TP points are no longer required for the correspondence matching and are therefore removed from the result image such that they are not counted multiple times (Fig 4.9e.)
5. Next the location of all False Negatives (FNs) in the result image are determined, as these locations require a correspondence match. (Fig 4.9f)

6. Then identify and label unmatched edges (Fig 4.9f)
7. In the FN locations, establish a set of zones designated by the $T_{match}$ neighbourhood (here $5\times5$) in order to create a list of potential candidates for a correspondence match.
8. The algorithm aims to produce optimal matching with a minimised $d_i$ cost function. In order to achieve this, FN locations which have the fewest number of candidate matches need to be identified. Therefore for each zone, count the number potential correspondences between the FN locations and FP responses. These candidates are signified by blue points in Fig 4.10.
9. To minimise the cost of $d_i$, establishing closest distance matches should be prioritised. This is achieved by pairing FN locations with FPs in the lowest level available zone first.
10. In order to maximise the number of possible correspondence matches within the zone, paring are to be made starting with FN locations with the fewest available candidates. This is because candidates with more than one potential match have more paring options, and as the number of correspondence matches increases, this reduces the availability of remaining potential parings.
11. Once a paring has been made, remove the $T_{match}$ neighbourhood from that location and remove the FP response. (In this example, Zone 1 has no potential matches)
12. When all matches from the zone have been established, or all candidates from the zone have been exhausted, repeat the process with the next zone
13. Continue this process through each zone until all FN locations have been assigned a match or when the $T_{match}$ neighbourhood has been exhausted.
14. Unmatched FNs are determined to be actual FNs, while unmatched FPs are determined to be actual FPs and are counted as such.
15. Each pairing is assigned a cost which relates to the zone from which the paring was made, the cost is therefore the Euclidean distance, thus producing a distance function of the pairings. As 2 points have been matched from zone 2 and one from zone 3, this provides a distance function of [1.41,1.41,2] which can be used for a distance based metric.



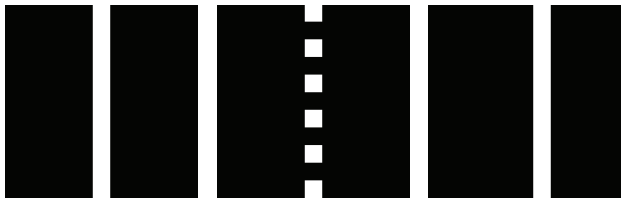
(a) Zone 1 (b) FN 1 (c) FN 2 (d) FN 3

(e) Zone 2 (f) FN 1 (g) FN 2 (h) FN 3

(i) Zone 3 (j) FN 1 (k) FN 2 (l) FN

Figure 4.10: An example case of one to one correspondence matching using a set of 2-D zones. Here each zone is signified by a blue grid, unmatched FNs are red, unmatched FPs are white. While matched FNs are green and matched FPs are shaded grey. ($T_{match}$: 5×5). In the example case, Zone 1 has no candidates for a match. Zone 2 FN1 and FN2 locations each have candidate matches. (f) FN1 has 2, (g) FN2 has 1, (h) while FN3 has zero. Since FN2 has the fewest potential candidates (1) it is assigned first, then the remaining candidate FN1 is assigned. (i) Zone 3 is processed next, where the final remaining FN is assigned a match and the pairing process is completed

By using one-to-one correspondence matching, a more reliable measure of performance can be established. To highlight the difference in outcome between the one to many PFOM and that of the EPS method, the earlier example is re-examined in Fig 4.11. Here $d_i$ is established using both EPS method and the traditional PFOM method in the Pratt figure of merit calculation (Eq. 4.3). The difference in the metric score for a fragmented edge is significant, while the advantages of accounting for displacement are maintained.



(a) Ground Truth (b) Fragmented edge (c) Displaced edge

Figure 4.11: Comparison of metric scores between the one to many Pratt Figure of Merit and the one to one EPS on two types of incorrect edge results. Fragmented edge: (PFOM 0.9545 EPS 0.5455). Displaced edge: (PFOM 0.900 EPS 0.900). The results indicate that the EPS penalises a fragmented edge while the Pratt figure of Merit does not leading to an unreliable objective performance score.

## 4.5 Evaluation of Correspondence Matching Techniques

### 4.5.1 Analysis of Accuracy

The EPS method presented is a fast alternative to the Hungarian algorithm for one to one correspondence matching, this technique allows for the rapid evaluation of surface detection algorithms for large 3-D datasets. The Hungarian algorithm is considered optimal since it solves the assignment problem with the smallest possible aggregate cost function $d_i$. Thus the accuracy of the EPS method and alternate methods for one to one correspondence therefore require direct comparison against the Hungarian algorithm.

#### 4.5.1.1 Accuracy Test Data

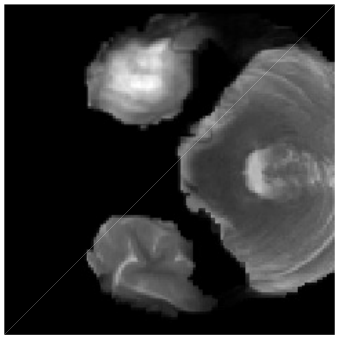
In order to assess the accuracy of the EPS technique and that of other performance metrics, a large sample dataset of realistic edge results and reference images is preferred. In order to create a large dataset of typical edge filter responses, an edge filter was first applied, with NMS and hysteresis thresholding to a set of 5 MRI volumes for which reference data was already available. The MRI volumes and reference images were obtained from the BRATS data set (Menze et al., 2015).

To increase the number of example images in the dataset, small sample regions were extracted from the reference data and the same corresponding sample location is also extracted from the filter result to form a 'sub image pair'. These sub-image pairs are small 2-D neighbourhoods of varying sizes, extracted from every edge point location in the reference data and paired with the same neighbourhood location in the filter result (Fig 4.12).

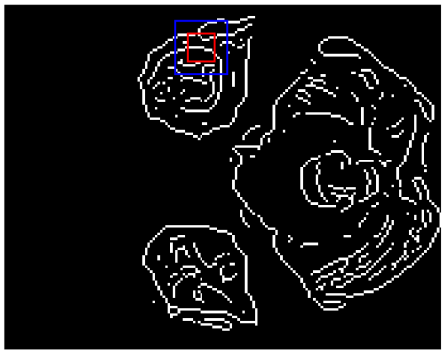
The sub-image sizes range from $11 \times 11$ and increase in odd increments through to $29 \times 29$. By varying the sub-image size, the complexity of performance measurement is altered,

more edge points are introduced as the size of the sub-image increases. This provides an indication of whether the complexity of the data affects the accuracy of the methods being evaluated. The sub-images pairs are required to be 2-D as opposed to 3-D, as the methods are to be compared against the Hungarian algorithm, which is only practical for 2-D performance evaluation.

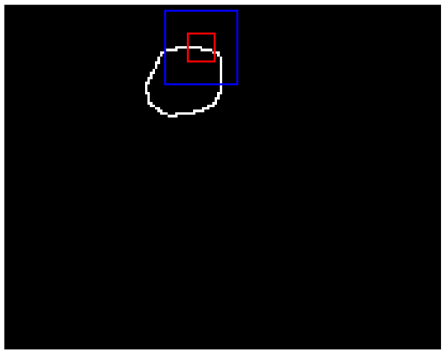
There are a total 12424 ideal points in the 5 MRI Volume reference images, therefore 12424 2-D sub-images pairs were created at each of the 10 aforementioned neighbourhood sizes. Using 10 different sub-image sizes creates a total of 124240 reference sub-images with corresponding filter result sub-images for the evaluation dataset.



(a) MRI layer



(b) Edge Filter Result

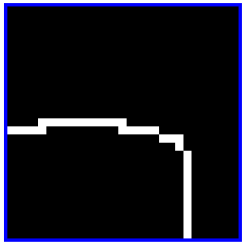


(c) Ground Truth (GT)

(d) Filter sub-image (11×11)

(e) GT sub-image (11×11)

(f) Filter sub-image (29×29)



(g) GT sub-image (29×29)

Figure 4.12: Examples of sub-image pairs extracted from MRI volume. Each sub-image is a small region extracted from the Edge Filter result and reference image. Sub-images were extracted for every position in the image with a positive reference point in its central index location. For a total of 12424 2-D sub-images pairs at each scale.

#### 4.5.1.2 Method

The evaluation of accuracy for correspondence matching, requires comparison against the Hungarian algorithm, as this approach always produces the optimal cost function $d_i$. The different approaches to correspondence matching are applied within a singular performance metric such that the only variable assessed is the cost function $d_i$. The performance metric selected is comparable to widely used Pratt Figure of Merit (Eq 4.4).

$$MetricScore = \frac{1}{\max\{I_D, I_i\}} \sum_{i=1}^{I_D} \frac{1}{1 + \alpha(d_i)^2} \tag{4.4}$$

In the evaluation four different approaches to correspondence matching are assessed to provide the $d_i$ distance metric in their own form. Firstly One to many correspondence matching is assessed relative to optimal one to one matching using the standard Pratt Figure of Merit. Three one to one correspondence matching approaches are then compared to the optimal Kuhn-Munkres Hungarian algorithm approach. These include the Closest Distance Match of Bowyer et al. (2001), the CSA assignment method of Goldberg and Kennedy (1995) and the novel Efficient Paring Strategy (EPS) presented in this chapter.

The performance measures are applied to each of the sub-image pairs and the scores are measured for accuracy by comparing against the optimal Hungarian algorithm using Pearson's pairwise correlation, the results of which are presented in table 4.1.

#### 4.5.1.3 Results

| Metric | 11×11 | 13×13 | 15×15 | 17×17 | 19×19 | 21×21 | 23×23 | 25×25 | 27×27 | 29×29 |
|---|---|---|---|---|---|---|---|---|---|---|
| PFOM | 0.80 | 0.87 | 0.91 | 0.93 | 0.95 | 0.96 | 0.96 | 0.97 | 0.97 | 0.97 |
| CDM | 0.86 | 0.91 | 0.94 | 0.95 | 0.96 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| CSA | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| EPS | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |

Table 4.1: Pearson Pairwise Correlation of performance metric results obtained through optimal Hungarian method calculation of $d_i$ with that of results obtained from sub-optimal correspondence matching methods. Including the standard Pratt Figure of Merit, in addition the following one to one correspondence matching techniques; Closest Distance Match, CSA Assignment and Efficient Paring Strategy calculation. 12424 example images were used at odd sub-image sizes from 11×11 to 29×29. Pvals for all results were 0

The results of the comparison show that the presented EPS is strongly correlated to the Hungarian algorithm solution over a range of different sub image sizes with a coefficient of 0.99. The CSA method is also strongly correlated to the Hungarian solution achieving

coefficient scores ranging from 0.99-1.00. This indicates that for optimal paring the CSA or EPS method are highly preferred over the other methods, offering greater accuracy, and therefore greater reliability when compared against the CDM and PFOM methods.

### 4.5.2 Analysis of Efficiency

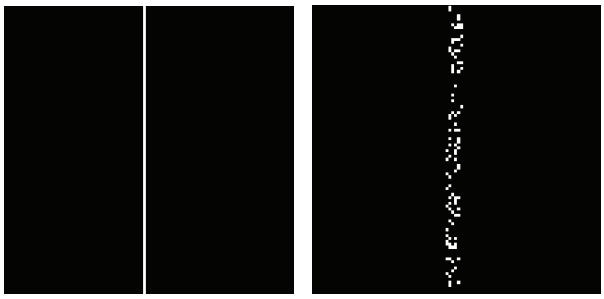
While the Hungarian algorithm always provides an optimal cost function for correspondence matching, it is very computationally expensive and not practical for large datasets or complex 3-D performance analysis. The alternate methods assessed show that CSA and EPS approaches both can achieve results very close to optimal. The following analysis evaluates the computational cost in terms of completion time in order to assess whether or not they are suitable for large data sets and 3-D data.

#### 4.5.2.1 Efficiency Test Data

In order to assess edge and surface detection algorithms, the performance metrics again needed to be applied to large datasets, and the computation time recorded. Since the complexity of the data affects the computational efficiency of the methods in question, the procedure for measuring the efficiency of the methods required running the performance measures on images on controlled levels of complexity.

In this analysis, the complexity of the data is increased in a linear fashion, this requires careful construction of the data such that the number of edge and surface points is controlled. Increasing the complexity of the data is achieved by increasing the number of potential correspondence matches in each 2-D and 3-D example case. This is achieved by increasing the size of each image, and increasing the number of edge and surface points linearly as the size of the images increases.

In the creation of the data, the number of edge points in the ideal image was made to precisely match the number of edge points in the result image, this is to ensure conditions which allow for precise one to one correspondence matching of all edge points in 2-D and surface points in 3-D. The edge and surface points in the test images were pseudo randomly generated and only existed within the tolerance zone ($T_{match}$) of an ideal edge. An example of a reference image and test image pair is presented in Fig. 4.13. Here $T_{match}$ was set to accommodate a $5 \times 5$ window around the ideal points, therefore all FP points are located within the tolerance governed by $T_{match}$ and the number of edge points in the reference is the same as the test image.



(a) Reference Example (b) Test Image Example

Figure 4.13: Example of test images and ground truth images for time analysis with 100 potential correspondences. Edge points are created within $T_{match}$ neighbourhood such that the number of edge points in both ground truth and test image are equal in order to allow for correspondence matching of all points.

An equivalent 3-D dataset was created in order to assess the efficiency of the correspondence matching approaches for surface detection evaluation. Here $T_{match}$ was set to accommodate a $5\times5\times5$ neighbourhood. Complexity was again increased by increasing the number of potential correspondence matches within data by increasing the image size and the number of surface points linearly.

#### 4.5.2.2 Method

Again the correspondence matching approaches are evaluated using a singular performance metric employing the different approaches to calculating $d_i$ (Eq 4.4). The performance metric is applied to the reference and test pairs, and the average calculation time recorded across 20 examples, before the complexity of the test data is increased. In the 2-D analysis, the number of edge correspondences ranges from 0 to 600, while for 3-D the number of surface correspondences ranges from 0 to 10000, the discrepancy between ranges is due to the fact that 3-D data will possess more surface points than an equivalent 2-D image contains edge points.

#### 4.5.2.3 Results

The CSA (Goldberg and Kennedy, 1995) and the EPS performance measures were compared first in 2-D against the Munkres (1957) Hungarian algorithm. The results are shown in Fig 4.14, here it can be seen that for 2-D performance measures the Munkres Hungarian completion time increases exponentially as the complexity of the data increases through broadening the number of potential correspondences. The Munkres Hungarian algorithm is shown to be significantly slower than the alternate accurate CSA and EPS methods.

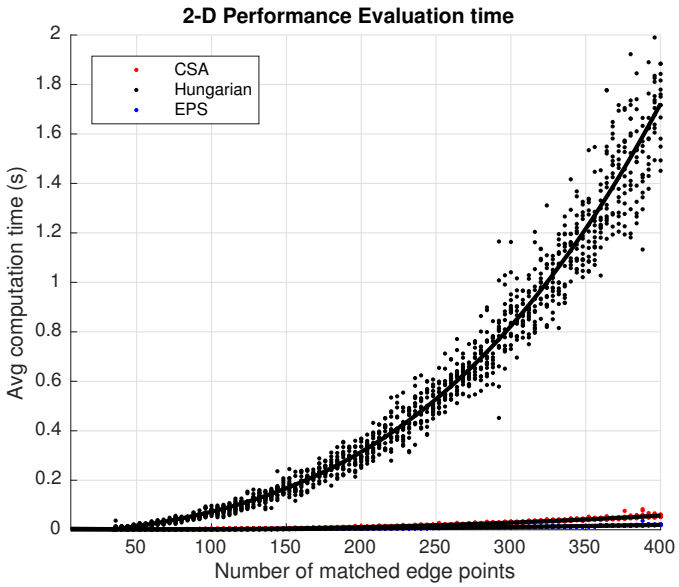



Figure 4.14: (Computational time analysis of 2-D performance measures, including Hungarian Kuhn, 1955, CSA Goldberg and Kennedy, 1995 and the EPS method.

However, it is not clear from this figure that the EPS method is efficient enough to be a viable replacement for the Pratt Figure of Merit or the Closest Distance match. Fig 4.15 shows a comparison of the EPS technique against the efficient sub optimal PFOM and CDM techniques, revealing that the EPS is able to perform the task in comparable time.

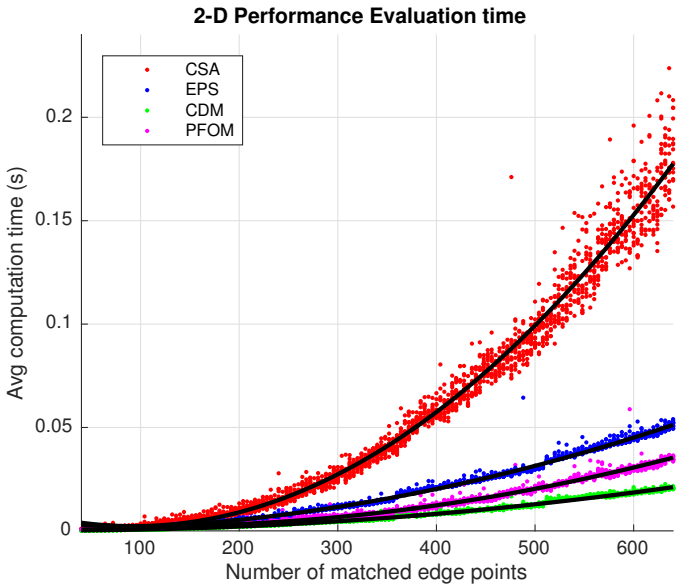
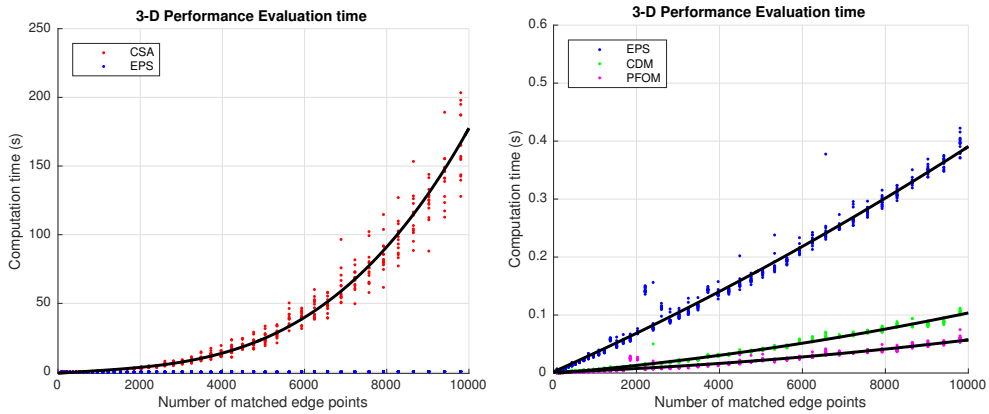



Figure 4.15: Computational time analysis of alternative performance measures, including CSA, EPS, CDM and PFOM.

While the CSA method could be considered suitable for 2-D data sets (Fig 4.15), for 3-D data it is shown to be significantly more computationally expensive when compared to the EPS method (Fig 4.16a). The additional computational complexity introduced by 3-D data and surface information, leads to a significant increase in completion time for the CSA method. However Fig 4.16b shows that in the context of 3-D, the time complexity of the problem remains linear for the EPS, CDM and PFOM methods. This means that the EPS is the only method which is both accurate and computationally efficient.




(a) 3-D CSA v EPS

(b) 3-D EPS v linear time methods

Figure 4.16: (a) Computational time analysis of optimal 3-D performance measures, CSA Goldberg and Kennedy, 1995 and the proposed EPS method. (b) Computational time analysis of fast sub-optimal 3-D performance measures, including EPS, CDM and PFOM. The results of a) show the EPS to be significantly more efficient than the Hungarian algorithm approach, while the results of b) show the EPS to be comparable to the sub-optimal CDM and PFOM approaches

### 4.5.3 Summary

The Efficient Pairing Strategy assisted metric offers increased accuracy over a number of classical performance metrics, notably the commonly applied one to many correspondence PFOM technique, and reliably presents an objective measure that more closely reflects the visual image results by adequately penalising fragmented edges and surfaces.

The EPS results are shown to be consistently accurate, with a 0.99 Pearson correlation against test assignment cases solved by the Hungarian algorithm, improving over the existing CDM and standard PFOM methods for correspondence matching (Table 4.1). Furthermore the EPS method was shown to provide results comparable to the Hungarian and CSA methods in terms of accuracy of one to one correspondence matching with significantly less computational cost.

Finally, it should be noted that the EPS in this form does not offer a general solution for the assignment problem. However, the EPS does provide a fast and accurate alternative to the Hungarian and CSA algorithms in the context of correspondence matching reference solutions with edge and surface detection results, finally allowing for practical fast one to one correspondence matching that is suitable for both large 2-D and 3-D data sets.

# Chapter 5

# Objective Analysis of 3-D Surface Detection

## 5.1 Introduction

Evaluation of 3-D statistical surface detection methods outside of this work has yet to be undertaken. In this chapter, the performance of the Vector Magnitude statistical surface detection method presented in Chapter 3, is evaluated across a range of experimental tests. The novel statistical approach is compared against two baseline traditional methods, the 3-D Canny surface detection algorithm, adapted from the groundbreaking 2-D Canny (1986) edge detection algorithm, and the optimal 3-D Steerable filter method of Aguet et al. (2005) for detecting surface features.

This chapter provides an objective analysis across a range of evaluations that utilise synthetic datasets with ground truth reference data. Objective assessments require a performance metric which grades the quality of detection with a numerical value, thus allowing for repeatability in the evaluation. In order to undertake a multi-variable analysis, a large number of performance metric calculations need to be made, therefore the Efficient Paring Strategy for one to one correspondence matching presented in Chapter 4 is utilised to allow for a comprehensive analysis within a practical time frame.

The design of the evaluations seek to address the common issue of a lack of correlation between performance results obtained with synthetic datasets and that of performance in a real application. The findings from these evaluations can be utilised to inform appropriate application of the techniques in other contexts. Observed are the effects of the statistical test selection (described in Chapter 3), the effects of scale, and the preprocessing Gaussian filter stage in the case of the baseline techniques.

A further contribution from this work is that the methodology adopted here can also serve as a framework for the evaluation of novel surface detection techniques developed in the future, allowing for a reliable comparison with newly developed methods.

## 5.2 Evaluation Methodology

As discussed in Chapter 4 there are various advantages and disadvantages to objective analysis. The primary advantage is that an objective analysis provides a repeatable result which allows for direct comparison across different bodies of work. The main shortfall however is that objective analysis requires “ground truth" or “reference" data in order to generate a reliable metric. This problem is two-fold, for evaluation with synthetically created data, determination of the ground truth solution is trivial and easily controlled, however the reliability of the synthetic data to simulate real application data is a non trivial problem due to the complexity of real imagery. On the other hand if real application data is used, obtaining reliable reference solutions is non-trivial since subjective analysis of the imagery is required.

In order to objectively analyse the surface detection methods in a reliable fashion, careful consideration was required in the design of the evaluations. This chapter provides an objective analysis of surface detection with evaluations that utilise synthetic data for repeatability in the analysis. Chapter 6 presents a subjective approach using a real application case to assess whether the findings of the objective analysis give a reliable indication of performance in real application data.

Bowyer et al. (2001) specified three conditions for why synthetic data does not always translate accurately to real application performance, this chapter addresses the following three conditions in the design of the methodology.

- The simplicity of the data.
- The lack of topological considerations of the interface, namely lack of curvature.
- The number of different types of interfaces with different relative strengths or scales contained within the data.

### 5.2.1 Evaluation Aims

To address the correlation problem stated by Bowyer et al. (2001), this analysis is broken down into 3 separate evaluations, in order to assess the independent effects of the aforementioned issues. The evaluations are as follows:

1. Measurement of the resolving power of statistical surface detection on interfaces between regions of different region profiles.

2. Measurement of effects of surface topology on the detection performance.
3. Measurement of performance with data that possesses multiple distinct region profiles with different strength interfaces.

#### 5.2.1.1 Evaluation 1

Evaluation 1 is to act as a control, measuring the resolving power of surface detection methods independent of the effects of topology and relative surface strength by using a single uniform interface between two regions. The resolving power is the ability of a filter to resolve an edge or surface between two or more regions of varying similarity. Therefore resolving power is a measure of the signal to noise ratio of the surface to is surroundings and the rest of the image. A high signal to noise ratio of correct surface points to non surface points indicates good resolving power. Maximising resolving power requires the filter to resolve the boundary with a high value output with respect to background noise and clutter. For this evaluation, the surface interface topology is structurally uniform in the spatial domain (flat) and invariant between 2-D and 3-D. Thus, the topology of the surface in each layer is unchanging, as topology features influence filter performance.

#### 5.2.1.2 Evaluation 2

Evaluation 2 explores the impact of interface topology. While evaluation 1 uses scale invariant image volumes with a surface structure that is uniform and invariant through the layers, this is not typical of real imagery. In real imagery, interfaces exist with a wide range of structures with varying degrees of complexity (Milan et al., 1993). The topology of a surface introduces intricacies which affect the ability of a filter to resolve a surface accurately, independent of the separation in terms of the statistical differential between region profiles. The variations in surface topology can exist at varying scales, where details exist which are smaller in scale than the neighbourhood mask, these cannot always be resolved, leading to a smooth surface with some loss of the finer details Witkin (1984). In 2-D imagery, the topology can affect the performance of the filter, notably the Canny edge detector can perform sub optimally in the vicinity of corners and junctions (Rothwell et al., 1995; Ding and Goshtasby, 2001). This is an inherent issue of measuring the intensity gradient where boundaries meet, and remains true for 3-D imagery since the direction of the gradient at these locations is unclear. Non maximum suppression is also negatively affected in these locations due to the ambiguity of the gradient direction in these positions. While evaluation 1 utilises the statistical properties of MRI data to establish the interface properties, the topological considerations included in the evaluation were established through precedent and are based on the criticisms made by Bowyer et al. (2001), these criticisms call for curvature

to be assessed, as well as junction and corner interfaces in addition to the commonly used test interfaces which typically only contain uniform flat structures.

Evaluation 2 therefore evaluates the effects of surface topology on filter performance, and introduces four different scale variant surface topology image volumes each using the same region profile distributions from evaluation one. 2-D synthetic shape images are often used for evaluation methods, for example square edge interfaces (Lim, 2006) and circles and triangles (Setayesh et al., 2011). While 3-D surface detection methods can be tested using 3-D structures such as spherical synthetic interfaces such as in the work of Meinhardt et al. (2008). Here four surface topologies are used in the assessment, a cuboid structure for corner and junction features, a spherical structure for curved interfaces, and two 'staircase' structures with different corner features at different step scales. Each topology image is repeated four times with the same region profiles used in Evaluation 1, and each volume has four versions for a Monte-Carlo style of analysis.

#### 5.2.1.3 Evaluation 3

Evaluation 3 explores the impact on performance when multiple different interfaces exist within a single image. In real application data, such as CT or MR, seldom is ever only one interface present in the image. The complexity of real images mean there are several different regions within an image with their own unique properties. Sometimes these properties change abruptly enough to indicate the presence of an interface between distinct regions. The abruptness of the change is often more significant in some parts of an image than others. Often the types of image properties which change can differ, such as a strong intensity shift from one region to another in one part of an image, or a strong texture shift in another part of an image. The consequence of this are that there are interfaces within an image that are resolved to different degrees by different detection methods. A disparity arises in the strength of the interface, this creates a potential problem for defining a threshold for what constitutes a surface and what does not.

Defining parameters which are optimal for strong surfaces responses will typically result in weaker surfaces being discarded. Since these algorithms are techniques which do not utilise semantic information, the discarded interfaces could be of importance. Conversely, optimising parameters which resolve the weaker interfaces can result in an excess of spurious responses which will clutter the image and manifests as increased noise in the output and thereby decreasing the signal to noise ratio of the detected surfaces.

Placing interfaces of different relative strengths in the same image volume will assess local normalisation aspects of the statistical filters, which contribute to the overall "robustness" of the filter with respect to sensitivity and thresholding. This evaluation more closely resembles real imagery since it is more typical for real images to possess more than one region. However, as a synthetic image, it is highly simplified in comparison to real image structure and texture properties, which maintains the ability to perform a quantitative analysis of filter performance.

## 5.3 Evaluation Dataset

In the subsequent Chapter 6 a domain specific case study is presented which assesses different surface detection responses to particular important interfaces that are present in MRI datasets. The interfaces of interest in the case study are of various brain tumour and cyst boundaries in different Magnetic Resonance Imaging modalities. Specifically T1-weighted, T2-weighted and contrast enhanced T1-weighted images of paediatric patients with Pilocytic Astrocytoma (PA) tumour pathologies. However, first the general characteristics of the surface detection methods need to be objectively established, this requires synthetic data. A criticism of objective analysis with synthetic data is that the data is not representative of real imagery, thus making synthetic data unreliable predictors of success for real applications. For an objective analysis to predict the success of surface detection in a real context, the datasets should be general enough that characteristics of the results can be inferred for a wide range of applications, but realistic enough that the results bear some resemblance to an actual real application. To facilitate this, the creation of the data needs to be a compromise between realism and universality (Fernandez et al., 2015). Outlined prior are 3 evaluations which have been designed to improve correlation between objective performance analysis using synthetic data and performance on real data. To facilitate the aims of each evaluation, careful consideration is required in the generation of the datasets.

For creation of the synthetic datasets, some general image features and properties were extracted from the case study data presented in chapter 6 to inform the development of the synthetic datasets. Fig 5.1 illustrates regions in close proximity to an interface from which the region properties were analysed. Two statistical features were extracted from the regions. Firstly the arithmetic mean as defined in eq 5.1, where $X$ is the sample region comprised of voxel intensity values ($x$), $n$ is the sample size and $\bar{x}$ is the arithmetic mean. Secondly the variance ($\sigma$) of the region was determined, as defined in eq 5.2, these values were normalised to fall within an 8bit 0-255 range. Further complex texture properties were not incorporated as these could result in test images overly suited to a narrow purpose. This is inline with the data creation protocol used

by Williams et al. (2014), where properties of histological images were used to inform the creation of 2-D datasets.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

Arithmetic Mean ($\bar{x}$): Where $x_i$ is the intensity value of the voxels in the region, and $n$ is the size of the sample

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.2}$$

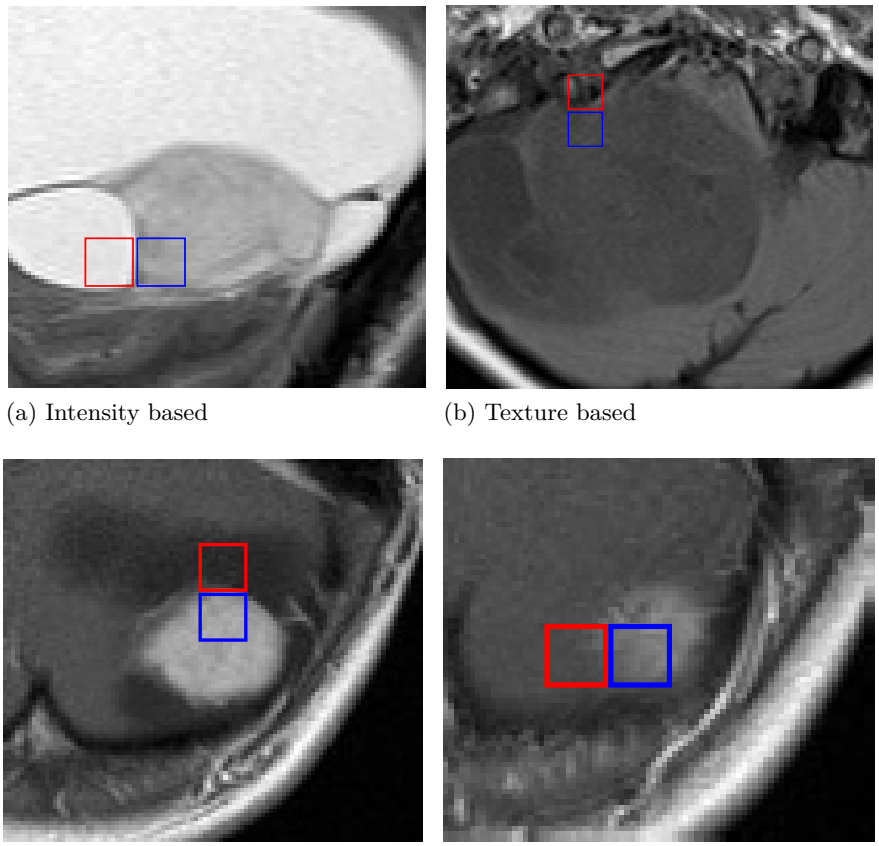
Variance ($\sigma$): Where $x_i$ is the intensity value of the voxels in the region, and $n$ is the size of the sample and $\bar{x}$ is the Arithmetic Mean of the region

### 5.3.1 Evaluation 1 Data

For 2-D edge detection evaluation, Williams et al. (2014) also utilised synthetic data in the analysis, and set a precedence for how synthetic data should be generated. It was established that multiple versions of each test image type should be introduced, this 'Monte-Carlo' style of approach ensures that the results from a particular test are consistent across multiple instances of the same test image type. This rules out abnormal or peculiar responses which are not representative in a broader context. The same general approach is utilised here for 3-D data, each test image volume is generated 4 separate times using pseudo-randomly generated values, each with the same general region properties, but are unique in their own right. For this 4 control interface types were utilised, consisting of a predominantly intensity based interface derived from Fig 5.1a (IntFl), a predominantly stochastic texture based interface derived from Fig 5.1b (TexFl), and two combination interfaces which utilises different mean and variance properties derived from Fig 5.1c (Com1Fl) and Fig 5.1d (Com2Fl).

IntFl (Fig. 5.2a): This test volume simulates a typical step-change in intensity with a flat interface, 1st derivative methods of edge and surface detection are designed to be optimal on this kind of boundary. Thus, indicating whether statistical methods of surface detection can offer comparative results where traditional derivative based methods of surface detection methods typically excel (Williams et al., 2014).

TexFl (Fig. 5.2b): This test volume presents a change in the variance property of the regions, this aims to simulate the most simple kind of stochastic texture interface with a single uniform flat interface. These types of boundaries are present in highly textured and noisy images. 1st derivative methods are not specifically suited to resolve this



(a) Intensity based

(b) Texture based

(c) Combination with strong intensity component

(d) Combination with weak intensity component

Figure 5.1: Types of interfaces in real imagery. The following measurements are properties of the regions of interest around the interface. Intensity based ($\bar{x}$ 1420-1010. $\sigma$ 94-118). Texture based ($\bar{x}$ 326-392 $\sigma$ 122-30). Combination 1 ($\bar{x}$ 230-751 $\sigma$ 71-120). Combination 2 ($\bar{x}$ 387-545 $\sigma$ 37-75)

kind of interface and as such, typically produce less than optimal results (Williams et al., 2014). This interface type examines which statistical methods, if any, can offer improvements where derivative based methods typically fail.

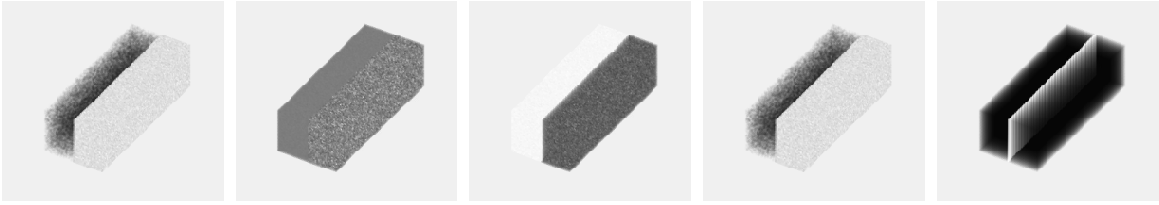
Com1Fl (Fig. 5.2c): This test volume presents an interface combining both intensity and texture which is more typical of surfaces in real imagery where simple textures are present. Com2Fl (Fig. 5.2d): This test volume also presents an interface combining intensity and texture differences. However with a smaller intensity change component. Region properties are presented in Table. 5.1.

FlGT (Fig. 5.2e) Is the reference image generated which provides the exact location of the surface, and is used as a basis for comparison in the performance metric.

Each region is created pseudo-randomly with controlled statistical distributions to ensure there is no repeating pattern or noise. Each image volume is $100\times40\times30$ voxels in size and each region is $100\times20\times30$ positioned adjacent to each other. The interfaces properties are described in Table 5.1 and are illustrated in Fig 5.2 and 5.3.

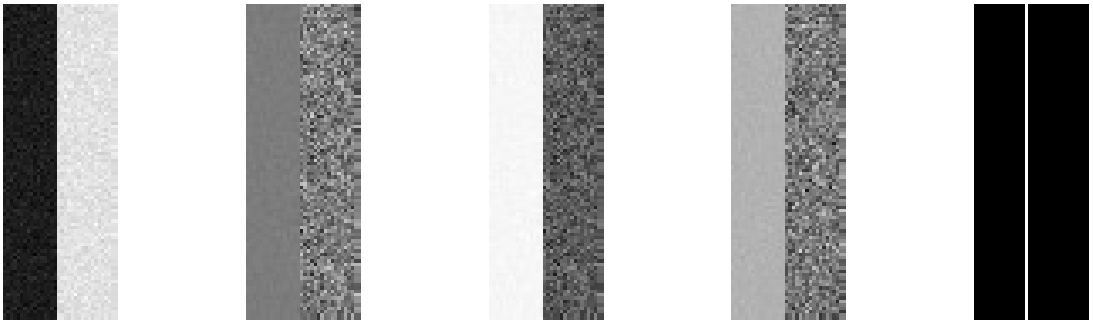
| Image | Mean | | Variance | | Topology | # of Regions | Figure | Reference |
|---|---|---|---|---|---|---|---|---|
| IntFl | 6 | 249 | 9 | 9 | Flat | 2 | Fig 5.2a | FlGT |
| TexFl | 125 | 125 | 1 | 11 | Flat | 2 | Fig 5.2b | Fig 5.2e |
| Com1Fl | 125 | 50 | 1 | 11 | Flat | 2 | Fig 5.2c | |
| Com2Fl | 75 | 54 | 1 | 11 | Flat | 2 | Fig 5.2d | |

Table 5.1: Table containing information about the test images used in Evaluation 1. The table provides the naming convention and region statistics of test interfaces. The topological structure of the test image, The number of regions in the test image is stated. In addition the figure number of the each test image volume is provided along with the corresponding ground truth reference image name



(a) IntFl (b) TexFl (c) Com1Fl (d) Com2Fl (e) FlGT

Figure 5.2: Evaluation 1 dataset



(a) IntFl (b) TexFl (c) Com1Fl (d) Com2Fl (e) FlGT

Figure 5.3: Evaluation 1 images cross section

### 5.3.2 Evaluation 2 Data

Evaluation 2 assesses the impact of interface topology. 4 unique topological interface structures are utilised, each of these topological structures incorporate the previously described region properties from evaluation 1 for a dataset containing 16 unique image volumes. The topologies are as follows:

#### 5.3.2.1 Topology 1:Cuboid (Cu)

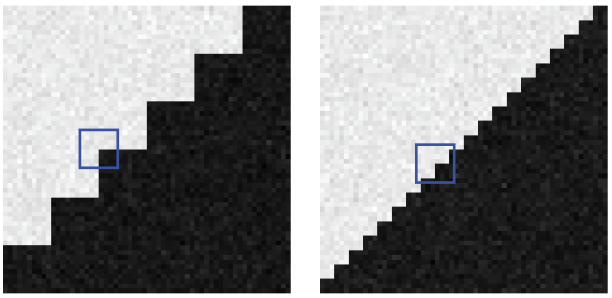
The cuboid structure (Fig. 5.5a-e) introduces the effects of corners. In 2-D edge detection methods, corners and junction are often not correctly resolved (Rothwell et al., 1995; Ding and Goshtasby, 2001), this property of the edge detection filters continues to apply with surface detection methods. Corners are locations of a structure where the edge or surface gradient points in different directions, the gradient directions of the cuboid corners are orthogonal, the opposing nature of the gradient directions often leads to a surface measure which is relatively small. Low value points are considered non-surface points, as a result the edges along the perimeter of the surface, as well as corners, may not be resolved by the detection method. The cuboid structure is not invariant throughout the volume, and it features flat surfaces, including those which exist only in the z-plane, these regions of the surface interface will only be resolved by filters that process the voxels between layers, thus 3-D specific filters are required (Fig. 2.19). The image volumes are constructed with a cuboid structure of size $60\times30\times30$ positioned centrally in a image volume of size $80\times50\times50$. The exterior of the cuboid interface was generated with the mean and variance properties of region 1 in Table 5.2, while the interior of the cuboid was constructed with region 2 properties.

#### 5.3.2.2 Topology 2: Spherical (Sp)

The second image type is a spherical interface (Fig. 5.5f-j). The sphere was produced with a high degree of curvature in relation to the neighbourhood mask size. The high degree of curvature therefore will be strongly affected by the scale parameter, penalising larger mask sizes, thus the trade-off between resolving power and accuracy can be analysed. The image volumes are constructed with a sphere of radii 30 voxels, and is positioned centrally in an image volume of size $75 \times 75 \times 75$. The exterior of the sphere was generated with the mean and variance properties of region 1 in Table 5.2, while the interior of the sphere was constructed with region 2 properties.

#### 5.3.2.3 Topology 3 and 4: Coarse Staircase(CSt) and Fine Staircase (FSt)

The two remaining interface types both possess a 'staircase' structure (Fig. 5.5k-o, for large steps and Fig 5.5p-t with small steps). The purpose of using steps with different sizes relative to the neighbourhood mask is to assess the affect topology details smaller than the neighbourhood mask have on performance. A staircase structure also possesses a regular occurrence of corners, which should further highlight any corner feature problems. Both the staircase volumes are constructed with a size of $60 \times 60 \times 30$ voxels, the 'staircase' structure runs diagonally throughout the volume. The large stepped volume has steps that are 10 voxels tall, 10 voxels wide, and each step runs through all 30 layers of the image volume. The small stepped volume has step dimensions of $3 \times 3 \times$ 30. Resulting in an interface with details small than the smallest neighbourhood window use by the surface detection methods (Fig 5.4). Each interface volume has two regions, and the statistical properties of these regions are shown in table 5.2. This test will identify the effect on resolving the boundary at different neighbourhood mask sizes or with different amounts of smoothing in the case of Canny and Steerable.

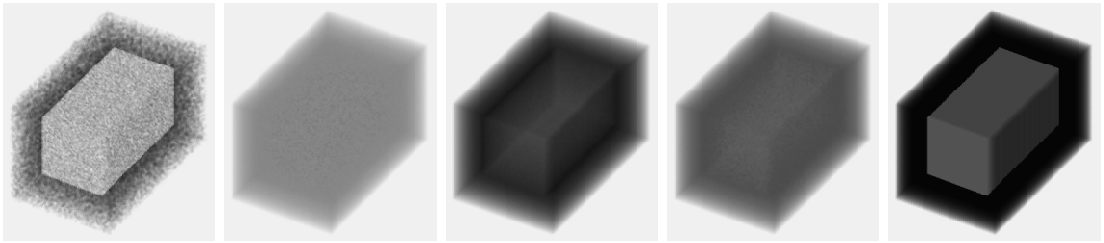


(a) Large Step (b) Small Step

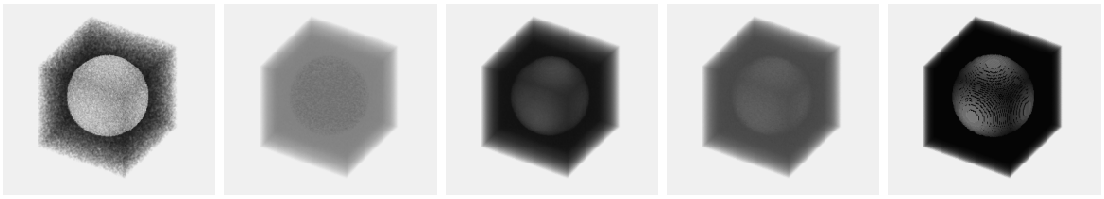
Figure 5.4: 2D representation of scale using an $11 \times 11 \times 11$ neighbourhood mask in relation to interface step size

| Image | Mean | | Variance | | Topology | # of Regions | Figure | Reference |
|---|---|---|---|---|---|---|---|---|
| IntCu | 6 | 249 | 9 | 9 | Cuboid | 2 | Fig 5.5a | CuGT |
| TexCu | 125 | 125 | 1 | 11 | Cuboid | 2 | Fig 5.5b | Fig 5.5e |
| Com1Cu | 125 | 50 | 1 | 11 | Cuboid | 2 | Fig 5.5c | |
| Com2Cu | 75 | 54 | 1 | 11 | Cuboid | 2 | Fig 5.5d | |
| IntSp | 6 | 249 | 9 | 9 | Spherical | 2 | Fig 5.5f | SpGT |
| TexSp | 125 | 125 | 1 | 11 | Spherical | 2 | Fig 5.5g | Fig 5.5j |
| Com1Sp | 125 | 50 | 1 | 11 | Spherical | 2 | Fig 5.5h | |
| Com2Sp | 75 | 54 | 1 | 11 | Spherical | 2 | Fig 5.5i | |
| IntCSt | 6 | 249 | 9 | 9 | Coarse Staircase | 2 | Fig 5.5k | CStGT |
| TexCSt | 125 | 125 | 1 | 11 | Coarse Staircase | 2 | Fig 5.5l | Fig 5.5o |
| Com1CSt | 125 | 50 | 1 | 11 | Coarse Staircase | 2 | Fig 5.5m | |
| Com2CSt | 75 | 54 | 1 | 11 | Coarse Staircase | 2 | Fig 5.5n | |
| IntFSt | 6 | 249 | 9 | 9 | Fine Staircase | 2 | Fig 5.5p | FStGT |
| TexFSt | 125 | 125 | 1 | 11 | Fine Staircase | 2 | Fig 5.5q | Fig 5.5t |
| Com1FSt | 125 | 50 | 1 | 11 | Fine Staircase | 2 | Fig 5.5r | |
| Com2FSt | 75 | 54 | 1 | 11 | Fine Staircase | 2 | Fig 5.5s | |

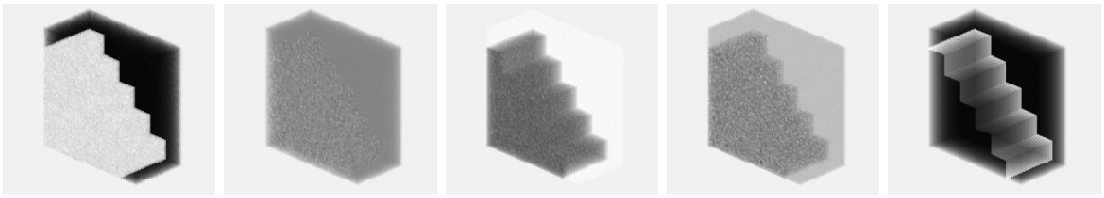
Table 5.2: Table containing information about the test images used in Evaluation 2. The table provides the naming convention and region statistics of test interfaces. The topological structure of the test image, The number of regions in the test image is stated. In addition the figure number of the each test image volume is provided along with the corresponding ground truth reference image name



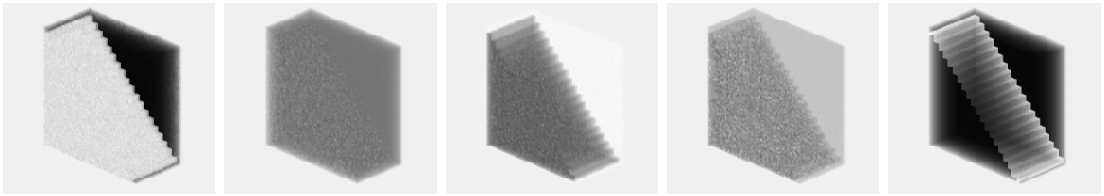
(a) IntCu (b) TexCu (c) Com1Cu (d) Com2Cu (e) CuGT



(f) IntSp (g) TexSp (h) Com1Sp (i) Com2Sp (j) SpGT

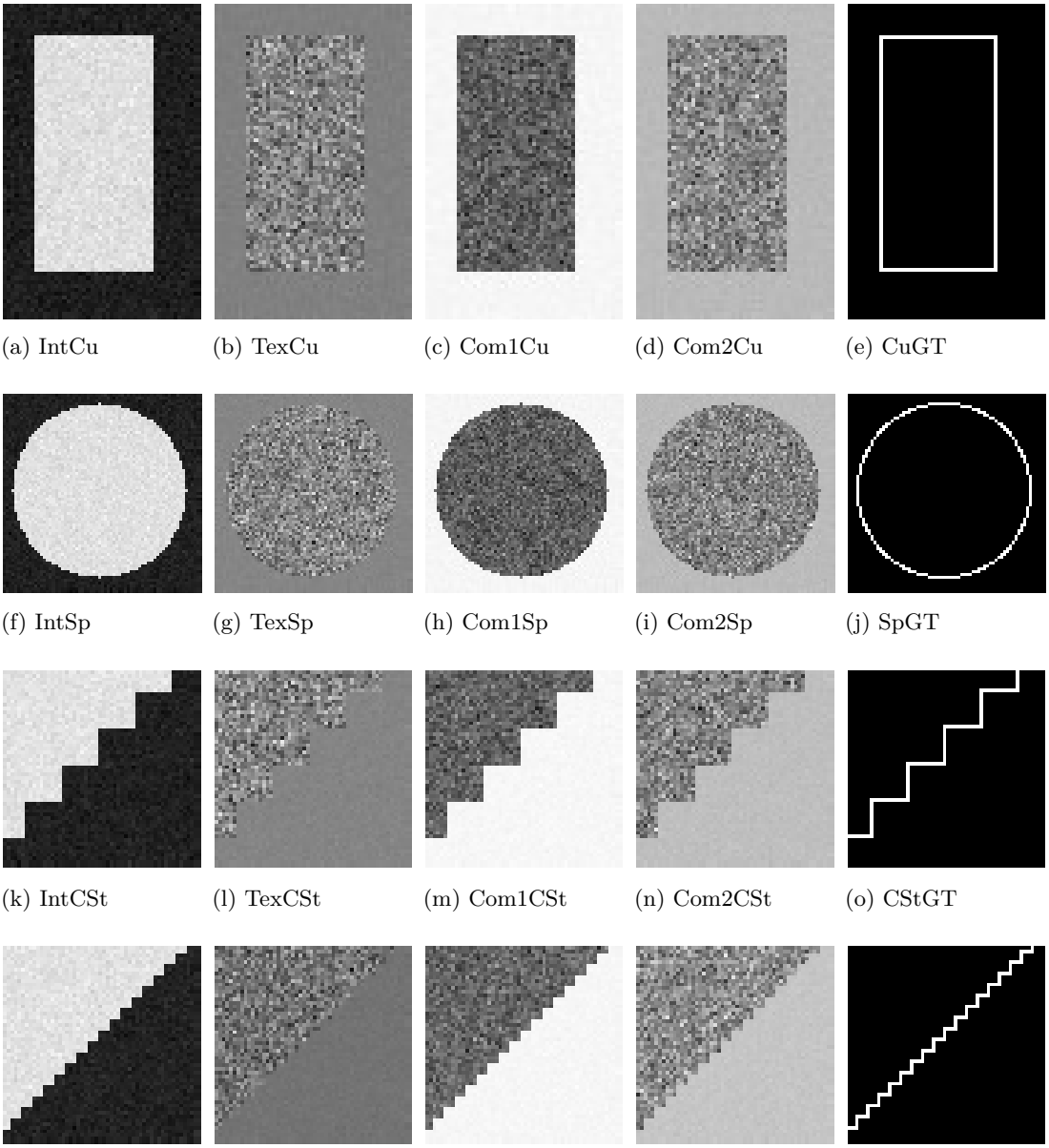


(k) IntCSt (l) TexCSt (m) Com1CSt (n) Com2CSt (o) CStGT



(p) IntFSt (q) TexFSt (r) Com1FSt (s) Com2FSt (t) FStGT

Figure 5.5: Evaluation 2 test image volumes with different interface topologies and region properties



(a) IntCu (b) TexCu (c) Com1Cu (d) Com2Cu (e) CuGT

(f) IntSp (g) TexSp (h) Com1Sp (i) Com2Sp (j) SpGT

(k) IntCSt (l) TexCSt (m) Com1CSt (n) Com2CSt (o) CStGT

(p) IntFSt (q) TexFSt (r) Com1FSt (s) Com2FSt (t) FStGT

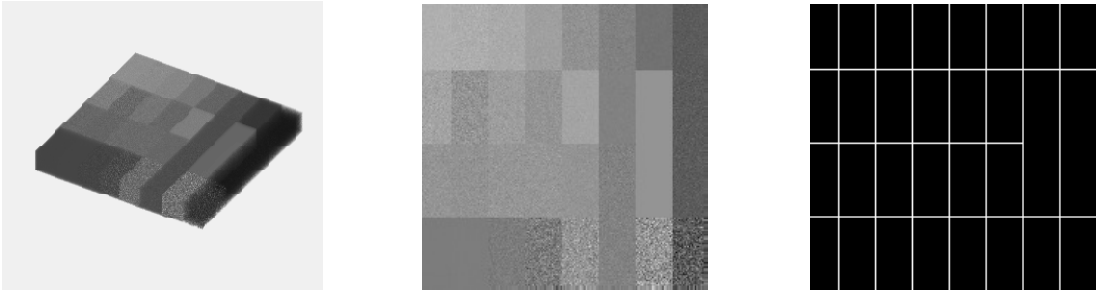
Figure 5.6: Evaluation 2: 2-D cross sections of test image volumes for additional clarity on internal structure.

### 5.3.3 Evaluation 3: Test Images

One of the issues outlined by Bowyer et al. (2001) are the limited number of different types of interfaces with different relative strengths or scales typically contained within the data. In order to address this problem, this evaluation consists of two image volume types labelled as MultiFlat and MultiCurve, which utilises multiple different region profiles within the same image volume.

#### 5.3.3.1 MultiFlat

The MultiFlat image contains 23 different region profiles, the properties of which are presented in Table. 5.3 and the arrangement of regions provide 33 unique flat interfaces in total. Additional region profile properties were extracted using the same process and dataset illustrated in Fig 5.1. However, the proximity of one region to another in the synthetic volume is not based upon their proximity in the real data.



(a) 3-D Multiflat

(b) 2-D Cross sectional view

(c) 2-D Reference (MultiFlatGT)

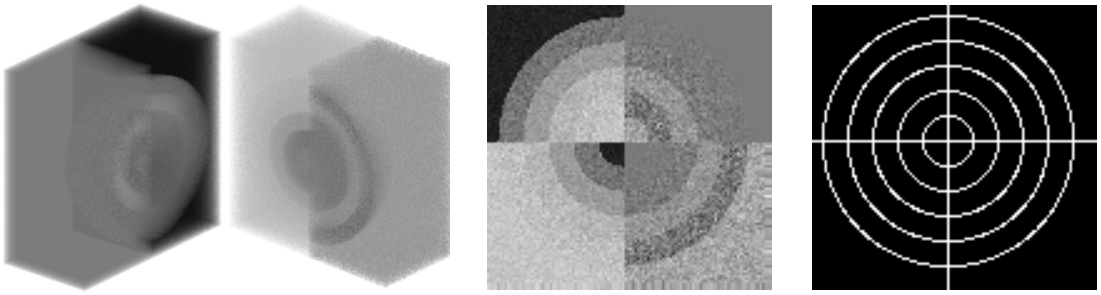
Figure 5.7: 3-D MultiFlat image volume containing multiple image regions.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| mean | 250 | 245 | 234 | 211 | 175 | 134 | 93 | 40 |
| variance | 8 | 12 | 10 | 8 | 12 | 10 | 5 | 10 |
| mean | 220 | 180 | 203 | 180 | 220 | 135 | 180 | 40 |
| variance | 17 | 24 | 14 | 14 | 7 | 5 | 0 | 15 |
| mean | 181 | 176 | 185 | 186 | 180 | 135 | 180 | 40 |
| variance | 14 | 14 | 15 | 15 | 10 | 15 | 0 | 15 |
| mean | 120 | 122 | 120 | 120 | 180 | 134 | 181 | 39 |
| variance | 1 | 1 | 15 | 40 | 40 | 15 | 60 | 60 |

Table 5.3: Region properties for for MultiFlat image volume. Cells represent spatial locations of regions from top left to bottom right.

#### 5.3.3.2 MultiCurve

Here the volume is comprised of a layered sphere divided up into four quadrants creating a total of 24 regions. Since the interfaces in this image are different in size, 12 unique region properties are used and are repeated in reversed order in an adjacent quadrant. By repeating interfaces the size discrepancy of particular interface types is reduced. This minimises, but does not eliminate, the biasing of the results toward a particular interface type. The regions properties used in for the MultiCurve volume are presented in Table 5.4.



(a) 3-D (b) 2-D (c) 2-D Reference (MultiCurveGT)

Figure 5.8: 3-D MultiCurve Image Volume. Containing multiple interfaces with curved topology, the 3-D volume visually separated into two halves in order to reveal internal structure.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| mean | 250 | 245 | 234 | 211 | 175 | 93 | Upper left quadrant |
| variance | 8 | 12 | 10 | 8 | 12 | 5 | |
| mean | 93 | 175 | 211 | 234 | 245 | 250 | Lower left quadrant |
| variance | 5 | 12 | 8 | 10 | 12 | 8 | |
| mean | 220 | 181 | 203 | 180 | 180 | 179 | Upper right quadrant |
| variance | 17 | 24 | 14 | 14 | 7 | 0 | |
| mean | 179 | 180 | 180 | 203 | 181 | 220 | Lower right quadrant |
| variance | 0 | 7 | 14 | 14 | 24 | 17 | |

Table 5.4: Region properties for MultiCurve image volume. Each row in the table represents the properties of the regions in each quadrant. Columns from left to right indicate from central to outer layers

## 5.4 Testing Methodology

A basic overview of each evaluation is as followed

**Section 5.3** Generation of synthetic test image volumes with controlled interfaces

**Section 5.4** Application of statistical surface detection filters, 3-D canny and Steerable filters, at different scales

**Section 5.4.2** Post processing preparation for quantitative analysis including linear normalisation, 3-D NMS and hysteresis thresholding

**Section 5.4.2** Apply objective performance measure.

Visual inspection of the 3-D surface maps to identify strengths and weaknesses and to identify localised problems

### 5.4.1 Filter Parameters

In the work of Williams et al. (2014), 2-D Statistical edge detection filters were evaluated across a range of filter kernel sizes, from 5×5 through to 19×19 in odd incremental steps. Comparisons were made against the baseline 2-D Canny edge detection method at different scales ($\sigma$1-4 of the Gaussian Filter). Due to the computational cost of processing data with 3-dimensions, a reduced range was applied. In this work 3-D Statistical surface detection filters are applied through a range of scales from 5×5×5 to 11×11×11 in odd incremental steps. However, due to similarity of the results across scales, for clarity only the 5×5×5 results are presented in chapters 5 and 6. In chapter 7 the characteristics of each filter method are explored at different scales, there the aspect of scale is assessed with the MultiFlat image type from Evaluation 3.

The 3-D Canny detection and 3-D steerable filters were applied to each volume using a scale parameter ranging from $\sigma$1-4. In the work of Williams et al. (2014) no significant benefit was observed for $\sigma$ values beyond this range for the Canny method, and therefore the upper limit of $\sigma$ =4 was maintained. In each evaluation, 8 different statistical filters were evaluated, 4 parametric tests, the Difference of boxes ($DoB$), Log-Likelihood ratio ($L$), Fisher test ($F$) and Student's $t$ test ($t$-test). Also evaluated are 4 non parametric tests, the local-$\chi^2$ test, Kolmogorov-Smirnov test($KS$), Mann-Whitney test ($u$-test) and Robust Rank Order test ($RRO$). Table 5.5 present an overview of the different evaluations, the test images were used, the filters applied and the scales used.

| Aims | Images | 3-D Filters | Normalised Thresholds | Performance measures |
|---|---|---|---|---|
| **Evaluation 1** To evaluate filter resolving power | IntFl<br>TexFl<br>Com1Fl<br>Com2Fl | Canny<br>Steerable<br>*DoB*<br>*t*-test<br>*F*<br>*L*<br>$\chi^2$<br>*KS*<br>*u*-test<br>*RRO* | Hysteresis<br>UT 1-99<br>LT 40%<br>increment of 1 | F1-score using EPS |
| **Evaluation 2** To evaluate impact of interface topology | IntCu<br>TexCu<br>Com1Cu<br>Com2Cu<br>IntSp<br>TexSp<br>Com1Sp<br>Com2Sp<br>IntCSt<br>TexCSt<br>Com1CSt<br>Com2CSt<br>IntFSt<br>TexFSt<br>Com1FSt<br>Com2FSt | | | |
| **Evaluation 3** To evaluate performance with multiple region profiles | MultiFlat<br>MultiCurv | | | |

Table 5.5: Evaluation 1-3 Test Parameters

### 5.4.2 Quantitative Evaluation

In order to apply an objective performance algorithm to assess the surface detection methods, voxels within the output image surface map first need to be classified into surface points and non surface points. This is achieved in a 3 stage process, normalisation of the output such that all voxel values are 32bit double precision floating point numbers scaled to the range 0-1. Application of 3-D Non-maximum suppression as described in Chapter 3. 3-D Hysteresis thresholding is then applied across a range of 100 discrete thresholds. The Upper threshold (UT) values ascend from 0 to 1 in increments of 0.01. Canny (1986) recommends a threshold ratio between 3:1 and 2:1 between upper and lower thresholds, this precedent was followed and the standard default lower threshold

to be a value of 40% of the UT was implemented. This generates 100 separate images for analysis for each test volume.

In an analysis of performance metrics used for evaluating edge detection methods, F-measure (F1-score) was shown to be the most reliable (Pont-Tuset and Marques, 2016). An optimal measure for 3-D surface detection has yet to be established, however characteristics of the F-measure also apply to 3-D. The F-measure is the harmonic mean of precision and recall and therefore first requires computation of precision and recall statistics. F-measure (Eq 5.3) is defined as:

$$F_1 = \left( \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5.3}$$

In order to accurately measure precision and recall, and therefore calculate the F-measure, one to one correspondence matching should be undertaken to get a more contextually applicable count of TP, FP, TN and FNs. The EPS technique discussed in detail in chapter 4 is embedded in the F1 score to provide an accurate metric calculation in each evaluation. F-measure is applied to each threshold output image. The F-measure scores are then plotted across the threshold (upper) range of 0-1, such as in Fig 5.13. A metric score value of 1 is indicates an accurate detection with no missed surfaces or any spurious responses. A score of 1 across all threshold ranges would be considered an ideal optimal filter response.

## 5.5 Evaluation 1: Results.

### 5.5.1 Evaluation 1: Aim

This test is to determine the resolving power of the filters on a controlled single interface between two regions of defined distributions with a uniform topology. Due to the simplified surface structure in comparison to the latter evaluations, and real imagery, these volumes can provide a controlled analysis of whether the detection method is able to successfully resolve the type of interface.

### 5.5.2 Evaluation 1: Visual Results

Presented in Figures 5.9,5.10,5.11 and5.12. are the surface map outputs from the 8 statistical surface detection methods and 2 baseline methods (3-D Canny and 3-D Steerable). Presented here are the 3-D results of the detection methods. For visual representation, the surface map outputs were normalised using a linear method to scale the intensity values between 0 and 255 and each of the surface maps are presented with

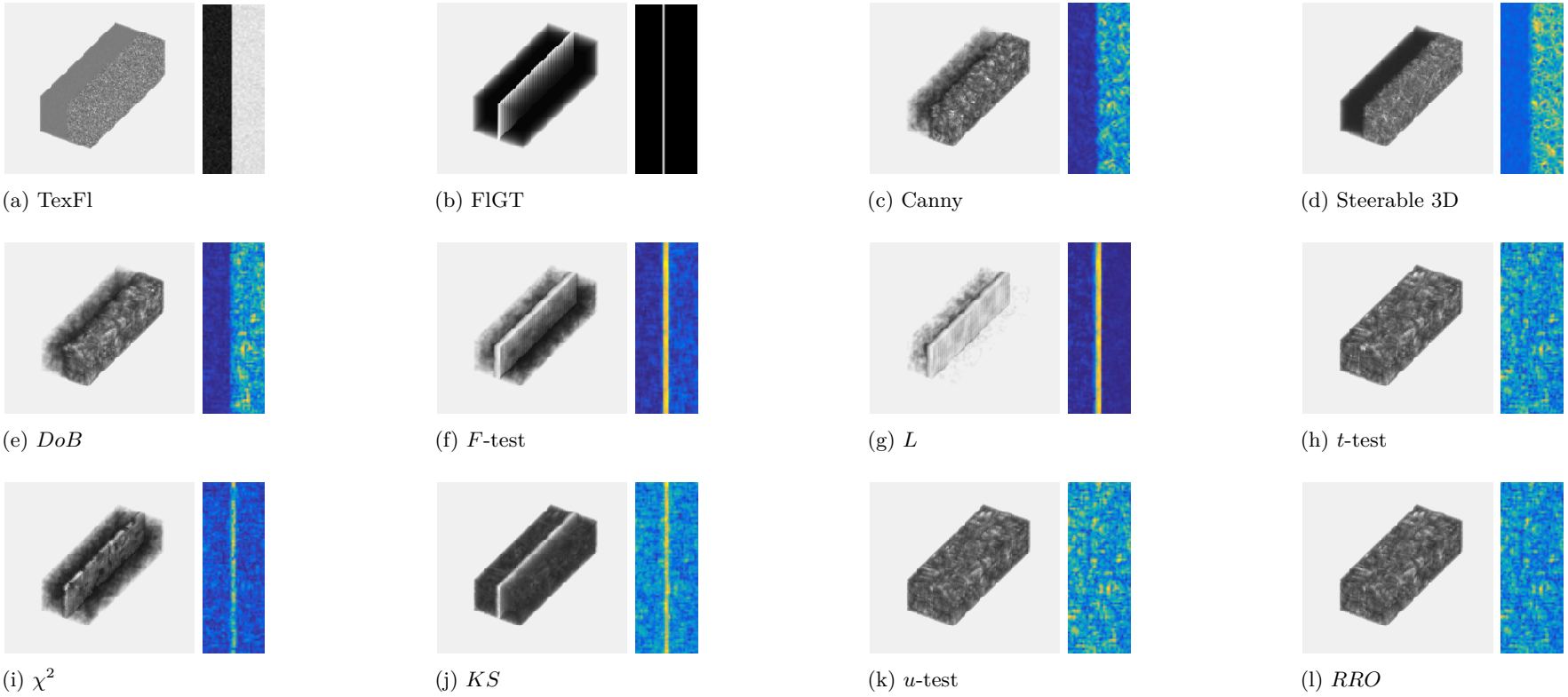
the same opacity. For added clarity in the visual assessment, a 2-D cross section is presented from the central layer of the image volume. The cross sectional layer uses Matlab's Parula colour space for enhanced visual differentiation between the minimum and maximum intensity levels.

(a) IntFl (b) FlGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) $F$-test (g) $L$ (h) $t$-test

(i) $\chi^2$ (j) $KS$ (k) $u$-test (l) *RRO*

Figure 5.9: Intensity based interface (IntFL) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

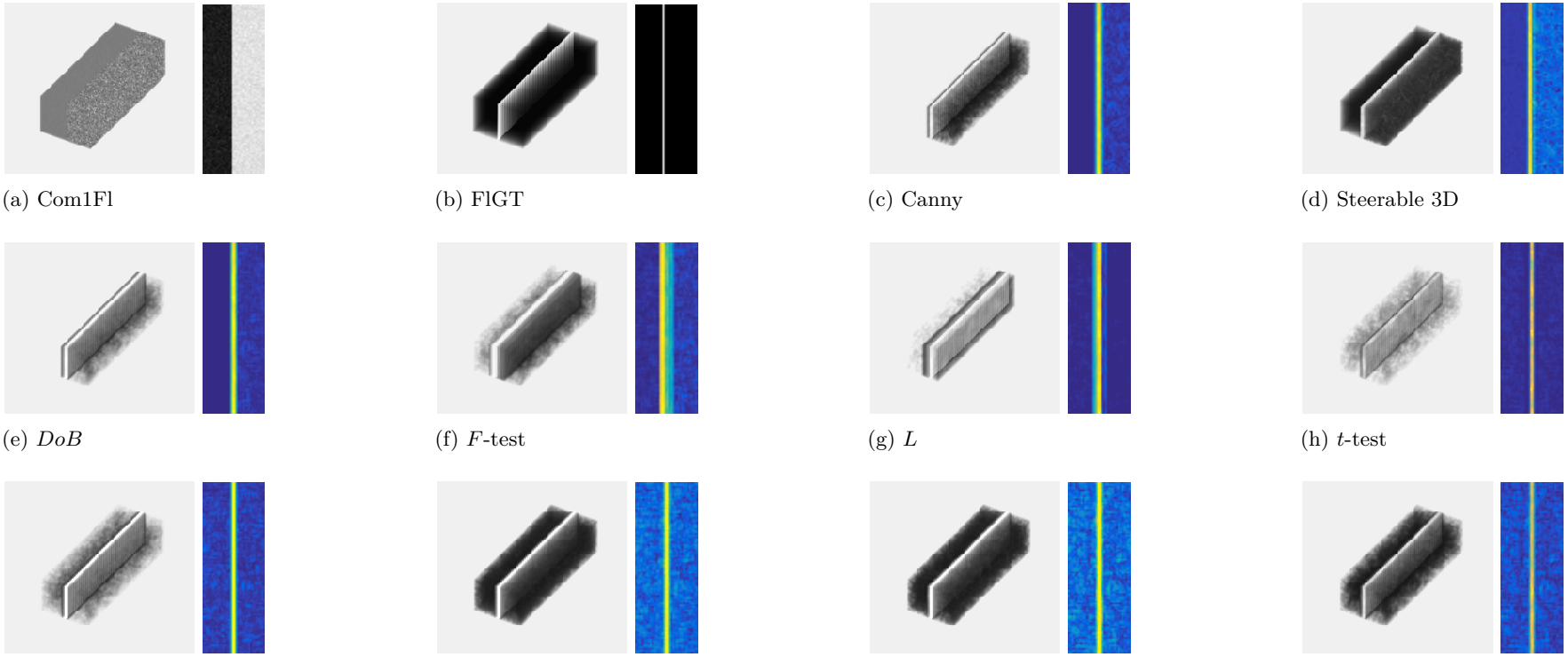


(a) TexFl (b) FlGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.10: Texture based interface (TexFL) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

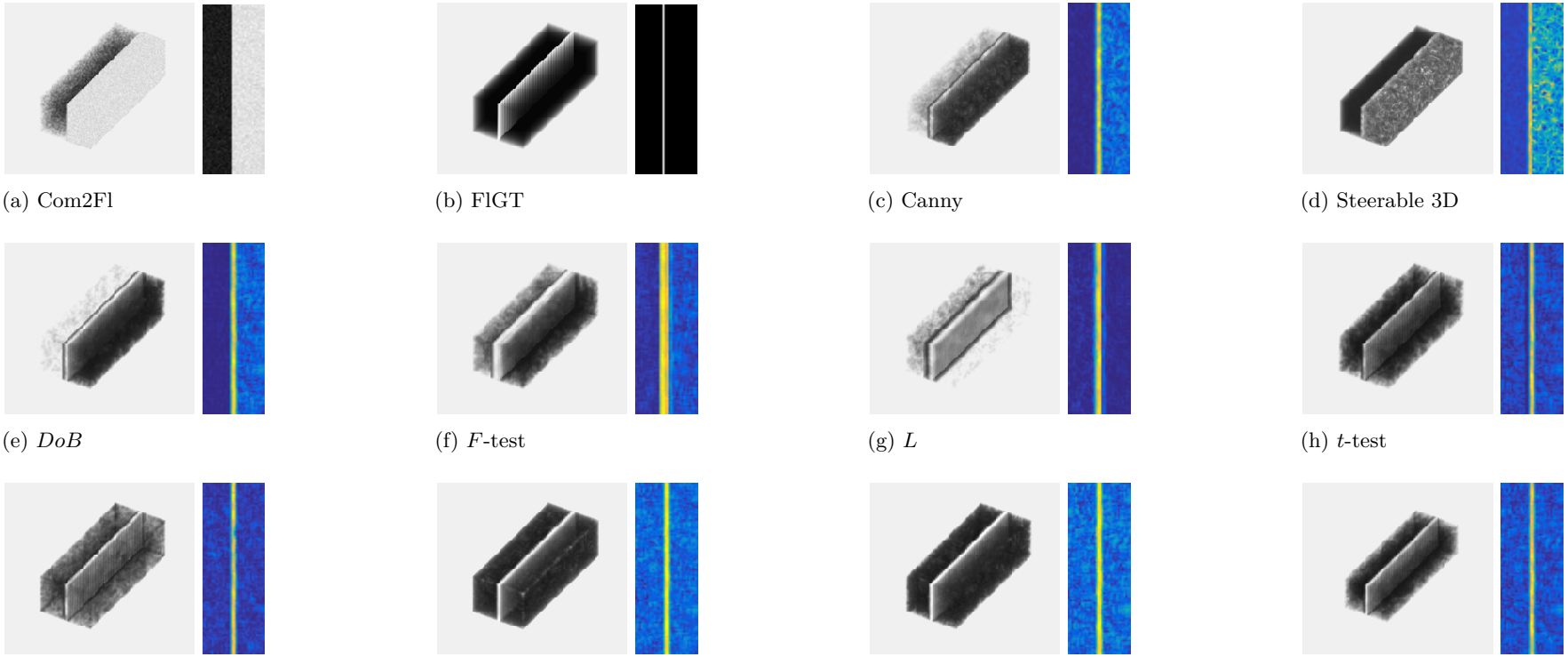


(a) Com1Fl (b) FlGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.11: Combinational 1 interface (Com1FL) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



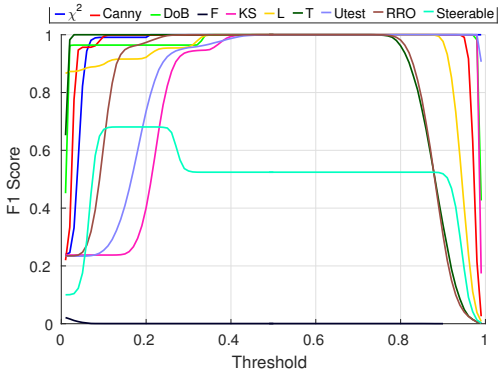
(a) Com2Fl (b) FlGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

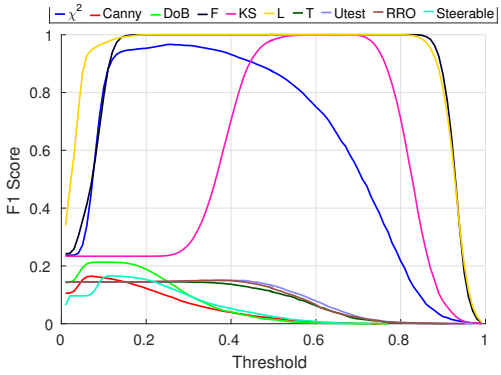
(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.12: Combination 2 interface (Com2FL) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

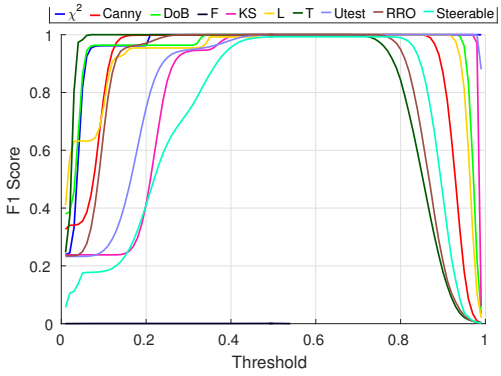
### 5.5.3 F-Measure Results



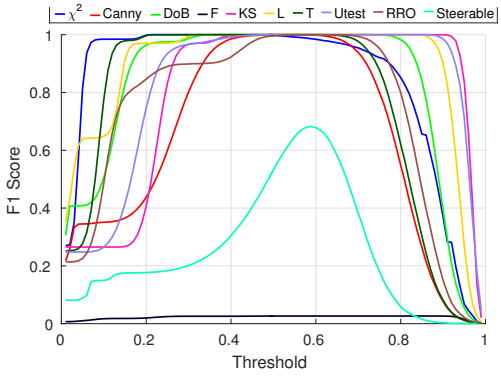

(a) IntFl




(b) TexFl




(c) Com1Fl




(d) Com2Fl

Figure 5.13: F-Measure analysis of Evaluation 1 tests. Results are gathered from a F1-score comparison between a normalised NMS filter result at 100 different hysteresis threshold values (lower threshold at 40% of upper) with that of the reference image.

### 5.5.4 Discussion

IntFl (Fig 5.2a) is primarily defined by a step change in average intensity between the regions. Upon visual inspection of the surface maps in Fig 5.9, it can be seen that all filter methods with an exception of the Fisher and likelihood tests successfully resolved the interface in its entirety, with all methods presenting good visual surface connectivity. The Canny operator, optimised for step changes in intensity performed strongly, signified by a solid connected boundary at an intensity level greater than the background noise (Fig. 5.9c). It should be noted that the $t$-test and $DoB$ test resolved the boundaries with the fewest spurious surface points, with the Canny method and $\chi^2$ method closely following. After NMS and thresholding, results without spurious responses were also achieved. The Likelihood method presented with good suppression

of spurious responses elsewhere in the image, giving rise to the possibility of rectifying the duplicate response in post processing.

The performance metric scores indicate optimal detection (F1-score of 1) for all tests barring the $F$-test and the Steerable filter methods (Fig. 5.13)a. The $F$-test method produced a duplicate boundary, either side of the ground truth interface location, contributing to their poor metric score. The $\chi^2$ test achieved a F1-score of 1 over the largest threshold range, followed closely by the Canny operator. From a visual standpoint, an F1 score 0.8 or above typically represents a successful detection result. Responses which produce a F1-score above 0.8 across a wide range of threshold values strongly indicate that the filter response has a strong signal to noise ratio. This is either due to a high magnitude detection of the surface in relation to any spurious responses, or due to few spurious responses in the output. However it should be noted that a wide threshold range was also achieved for the majority of the responses for the IntFl dataset.

Visual inspection of the TexFl surface maps indicate this to be a difficult surface interface to resolve (Fig. 5.10). There is a failure to resolve the interface on the part of the Canny, Steerable, $DoB$, $u$-test and $t$-test methods. There is a clear increase in over-detection across all filter methods indicated by a significant increase in the number of spurious responses.

A high variance region contains voxels with significantly different intensity values in relation to neighbouring values, intensity based operators produce high value outputs where significant changes in intensity exist, since these changes in intensity populate the entirety of the image volume, a high degree of spurious responses, or over-detection is present when gradient methods of surface detection are applied.

The best performing methods for TexFl are the $F$, $L$,and $\chi^2$ methods indicated by detection of the interface in Fig 5.10 whereas the other methods failed to detect the surface, and produced significant noise, particularly within the high variance region. The $RRO$ and $KS$ methods were able to resolve the interface, however a high degree of spurious surface points were present. Visually the likelihood method appears the least affected by over-detection, as well as presenting with good connectivity of the surface interface.

The performance metric scores indicate a clear success rate for $L, F$,$\chi^2$ and $KS$ signified by a peak F1 score of 0.9 or above (Fig. 5.13b). There was a clear failure to resolve the interface with the Canny, Steerable, $DoB$, $t$-test, $u$-test, and $RRO$ methods with an F1-score of 0.2 or below across the threshold range. The $L$ method performing best

overall, as well as offering the most robust output with regard to threshold selection, with the $F$-test also performing well.

Upon visual inspection of the Com1Fl surface maps, the $t$-test and $DoB$ test present fewer visible spurious surface points (Fig. 5.11) when compared against the other methods, as well as achieving good detection. The Canny and $\chi^2$ test also produce a strong surface response with good connectivity and relatively weak strength spurious surface points, which are removable through thresholding. The Steerable, $u$-test, $RRO$ and $KS$ methods also achieved successful detection of the interface, however over-detection artefacts are more prevalent. The $F$-test presented with over-detection artefacts to a greater extent than the other methods, including a more diffused surface boundary, existing over a larger voxel region. This is due to the aforementioned duplicated surface artefact, however with this interface type, the variance shift between regions also produced a high output at the surface location, leading to the appearance of over-detection at the surface interface. The likelihood method presented with a similarly displaced surface, but with less visible over-detection throughout the rest of the image volume when compared against the baseline methods and non parametric tests (Fig. 5.11g).

After thresholding, all methods including the likelihood method were able to obtain a successful detection of the surface interface of Com1Fl with the exception of the $F$-test which scored poorly (Fig. 5.13c). In terms of consistency, all the methods performed adequately over a wide threshold range with the $\chi^2$ being the most optimal, and the steerable having the narrowest threshold range. On this interface type it can be seen that the $t$-test is more suited to a low threshold range, while the $KS$ test is more suited to a high threshold range, with considerable overlap between the two methods (Fig. 5.13c).

The Com2Fl interface proved to be more challenging for the parametric and baseline filters to produce a successful visual output compared with Com1Fl (Figures. 5.11 and 5.12). While both combination images possess the same variance statistic, Com2Fl has a smaller average intensity difference ($\Delta\bar{x}75$ in Com1Fl and 21 in Com2Fl), resulting in an interface where a much greater overlap in the distribution of intensity values exist. This disparity in the results suggests that there is a tolerable amount of variance in the region properties for successful detection, providing the intensity difference between regions is great enough. However, when the intensity difference is reduced the interface is resolved at a weaker magnitude, effectively reducing the signal to noise ratio of the surface. Even though the variance in the region profiles in Com2Fl is the same Com1Fl, the fluctuations in the Com2Fl image causes an increased amount of relevant spurious responses which negatively impacts performance. As a consequence, over-detection is more significant. This is most noticeable in the steerable filter response which was

successful with Com1Fl, but performed poorly on Com2Fl (Figures. 5.11d and 5.12d). Here it can clearly be seen that the steerable filter suffered from a high degree of over-detection in the high variance region of the image volume, leading to an unsuccessful detection of the interface. The other statistical methods, and Canny method were still able to successfully resolve the interface, but with a higher degree of spurious responses, with the exception of the $F$ test which experienced a duplication of the surface with a large spatial displacement (Fig 5.12f). The $\chi^2$ method produced the clearest completion of the surface interface, with good connectivity (Fig 5.12i).

The performance metric scores for Com2Fl (Fig. 5.13d) support the visual results with successful completion for the majority of methods with a peak F1-score of 1, but this is with a reduced threshold range compared to Com1Fl (Fig. 5.13c) due to the lower signal to noise ratio of the filter responses. The most significant drop in performance between the Com1Fl and Com2Fl interfaces is that of the Steerable filter, which achieved a peak F1-score of 1 for the former and 0.68 for the latter.

### 5.5.5 Summary of Evaluation 1

Overall the $\chi^2$ test performed best, offering a combination of accurate detection across all interface types, as well as producing moderate to low spurious surface points. The surfaces were detected with good connectivity, and the performance metrics indicate good localisation of the detection. The $\chi^2$ test also performed well over a wide range of thresholds indicating a level of reliability as well as simplicity when setting parameters for detection. The results of Evaluation 1 indicate resolving power equal or surpassing the Canny and Steerable methods for the statistical methods, with the exception of the $F$-test. However the $F$-test performed well on the texture based interface where traditional gradient methods failed. The analysis showed that as the scale parameter is increased, the resolving power of the filter increased, this effect was greater with the non parametric $KS$ and $u$-test methods. Table 5.6 presents a success or failure assessment on each interface type, indicating that the $\chi^2$, $KS$ and likelihood methods were the only surface operators able to achieve successful detection across all single interface types. The strong resolving power from the $\chi^2$ and $KS$ methods, and success across the 4 interface types, indicate these to be preferential methods for most applications over the benchmark 3-D Canny surface filter. However, the Canny and Steerable methods can be applied using an effective small scale parameter ($3\times3\times3$), while the statistical methods effective minimum scale is $5\times5\times5$. Since a smaller scale parameter can resolve finer details of a more complex surface topology, there is a need to test on image volumes where the interface is not scale or layer invariant.

| Method | IntFl | TexFl | Com1Fl | Com2Fl |
|---|---|---|---|---|
| Canny | Accurate detection, connected surface, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, localised over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| Steerable | connected surface, displaced, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | connected surface, displaced, severe over-detection artefacts |
| $\chi^2$ | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, some discontinuity, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $DoB$ | Accurate detection, connected surface, no over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some localised over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| $F$-test | duplicated surfaces, some additional over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Inaccurate surface, duplication, some over-detection | Inaccurate, connected surface, duplicated surface and additional over-detection artefacts |
| $KS$ | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| $L$ | connected surface, no over-detection artefacts | Accurate detection, connected surface, very few over-detection artefacts | Accurate detection, duplicate surfaces are present, some localised over-detection artefacts | connected surface, duplicated surfaces, some over-detection artefacts |
| $u$-test | Accurate detection, connected surface, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| $RRO$ | Accurate detection, connected surface, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| $t$-test | Accurate detection, connected surface, no over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |

Table 5.6: Summary of evaluation 1 results

## 5.6 Evaluation 2: Topology test with known distributions.

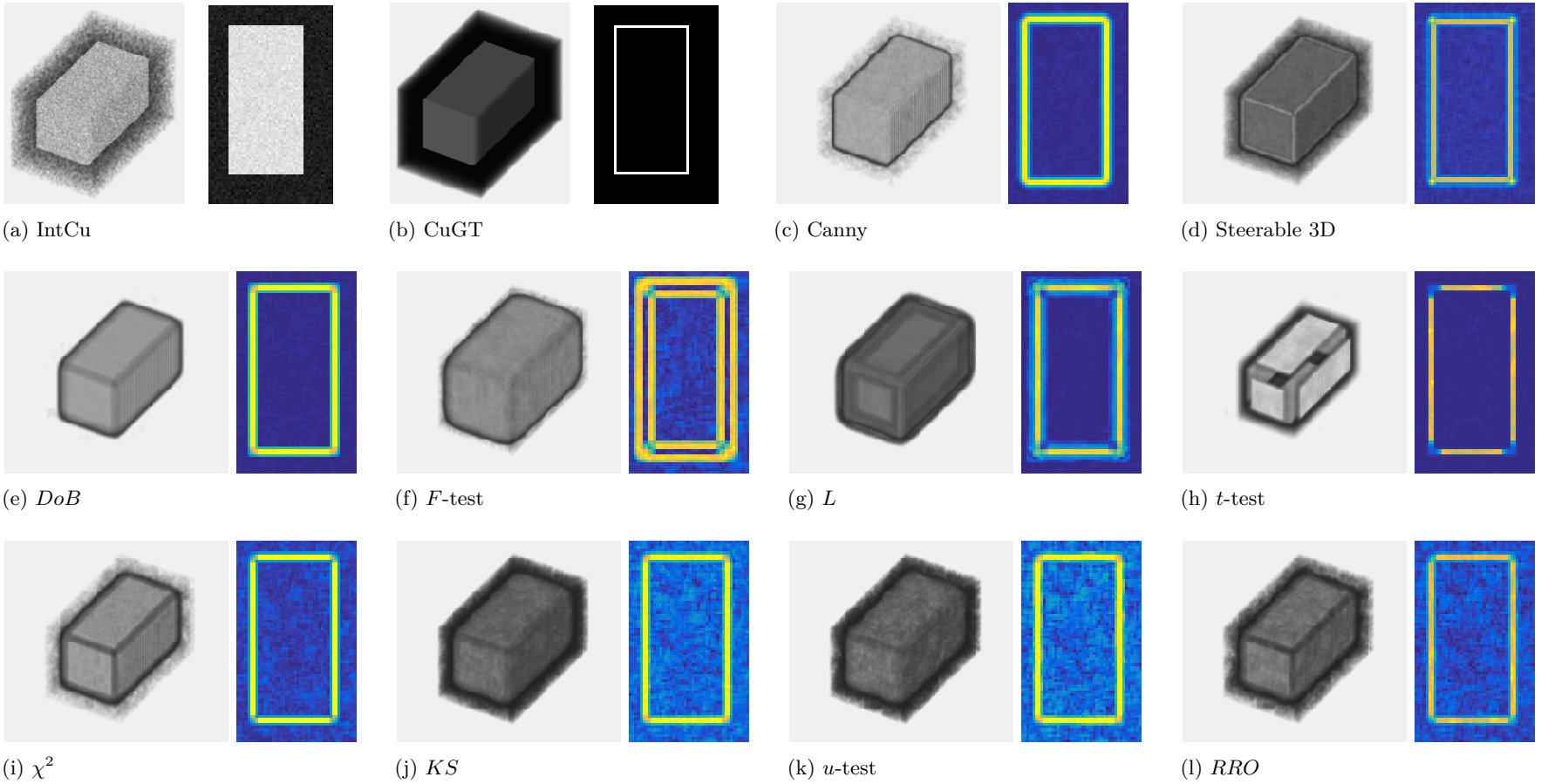
### 5.6.1 Evaluation 2: Aim

To fully exploit the capabilities of a true 3-D surface detection method, the topology or structure of an interface should be reflective of three dimensional data, and possess a surface interface that is not invariant throughout the layers. The statistical properties of the regions in evaluation 1 are repeated with the different topological structures, these structures are illustrated in Fig 5.5 and the properties are presented in Table 5.2. Therefore any differences in the results of evaluation 2 are due to the topological structure of the interface.

Evaluation 1 uses scale invariant image volumes in order to assess the resolving power of the detection methods, a consequence of this is the ability to increase the scale parameter of the detection method with minimal penalty in terms of inducing localisation error.

### 5.6.2 Topology 1: Cuboid (Cu)

#### 5.6.2.1 Visual Results (Cu)

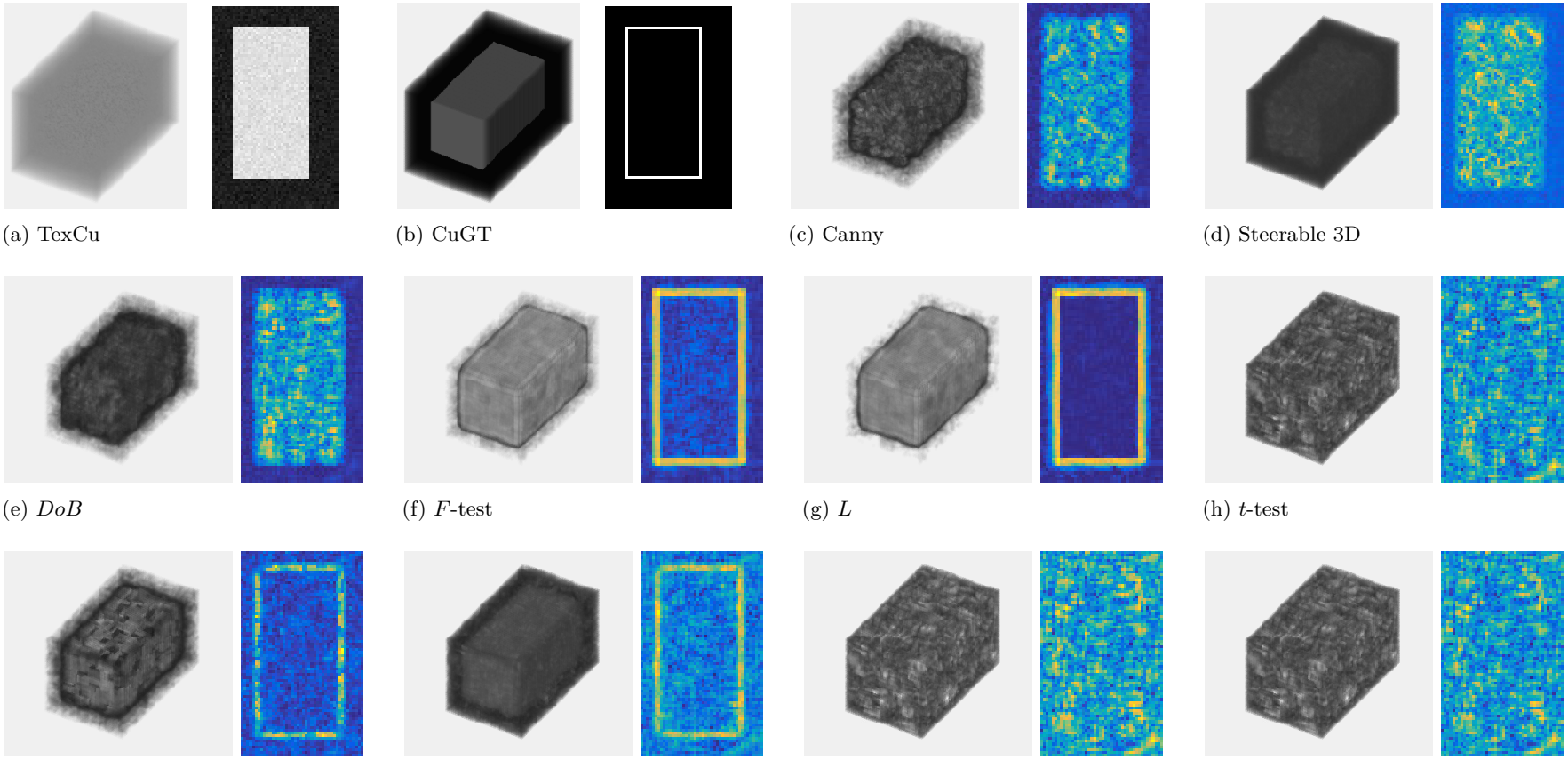


(a) IntCu (b) CuGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.14: Intensity based cuboid interface (IntCu) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

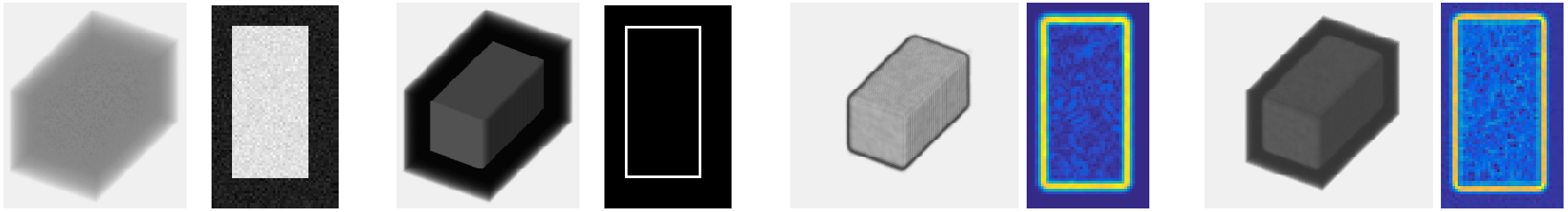


(a) TexCu (b) CuGT (c) Canny (d) Steerable 3D

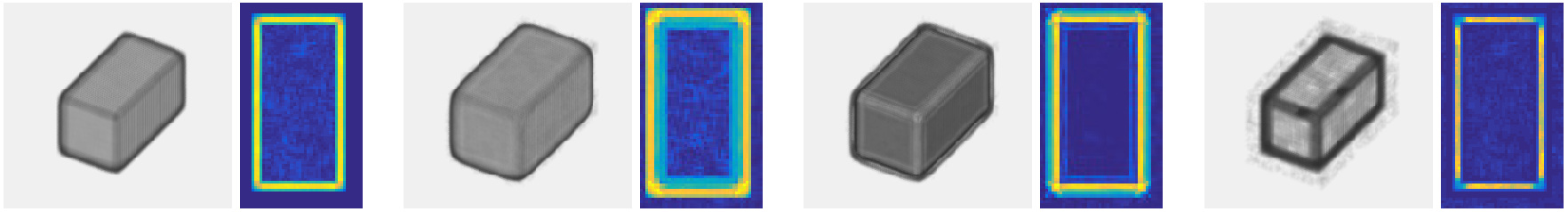
(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

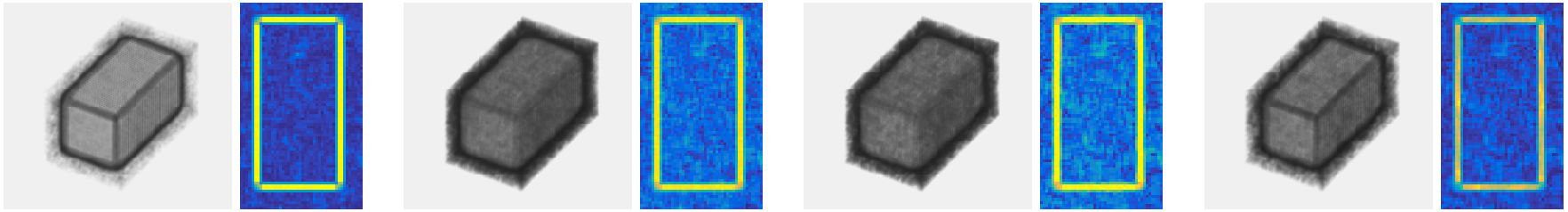
Figure 5.15: Texture based cuboid interface (TexCu) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



(a) Com1Cu (b) CuGT (c) Canny (d) Steerable 3D

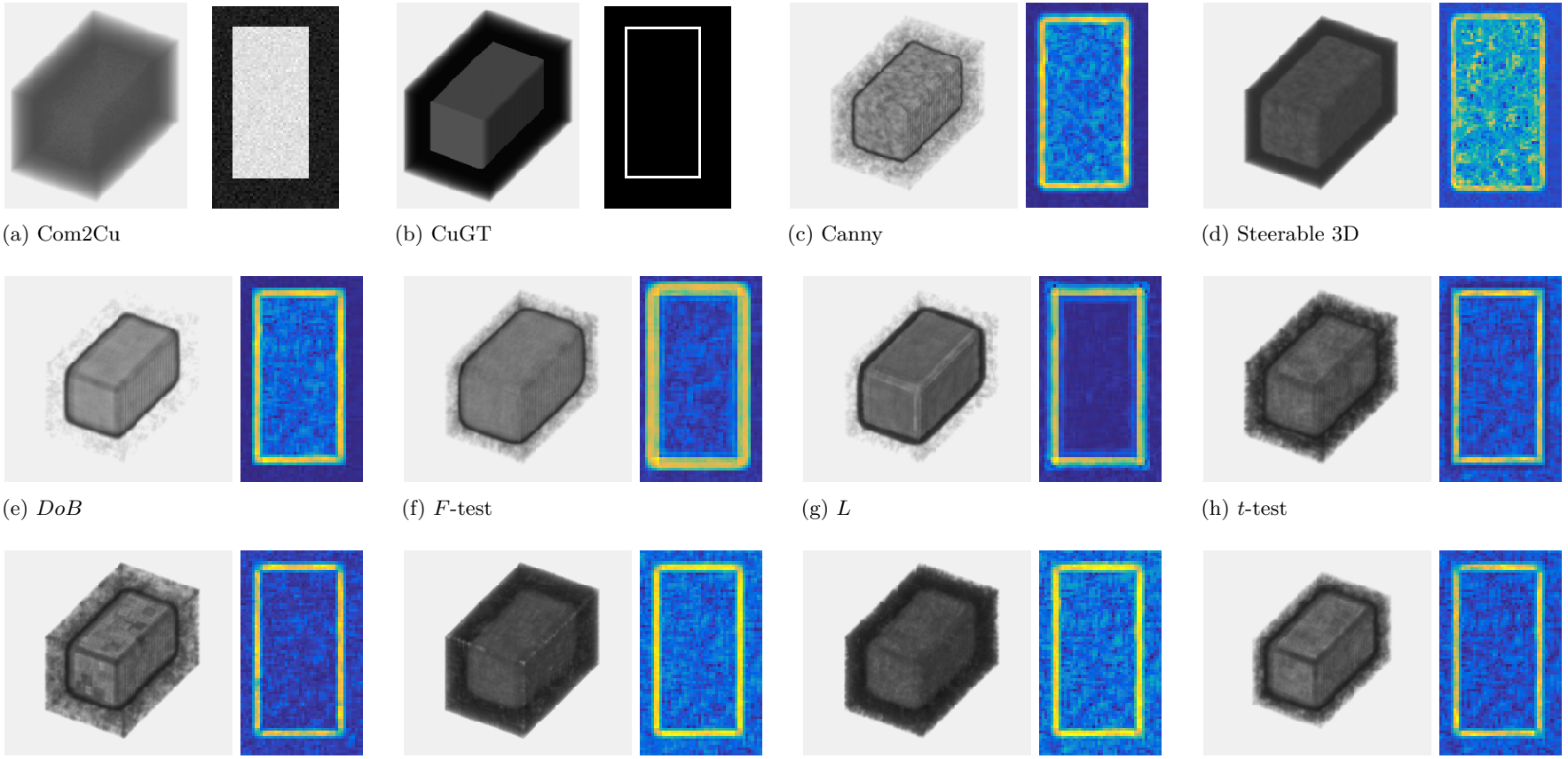


(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test



(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.16: Combinational 1 cuboid interface (Com1Cu) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



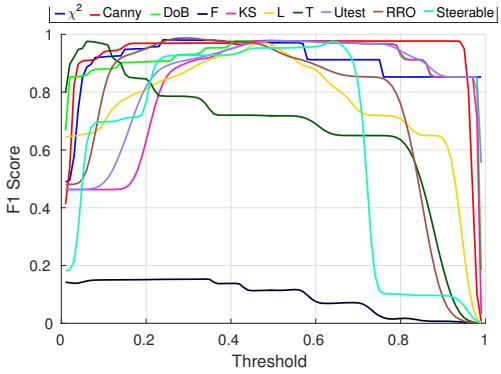
(a) Com2Cu (b) CuGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

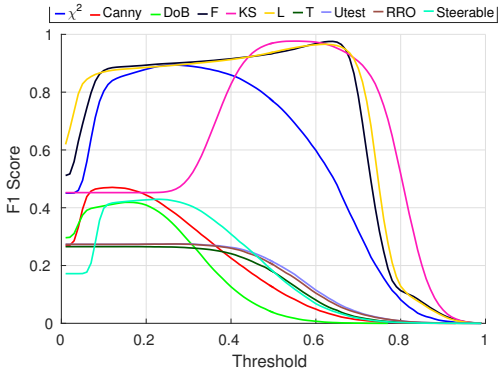
(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.17: Combination 2 cuboid interface (Com2Cu) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

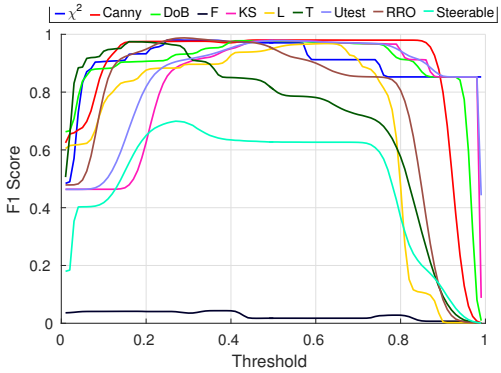
#### 5.6.2.2 F-Measure Results (Cu)



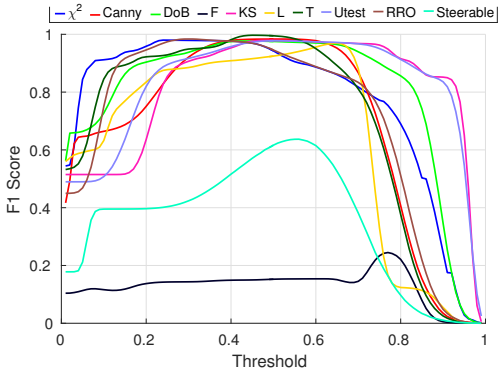

(a) IntCu




(b) TexCu




(c) Com1Cu




(d) Com2Cu

Figure 5.18: F-Measure analysis of Topology 1 (Cu) tests. Results are gathered from a F1-score comparison between a normalised NMS filter result at 100 different hysteresis threshold values (lower threshold at 40% of upper) with that of the reference image.

#### 5.6.2.3 Discussion

Visually, the results from the IntCu interface are similar to the IntFl results from evaluation 1 in terms of resolving power. With the *DoB* and $t$-test performing to a high standard with complete detection and minimal spurious surface points. Additionally the benchmark 3-D Canny operator also performed strongly. F-measure scores show that the Canny filter operated successfully over the widest range of threshold values (Fig.5.18a). The scores across all tests are lower than the single interface images presented in evaluation 1. This is due to some roughness or inaccuracies along the cuboid edges and corners of the surface interface. Upon further examination of the individual 2-D layers of the volume, it can be seen that $F$-test fails due to a duplicated surface response, and that the $t$-test is not optimal in areas where corners are situated (Fig.5.14).

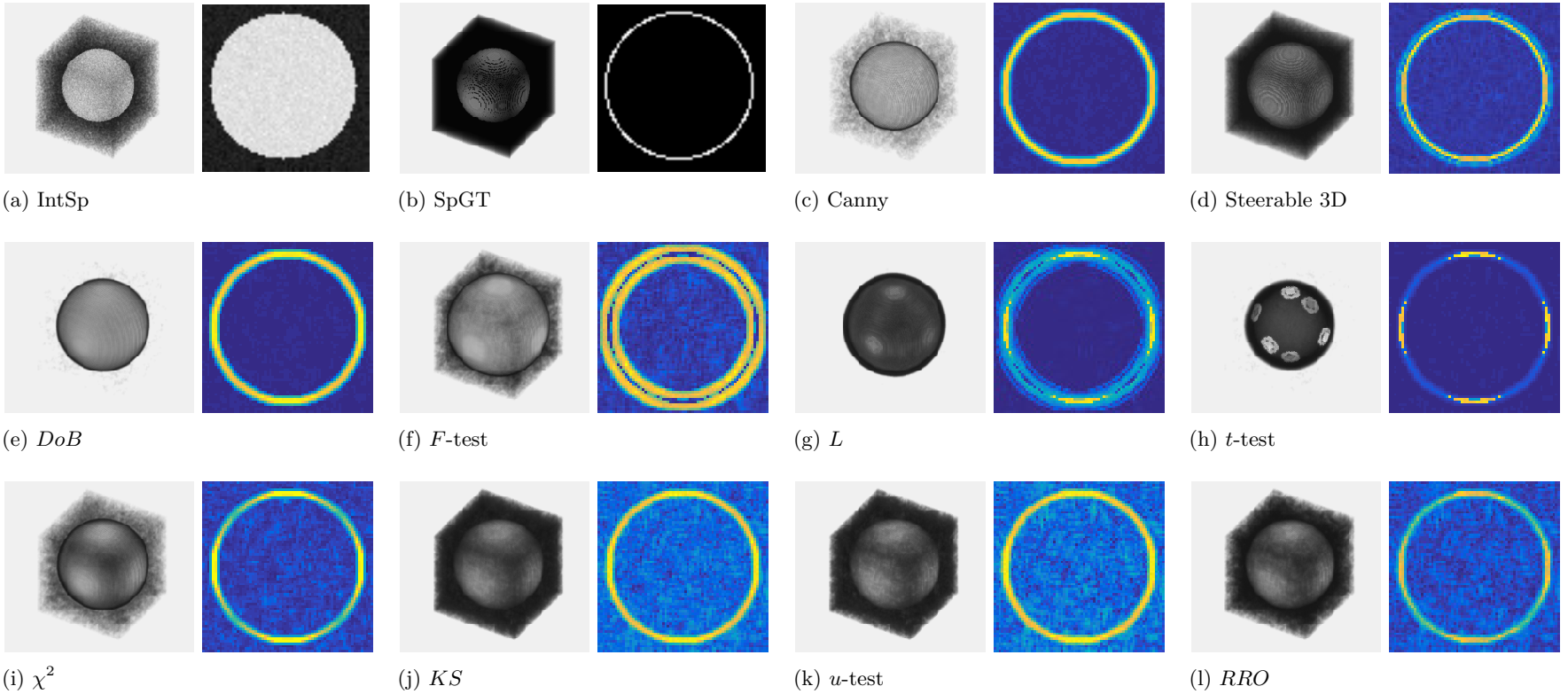
The texture based interface again proves more challenging for a number of the filters. A 3-D visual inspection of the results show that the $t$-test, $u$-test and $RRO$ filters were unable to differentiate between a surface region and non-surface region in its entirety (Fig 5.15h,k,l). A 3-D inspection of the Canny, Steerable and $DoB$ method would indicate a successful detection of the boundary (Fig. 5.15c,d,e 3-D ). However, further analysis reveals this not to be the case. Through examination of the individual 2-D layers of the image volume, it can be seen that these methods did not detect the surface, and instead produced a high value output for the central high variance region (Fig. 5.15c,d,e 2-D). It is clear that the most successful methods in this instance are the $L$, $F$, $\chi^2$ and $KS$ tests. With the $L$ and $F$-tests producing very clear accurate surfaces (Fig. 5.15f,g,i,j). However, while the $KS$-test accurately detects the complete surface it does suffer from some over-detection artefacts. Contrastingly, the $\chi^2$ test produces fewer noticeable spurious surface points, but does lack some connectivity in the detection of the surface interface. These findings are supported by the objective results with clear successes for the $L$, $F$, $KS$ and $\chi^2$ tests, and failure for the remaining filters (Fig 5.18b).

The 3-D visual results indicate all methods successfully detect the surface interface in a similar fashion to Evaluation 1 (Fig. 5.16). Analysis of the individual layers reveal the $F$-test produces a diffused boundary (Fig. 5.16). The 2-D layer results indicate that the $\chi^2$ method achieves the best trade-off between detection and over-detection, which is reflected the objective analysis results (Fig. 5.18c).

Figures 5.17 and 5.18d indicate that the Com2Cu interface also mimics the Com2Fl results from evaluation 1, with all methods producing a successful detection of the surface with the exception of the steerable filter. Visually the $F$-test appears to have successfully detected the surface boundary, however a duplicated interface is also produced. This is not clear in the 3-D visualisation since the internal boundary in occluded by the external outer boundary (Fig. 5.17e). Poor performance scores for the $F$-test indicate this to be the case.

### 5.6.3 Topology 2: Spherical (Sp)

#### 5.6.3.1 Visual Results (Sp)

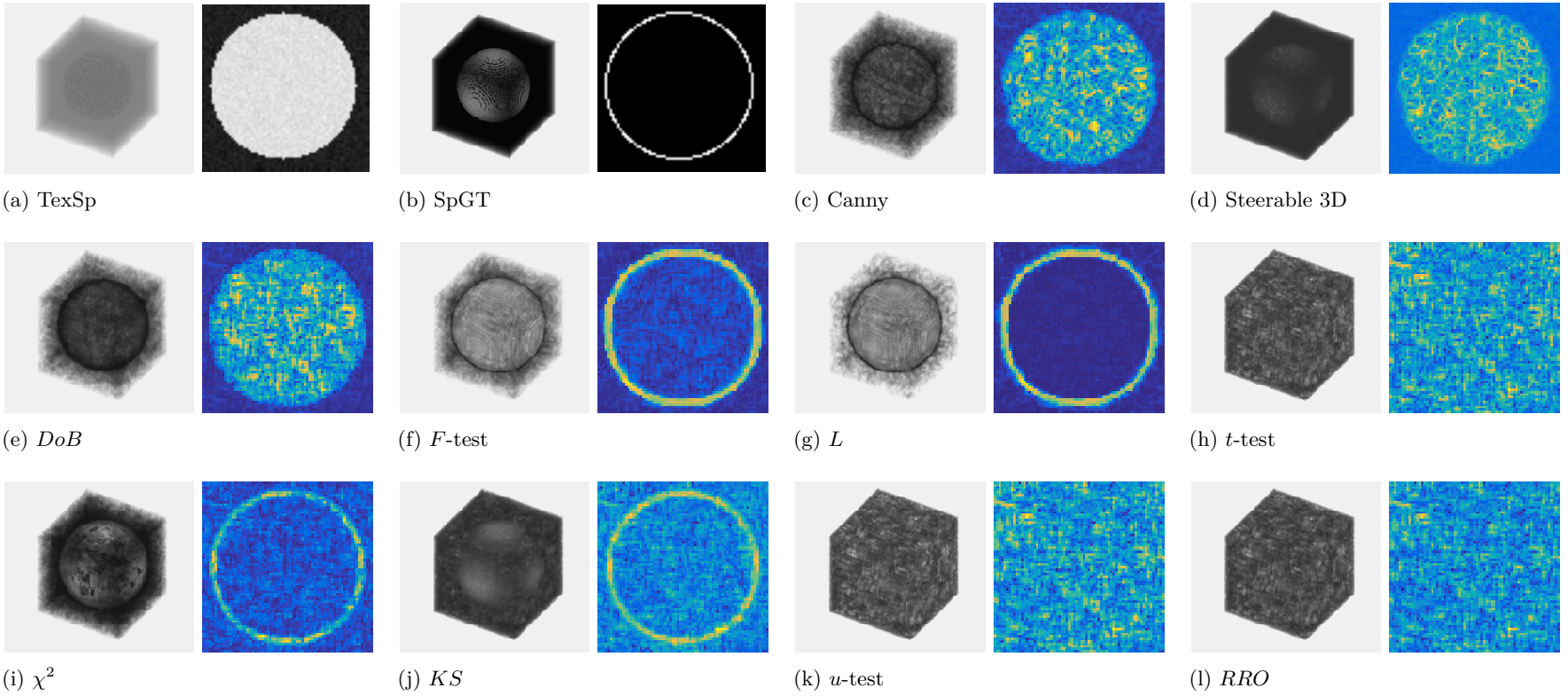


(a) IntSp (b) SpGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.19: Intensity based spherical interface (IntSp) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



(a) TexSp (b) SpGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.20: Texture based spherical interface (TexSp) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

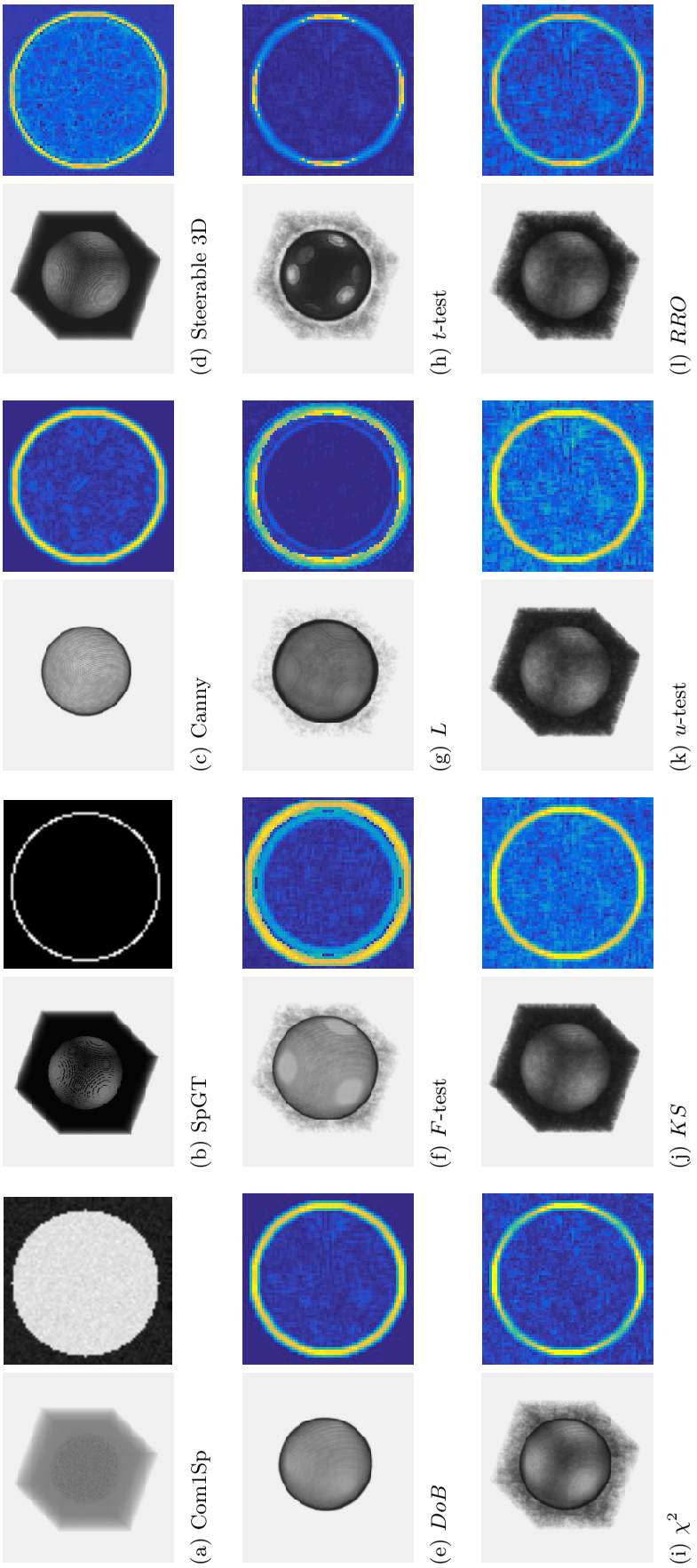
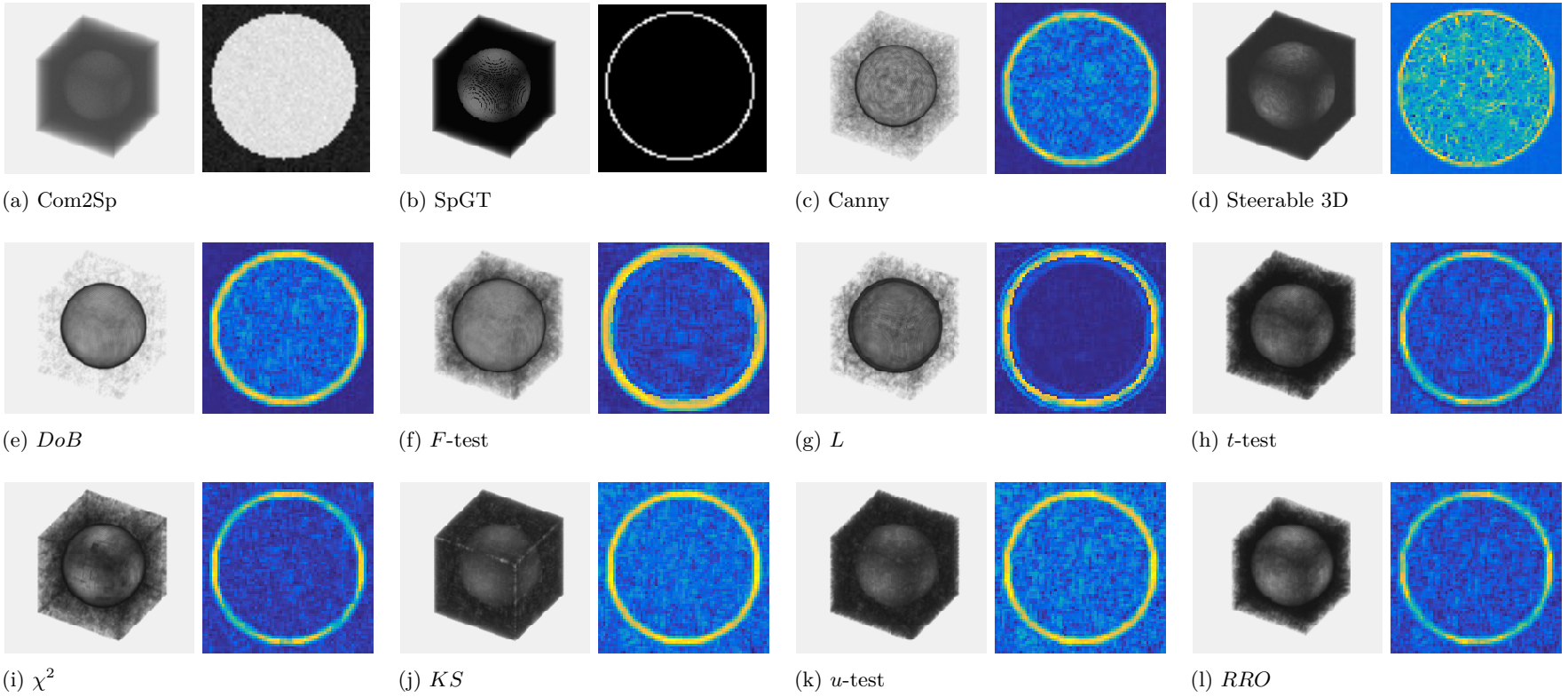



Figure 5.21: Combinational 1 spherical interface (Com1Sp) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



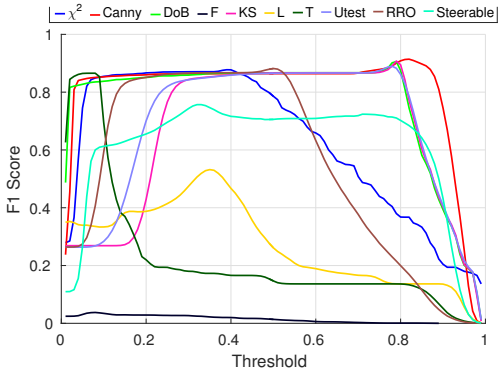
(a) Com2Sp (b) SpGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

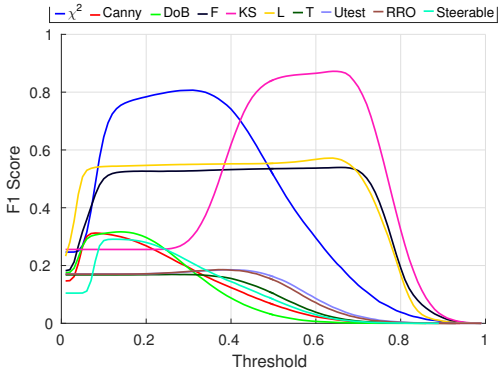
(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.22: Combination 2 spherical interface (Com2Sp) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

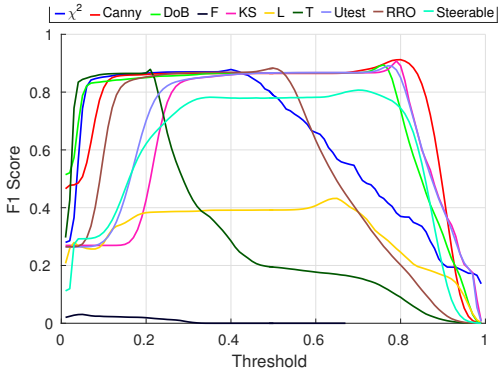
#### 5.6.3.2 F-Measure Results (Sp)



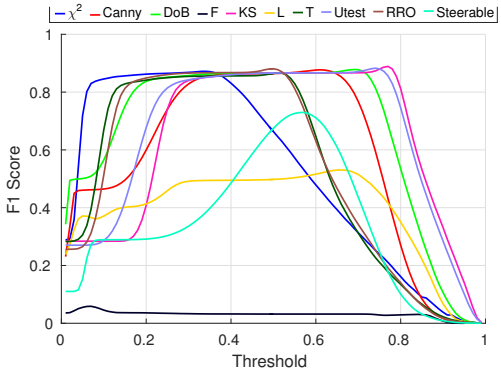

(a) IntSp




(b) TexSp




(c) Com1Sp




(d) Com2Sp

Figure 5.23: F-Measure analysis of Topology 2 (Sp) tests. Results are gathered from a F1-score comparison between a normalised NMS filter result at 100 different hysteresis threshold values (lower threshold at 40% of upper) with that of the reference image.

#### 5.6.3.3 Discussion (Sp)

The IntSp results are consistent with previous intensity based interface findings, with strong performances from all methods with the exception of the $F$ and $L$ tests which were less successful due to the duplicated surface responses, and a similar artefact presents in the Steerable result (Fig 5.19d,f,g).

Visually this difficulty can be observed by surfaces which do not have uniform magnitude and is most pronounced in Fig 5.19h, here the $t$-test produces very high magnitude surface values in specific regions of the sphere, in the cross sectional result these high magnitude responses are observed in the North, South, East and West surface locations, whereas low magnitude surface responses are observed in the North East, North West, South East and South West surface locations. This effect is lessened as variance increases in the image (Figures 5.21h and 5.22h). This is due to the scale of the mask in relation

to resolution of the image and the degree of curvature in the topology. It is a factor because the local maxima of a surface is defined where the differential between two neighbourhood mask regions is at its greatest. For this to occur, an equal number of voxels from each region should be present in the neighbourhood mask in order for that differential to be maximised. A discrepancy between the number of voxels from each region in the neighbourhood mask would indicate that the mask is in close proximity to a surface but not centred accurately on the interface. With a spherical interface, an unequal number of voxels from each region are usually present in the neighbourhood mask, but because the sphere is subject to the discrete resolution of the image, it is not perfectly spherical, and therefore this disparity is not uniform along the interface, and this leads to detection of a surface with an inconsistent magnitude. This is why the effect is lessened in Com1Sp and Com2Sp as increased variance in the region profiles of the datasets reduces the likelihood of 'spikes' in surface magnitude to occur. The Canny method overcomes this due to fact Gaussian filtering is utilised which smooths the curvature for a more even response, leading to good detection (Fig 5.19c). Performance measures indicate that curved surface topology of the spherical interface did negatively affect the performance of the detection methods when compared with the flat interfaces of the cuboid, and showed the Canny to be the least affected by the issue. (Figures 5.18a and 5.23a).

Fig. 5.20f,g indicates that the parametric variance methods ($F$-test and $L$) to be the strongest performing on the texture based interface (TexSp). Additionally, good detection of the interface was also achieved by the $KS$ and $\chi^2$ test methods with moderate over-detection and under-detection respectively, as can be seen by background noise in the $KS$ surface maps, and a surface with some discontinuities in the case of the $\chi^2$ test. Objective analysis places the $KS$ and $\chi^2$ methods ahead of the $L$ and $F$-tests (Fig. 5.23b).

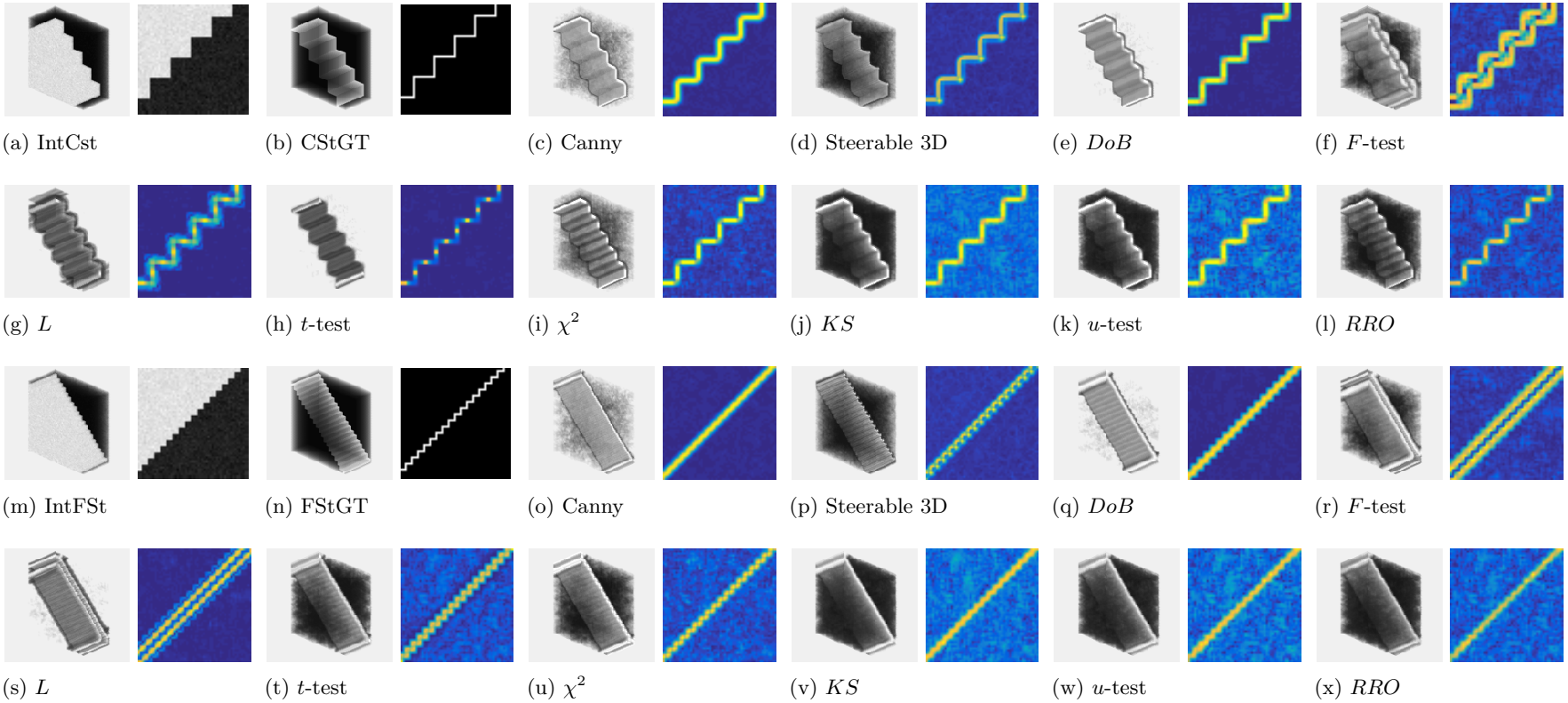
Figures 5.21 and 5.22 show the 3-D visual results from the combination interface types. The visual results with a spherical topography concur with the previous findings for this interface type. For Com1Sp, the $DoB$, $t$-test and Canny method are the preferred filter methods, visually these methods provide complete detection some spurious surface responses (Fig 5.21c,e,h). 3-D visualisation shows the $DoB$ method resolves the complete surface with the fewest spurious responses (Fig 5.21e). However all methods appear to successfully resolve the surface. For Com2Sp, over-detection is more apparent, with every filter method producing some spurious surface responses. The $DoB$ method offers the fewest noticeable spurious surface points, followed by the Canny method. The $L$ and $F$ methods show good performance visually with good suppression of noise (Figures 5.22f and 5.22g). However, not all findings from the visual analysis agree with

the objective results. The F-measure analysis indicate the $L$ test and $F$-test perform poorly on both combination interfaces (Fig 5.23c,d). When looked at more closely, it can be seen there is a duplication of the surface combined with a slight displacement of the interface (Figures 5.21f,g). This is the cause of the low F1-score. For Com1Sp the $KS$ test and Canny methods achieve the highest peak score both at an upper threshold of 0.79 (Fig 5.23c). For the Com2Sp interface type, the $KS$ test again achieves the highest peak score, however on this interface type a number of statistical methods score marginally higher peak F1-scores then the baseline Canny method, including $DoB$, $RRO$ and $u$-test methods. In terms of robustness to threshold sensitivity, the $DoB$ method offers the widest threshold level range at optimal performance (Fig 5.23d).

### 5.6.4 Topologies 3 and 4: Staircases (CSt and FSt)

#### 5.6.4.1 Visual Results (CSt and FSt)

Figures of the visual results are provided on the next page onwards.



(a) IntCst (b) CStGT (c) Canny (d) Steerable 3D (e) *DoB* (f) *F*-test

(g) *L* (h) *t*-test (i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

(m) IntFSt (n) FStGT (o) Canny (p) Steerable 3D (q) *DoB* (r) *F*-test

(s) *L* (t) *t*-test (u) $\chi^2$ (v) *KS* (w) *u*-test (x) *RRO*

Figure 5.24: Intensity based staircase interfaces (IntCSt, IntFST) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

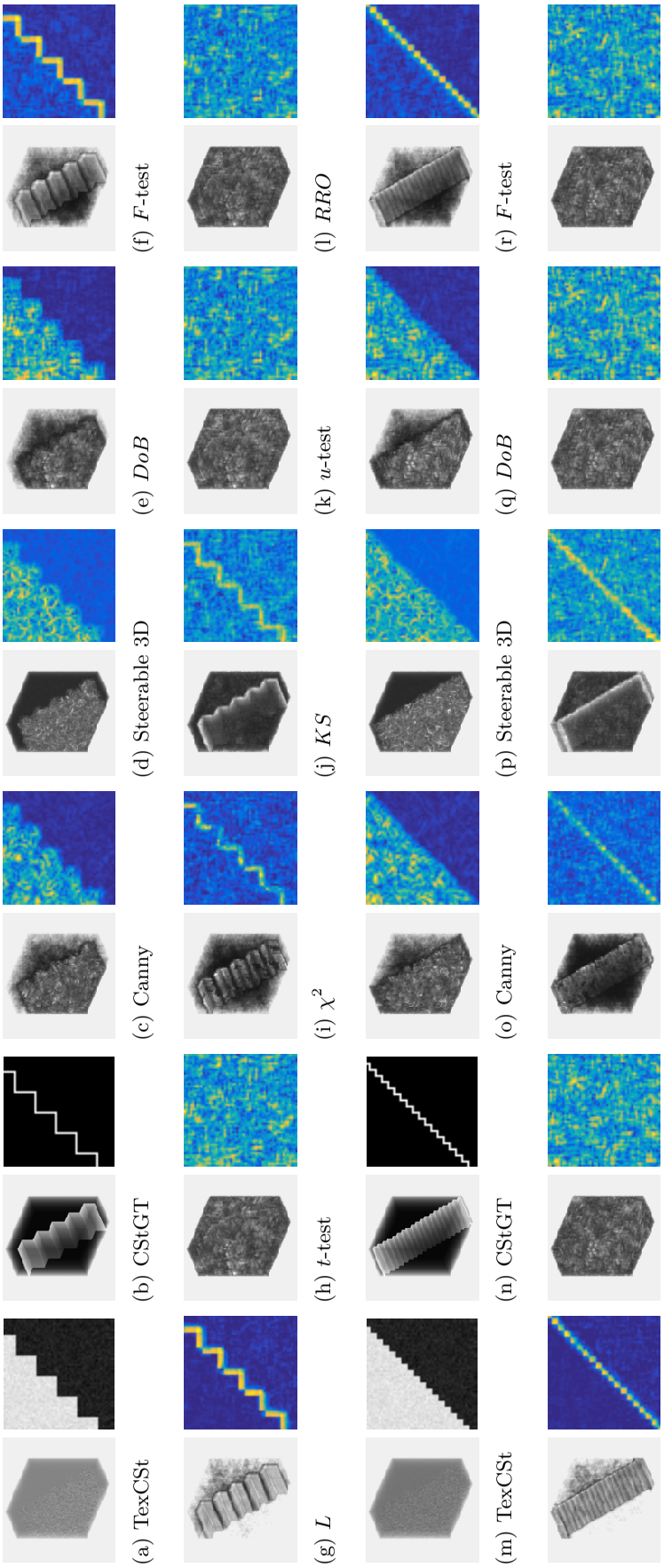
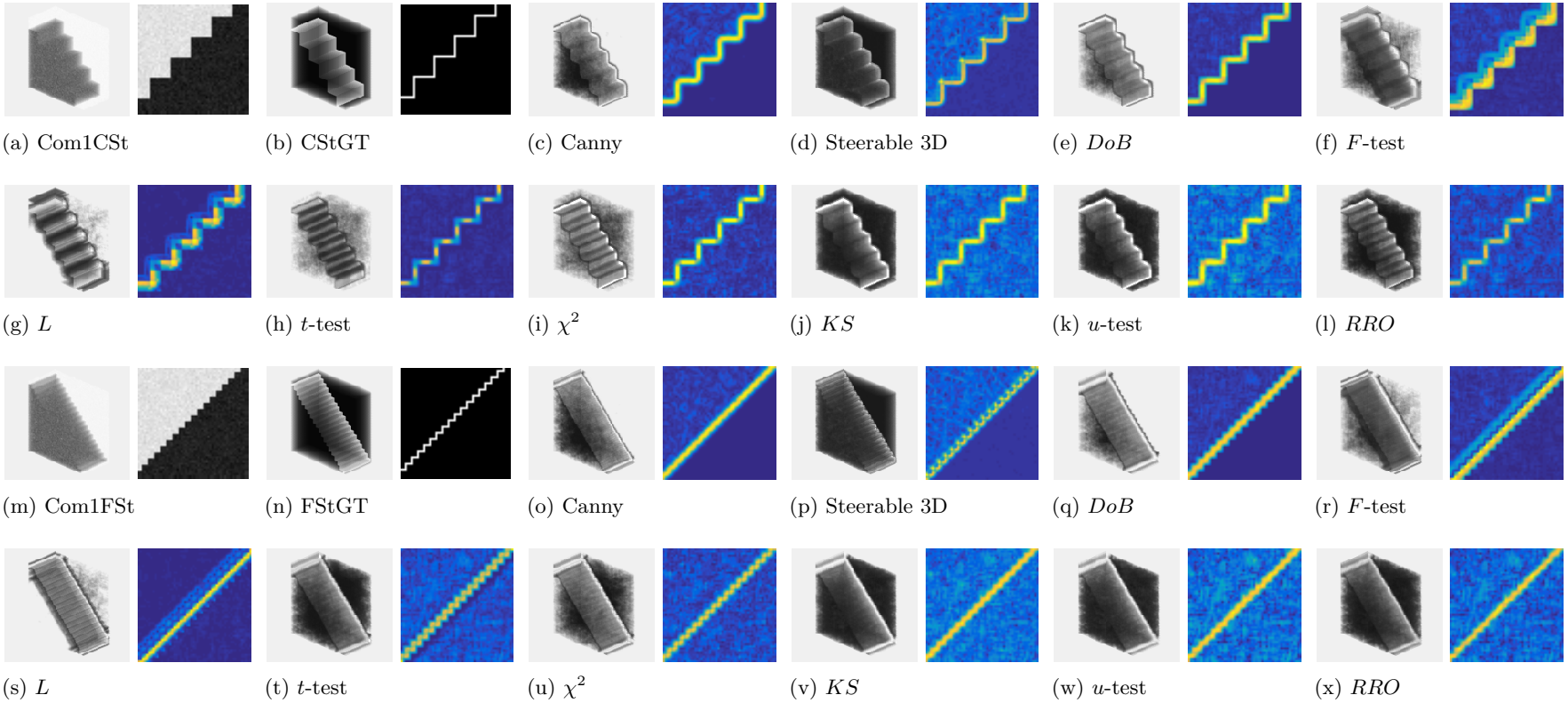



Figure 5.25: Texture based staircase interfaces (TexCSt, TexFST) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



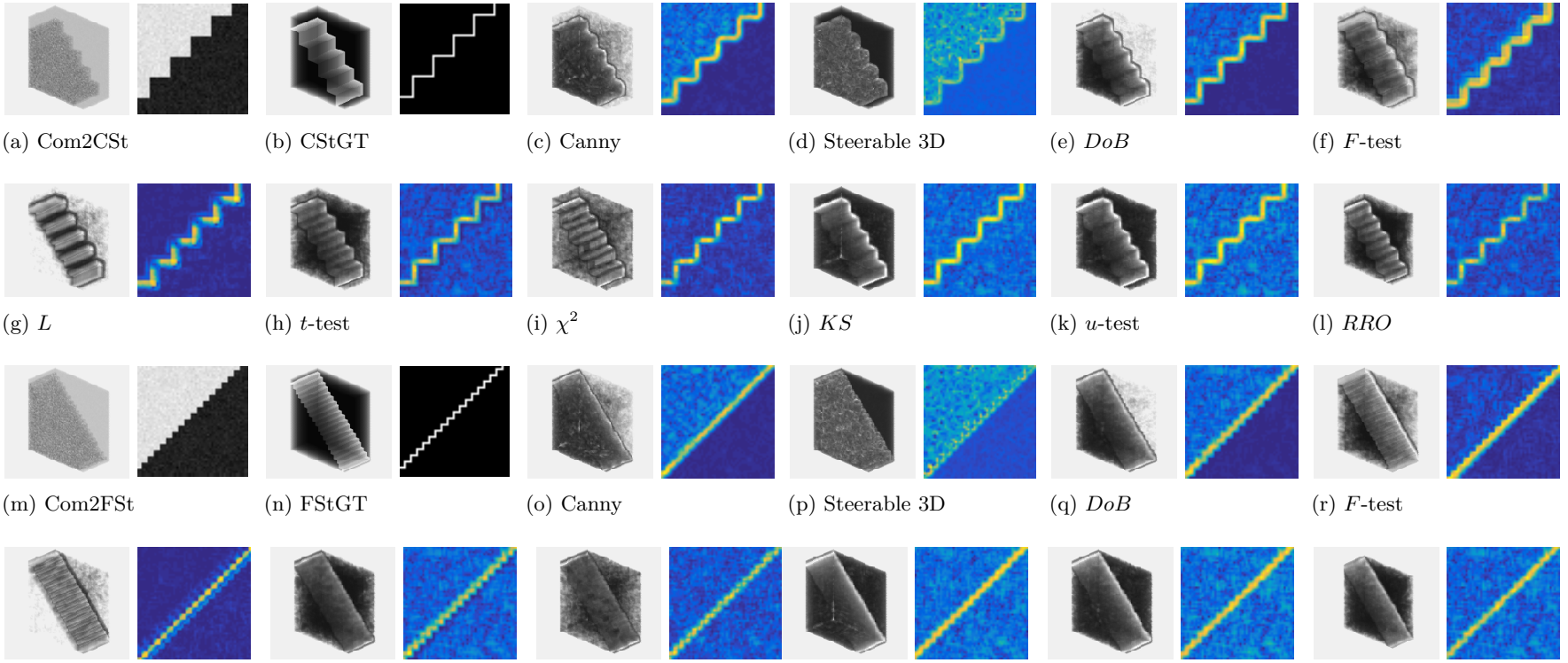
(a) Com1CSt (b) CStGT (c) Canny (d) Steerable 3D (e) *DoB* (f) *F*-test

(g) *L* (h) *t*-test (i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

(m) Com1FSt (n) FStGT (o) Canny (p) Steerable 3D (q) *DoB* (r) *F*-test

(s) *L* (t) *t*-test (u) $\chi^2$ (v) *KS* (w) *u*-test (x) *RRO*

Figure 5.26: Combination 1 staircase interfaces (Com1CSt, Com1FST) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



(a) Com2CSt (b) CStGT (c) Canny (d) Steerable 3D (e) *DoB* (f) *F*-test

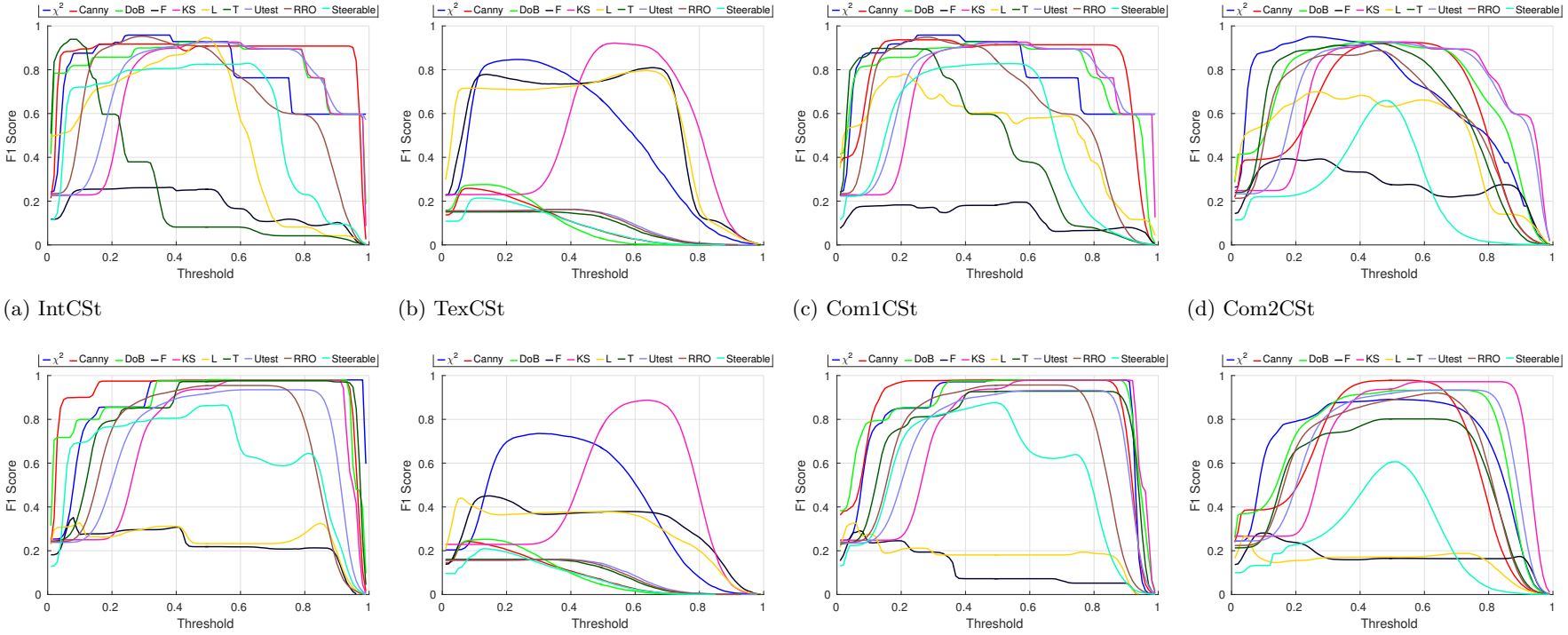
(g) *L* (h) *t*-test (i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

(m) Com2FSt (n) FStGT (o) Canny (p) Steerable 3D (q) *DoB* (r) *F*-test

(s) *L* (t) *t*-test (u) $\chi^2$ (v) *KS* (w) *u*-test (x) *RRO*

Figure 5.27: Combination 2 staircase interfaces (Com2CSt, Com2FST) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

#### 5.6.4.2 F-Measure Results (CSt and FSt)




(a) IntCSt (b) TexCSt (c) Com1CSt (d) Com2CSt

(e) IntFSt (f) TexFSt (g) Com1FSt (h) Com2FSt

Figure 5.28: F-Measure analysis of Topologies 3 and 4 (CSt and FSt). Results are gathered from a F1-score comparison between a normalised NMS filter result at 100 different hysteresis threshold values (lower threshold at 40% of upper) with that of the reference image.

#### 5.6.4.3 Discussion (CSt and FSt)

Across the IntCSt and IntFSt images, the *DoB* test performed best, characteristically only a minor amount of over-detection occurred while maintaining strong accurate detection of the interface (Fig. 5.24e,q). Good detection is also achieved by the Canny operator for the CSt image, however for the FSt volume, whilst noise in the output remains low, and the surface magnitude is strong, there is some positional detail lost in the result. This indicated by a smooth flat detected surface instead of a stepped structure and this is likely apparent due to the Gaussian smoothing stage of the Canny process which blurs finer details prior to detection (Fig. 5.24o). While this the Gaussian stage proved to be beneficial for IntSp interface type (Fig 5.19c) where a uniformly smooth curvature in the response was optimal, here it is shown to be detrimental in preserving the finer details of the image. The $\chi^2$ method produced a successful output but with more spurious responses than both the Canny and *DoB* methods, however this is offset by accurate detection of the finer details of the interface and a strong magnitude surface (Fig. 5.24i,u).

The $t$-test performed strongly from a visual perspective in the case of IntCSt image volume, with fewer spurious responses compared to the IntFSt results (Fig 5.24h,t). F-measure analysis suggests that the $t$-test achieves similar performance but at different threshold ranges. The IntCSt results provided a peak F1-score of 0.9 for the $t$-test, but in a narrow threshold range (Fig 5.28a), whereas the IntFSt results also achieves a peak F1-score of 0.9 across a higher but with a wider threshold range. This discrepancy is due to how the $t$-test performs in corner regions. For the IntFSt image the space between corners in smaller, this results in an interface which is lower in magnitude relative to the spurious responses, which is why the spurious responses appear with higher magnitude in the normalised results when compared with IntCSt. However with the IntCSt image, the surface interface between corners produces a very high magnitude response. As a consequence the magnitude of the surface is less uniform when compared with the IntCSt results. Due to this factor, lower hysteresis thresholds are required to preserve the surface. The non uniformity of the response therefore limits the threshold range where the $t$-test is optimal. The results of the $t$-test between the different staircase topologies is more similar between Com1FSt and Com1CSt (Fig 5.26h,t), and between Com2FSt and Com2CSt (Fig 5.27h,t). This is due to the fact increased variance in the data nullifies the high magnitude 'spikes' and as a consequence improves the uniformity of the surface magnitude responses. This results in outputs which more closely resemble that of IntFSt and therefore have a broader successful threshold range (Fig 5.28).

The $L$-test achieved good detection on the IntFl (Fig 5.24g) However, this is contrasted with the IntFSt image in which the $L$-test performed poorly (Fig 5.24s). This is the

result of duplicated surface points. The objective analysis also supports this with the $\chi^2$, *RRO* and *DoB* methods performed strongly.

Fig 5.25f,g indicates that the $L$ and $F$ tests offer the best detection on the TexCSt image, followed by the $\chi^2$ and $KS$ methods (Fig 5.25i,j). The poorest performing are the baseline methods and the statistical $u$-test, $t$-test and the *RRO* tests which were unsuccessful at resolving the interface(Fig 5.25h,k,l). F-measure analysis indicates that the $KS$ and the $\chi^2$ outperform the $F$ and $L$ test methods, with the non-parametric tests achieving higher peak F1-scores but with a narrower threshold range (Fig 5.28b). The results for TexFSt were similar to TexCSt, but the gap in peak performance between the non-parametric $KS$ and $\chi^2$ is even greater (Fig 5.28f).

Visually, the strongest performing methods on the Com1CSt image were the $\chi^2$, Canny, *DoB* and $t$-test methods (Fig. 5.26c,e,i,j). These accurately resolved the complete surface, while producing very few spurious responses, in the case of the Canny and Difference of boxes methods, these over-detection artefacts were confined to the region with higher variance. While the Canny method and *DoB* method both resolved a strong surface interface, some of the detail has been lost in the case of the Canny, which presented with a 'smoothed' appearance, where as the *DoB* method despite having a larger scale parameter ($5 \times 5 \times 5$ compared with $3 \times 3 \times 3$) produced much sharper steps in the case of Com1CSt (Fig 5.26c,e) and much clearer delineation of the steps in the case of Com1FSt (Fig 5.26,o,q).

The $KS$, $u$-test and *RRO* order methods also successfully resolved the complete interface surface, however, these methods proved more sensitive, and more spurious surface points were produced (Fig. 5.26j,k,l). The methods which performed weakest were the Steerable filter method, and the $F$ and $L$ test methods. Here the steerable filter produced a duplicated surface response with some inaccuracies, with further artefacts present at the corners (edges) of the steps (Fig. 5.26d). Additional over-detection artefacts were also present, this can be clearly seen in Fig. 5.26d in the high variance, lower portion of the output. The $F$ test method also produced a duplicated boundary, both of which are displaced to either side of the correct surface location (Fig. 5.26f), some spurious responses were also present throughout the surface map volume. The $L$ method did produce an output without general over-detection artefacts throughout the image volume, however it did present with some detection artefacts at the corners of the interface (Fig. 5.26g).

In the case of Com1FSt, visually the strongest performing methods were the statistical $\chi^2$ and $t$-test methods. These methods were able to resolve the surface and maintain some of the detail of the 'small steps' (Fig. 5.26t,u), while producing a complete con-

nected surface. However, these methods did present with some over-detection artefacts. The remaining non parametric statistical methods, all produced similar results on the Com1FSt image type, each of the $KS$, $u$-test and $RRO$ methods accurately detected the complete interface, however, there were some spurious responses throughout the image volume which can be categorised as over-detection artefacts (Fig 5.26v,w,x). The $L$-test did not suffer from over-detection artefacts, however there were some inaccuracies along the surface interface, including displacement of the surface and multiple duplicated responses (Fig. 5.26s). The $F$-test method again also presented with a duplicated surface (Fig. 5.26r).

The results of Com2CSt and Com2FSt are presented in Fig 5.27. Visually, the best performing method on the Com2CSt interface is the $\chi^2$ statistical detection. This detection method achieved an accurate detection of the complete boundary, with fewer spurious responses compared with the other tests (Fig 5.27i,u). However the strength of the surface did vary, this can be seen in Fig. 5.27i), which manifests itself as different intensities along the surface. However, due to the high signal to noise ratio of the surface interface to the background, the weaker parts of the surface are preserved during hysteresis thresholding. This finding was supported by the F-measure analysis with the$\chi^2$-test achieving the highest peak F1-score (0.95).

The $KS$, $u$-test and $RRO$ methods accurately resolved the complete boundary however more over-detection artefacts were present when compared against the $\chi^2$ and $t$-test detection methods (Fig 5.27j,k,l). The Canny and $DoB$ method also resolved the surface accurately, but there was regional over-detection artefacts which can be seen in clearly in the cross-sectional upper left regions of the images in Fig 5.27c,e. The Steerable filter failed to differentiate the surface, from spurious responses (Fig. 5.27d), while the $L$-test result presented with an incomplete surface at corner edges as did the $F$-test method (Fig. 5.26f,g).

Visually the best performing method for Com2FSt is also the $\chi^2$ method. However the loss of detail is more prevalent in Com2FSt, when compared against Com1FSt (Fig 5.26i,u and 5.27i,u). This is due to the regions being more statistically similar, thus a weaker surface response where the gradient orientation is not as well defined by the filter. The Canny method resolved the surface, but with some loss of detail due to Gaussian filtering. The Fisher method does appear to preserve the finer detail of the interface in superior fashion when compared against the other methods, however this surface is displaced, and there are some additional artefacts along the surface perimeter (Fig. 5.27r). The remaining non parametric tests again produced similar results, resolving the surface interface with some loss of detail, and some over-detection artefacts. The steerable method failed to resolve the surface (Fig 5.26p)

Objective results indicate that the best method for the large stepped image volume is the $\chi^2$ method while the $F$-test and Steerable are the poorest performing. The $KS$ scored best for the small stepped volume and the $F$ and $L$ methods performed the poorest (Fig. 5.28c,g).

### 5.6.5 Summary of Evaluation 2

Overall evaluation 2 aims to incorporate the effects of interface topology into the assessment of the various filter methods, and evaluate the filters quantitatively against ground truth solutions. Evaluation 2 reveals that the impact of topology is less compared with the impact of the statistical properties of the distributions that define the regions. However the topological differences are not inconsequential, a number of artefacts were caused by the different topologies, Notably the $t$- test which performed to a lower standard in localised areas such as at the corners and junctions of the interfaces, however this was due to high magnitude spikes in regions of low variance when the neighbourhood mask was completely aligned with the surface in one of the dimensional planes. These 'spikes' reduced the uniformity of the surface magnitude, which made the results less reliable with respect to thresholding.

The results from the cuboid(Cu) interface indicate similar findings to Evaluation 1. With the $\chi^2$ and $KS$ filters being the most highly rated, due to their success on all interface types and relatively wide threshold range for optimal performance. The cuboid interface style was primarily used since it has a 3-D topology which is not adequately resolved using 2-D filters and surfaces in the $z$-plane will be missed. The introduction of interface edges and corners in contrast to Evaluation 1 did create some conditions which proved to be more challenging for some of the filter methods. In general the $KS$ test achieved the best detection, however it was susceptible to over-detection, while the $\chi^2$ method offered detection only slightly behind the $KS$ method, while not quite suffering to same degree in terms of the production of spurious responses. A cross section examination of the surface maps indicate that the $L$ and $F$ tests can achieve good detection on texture interfaces and achieve good noise suppression, however they are susceptible to displacement issues as well as duplicated surface points.

The spherical interface volume introduces interfaces with a high degree of curvature in relation the filter neighbourhood size, here Gaussian smoothing of the Canny method was shown to be effective in improving the uniformity of the surface magnitude. But with the Staircase volumes, Gaussian filtering had a negative impact, by removing or smoothing the finer details of the topological structure.

#### 5.6.5.1 Characteristics (Cu, Sp, CSt, FSt)

Tables with summaries of the key characteristics are provided from page 157 to 160.

| Method | IntCu | TexCu | Com1Cu | Com2Cu |
|---|---|---|---|---|
| Canny | Accurate detection, connected surface, no over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| Steerable | Displaced surface, connected surface, localised inaccuracies at corners, some over-detection artefacts | No surface, over-detection artefacts | Displaced surface, connected surface, some over-detection artefacts | No resolved surface, over-detection artefacts |
| $\chi^2$ | Accurate detection, connected surface, no over-detection artefacts | Accurate detection detection, some discontinuities, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts, weaker corners | Accurate detection, connected surface, some over-detection artefacts |
| $DoB$ | Accurate detection, connected surface, no over-detection artefacts | No surface, over-detection artefact | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $F$-test | Missed surface, duplicated boundary either side, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Inaccurate surface position, duplicated surface, no general over-detection artefacts | Displaced graduated surface |
| $KS$ | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $L$ | Accurate detection, connected surface, some localised corner and edge artefacts, localised over-detection near surface | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, duplicated surface, no general over-detection artefacts | Displaced surface response, connected surface, some over-detection artefacts |
| $u$-test | Accurate detection, connected surface, some over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $RRO$ | Accurate detection, connected surface, some over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts, weaker corners | Accurate detection, connected surface, some over-detection artefacts |
| $t$-test | Accurate detection detection, weak corner resolution, no over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts, weaker corners and edges | Accurate detection, connected surface, some over-detection artefacts |

Table 5.7: Evaluation 2, Topology 1. Summary of Characteristics for Cuboid (Cu) interfaces

| Method | IntSp | TexSp | Com1Sp | Com2Sp |
|---|---|---|---|---|
| Canny | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| Steerable | Small displacement, connected surface, some localised artefacts near surface and some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, poor surface resolution, over-detection artefacts | poor differentiation between surface and over-detection artefacts |
| $\chi^2$ | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, some discontinuities in surface strength, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $DoB$ | Accurate detection, connected surface, no over-detection artefacts | No surface resolved, some over-detection artefacts | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface some over-detection artefacts |
| $F$-test | Duplicated surface either side of interface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Duplication of boundary, connected surface, some over-detection artefacts | Displacement of boundary, some over-detection artefacts |
| $KS$ | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $L$ | Duplicated surface, artefacts near interface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Displacement of boundary, connected surface, some over-detection artefacts | Displacement of boundary, some over-detection artefacts |
| $u$-test | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $RRO$ | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| $t$-test | Accurate detection, unconnected surface, no over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, some varying of surface strength, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |

Table 5.8: Evaluation 2, Topology 2. Summary of Characteristics for Spherical (Sp) interfaces

| Method | IntCSt | TexCSt | Com1CSt | Com2CSt |
|---|---|---|---|---|
| Canny | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some regional over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| Steerable | Displaced surface, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | connected surface surface,but some duplicates and positional inaccuracies, over-detection artefacts | Inaccurate surface, connected surface, poor differentiation from over-detection artefacts |
| $\chi^2$ | Accurate detection, connected surface some over-detection artefacts | Accurate detection, some discontinuities, some over-detection artefacts | Accurate detection, connected surface some over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts, weaker strength corners |
| $DoB$ | Accurate detection, connected surface, no over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some regional over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| $F$-test | Duplicated surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Incomplete detection, duplicate responses, over-detection artefacts | Disconnected surface surface with single voxel displacement, some over-detection artefacts |
| $KS$ | Accurate detection, connected surface some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| $L$ | Accurate detection, connected surface some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Incomplete detection, duplicate responses, over-detection artefacts | Disconnected surface surface with single voxel displacement, some over-detection artefacts |
| $u$-test | Accurate detection, connected surface some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| $RRO$ | Accurate detection, connected surface some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts, weaker corner strength | Accurate detection,connected surface. some over-detection artefacts |
| $t$-test | Accurate detection, connected surface, no over-detection artefacts, weaker strength corners | No surface resolved, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts, variable surface strength | Accurate detection,connected surface. some over-detection artefacts |

Table 5.9: Evaluation 2, Topology 3. Summary of Characteristics for Staircase (CSt) interfaces

| Method | IntFSt | TexFSt | Com1FSt | Com2FSt |
|---|---|---|---|---|
| Canny | Surface resolved, connected surface, some inaccuracies, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |
| Steerable | Surface resolved, connected surface, some inaccuracies, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved, inaccurate location, duplicated surface response, over-detection artefacts | Surface resolved with finer detail lost and slight displacement, over-detection artefacts |
| $\chi^2$ | Accurate detection, connected surface, some over-detection artefacts | Surface detected with some inaccuracies, surface discontinuities, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |
| $DoB$ | Accurate detection, connected surface, very few spurious surfaces | No surface resolved, over-detection artefacts | Surface resolved and connected with some loss of detail, regional over-detection artefacts | Surface resolved with finer detail lost some regional over-detection artefacts |
| $F$-test | Duplicate inaccurate surfaces, some over-detection artefacts | Surface resolved, localised artefacts, some additional over-detection artefacts | Inaccurate duplicated surfaces, some over-detection artefacts | Surface accurately resolved with small shift, some over-detection artefacts |
| $KS$ | Accurate detection, connected surface, some over-detection artefacts | Surface detected with some inaccuracies, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| $L$ | Duplicate inaccurate surfaces, very few spurious surfaces | Surface resolved, localised artefacts, some additional over-detection artefacts | Surface resolved, some displacement, very few spurious surfaces | Surface resolves with displacement, very few spurious responses |
| $u$-test | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved and connected with some loss of detail, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |
| $RRO$ | Surface resolved, connected surface, some inaccuracies, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved and connected with some loss of detail, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |
| $t$-test | Accurate detection, connected surface, localised artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, varied surface strength, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |

Table 5.10: Evaluation 2, Topology 3. Summary of Characteristics for Staircase (FSt) interfaces

## 5.7 Evaluation 3: Multiple textured interfaces

### 5.7.1 valuation 3: Aim

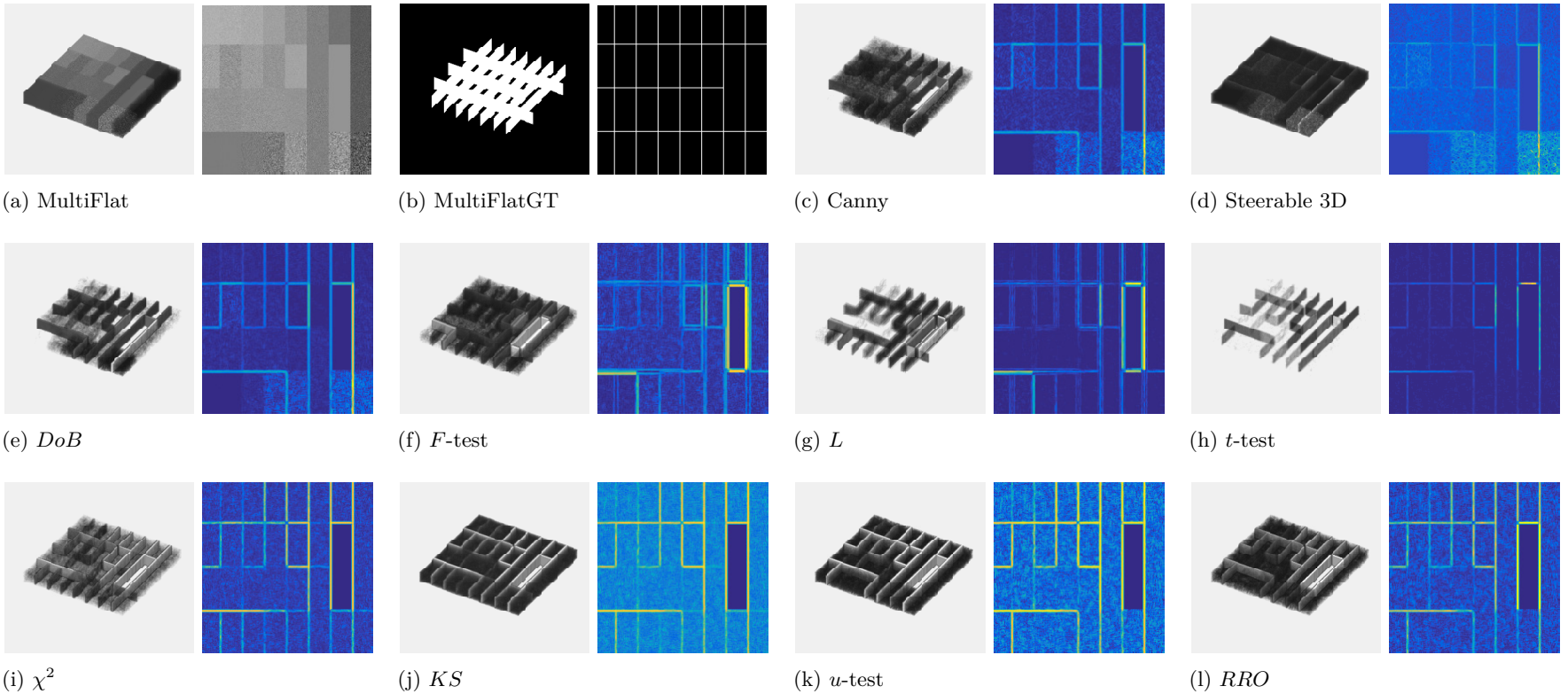
Evaluation 1 and 2 revealed that the statistical properties of a region have a more significant impact then the topological effects on the quality of the detected surface. Evaluation 3 aims to explore the impact of different region properties further. Here the filters are applied to data with a wide range of different regions which produce interfaces of different strengths. This creates less than optimal conditions for complete detection of a surface, and more closely approximates real image volume data which is inherently complex (Milan et al., 1993).

Evaluations 1 and 2 image sets are designed to analyse the ability of the filters to resolve a single surface between two controlled regions. However, real images seldom contain one interface. Image artefacts and object variation leads to the existence of multiple interfaces. Thus an image volume containing multiple region profiles, with several different interface types in the same image is more representative of real world imagery.

While many interfaces are present in an image volume, typically not all interfaces will be of the same strength due to the complexity of the textures and features that define particular regions. Thus, an effective synthetic test volume requires a multitude of regions capable of producing a range of interfaces with differing strengths. Gradient method operators and some parametric statistical tests are linear in their resolving power, meaning the weaker interfaces, if detected, should produce low output values, whereas stronger interfaces should result in higher outputs simultaneously in the same surface map. This may be of preference if the prioritised surface segmentation is of the stronger interface variety, however, if the weaker interfaces are critical surfaces required for a higher level process such as segmentation, these may be lost during post processing stages such as thresholding. Where this is the case, a method in which the local maxima could be normalised would be preferential.

The non parametric statistical tests which employ ranking of the sample data in essence locally normalise the output values while the image is processed. This local normalisation of output values has the effect of increasing the strength of weak surface responses to levels more often associated with stronger interface allowing for the detection of weaker boundaries. This aspect is referred to as sensitivity, and is assessed here.

### 5.7.2 Evaluation 3: Visual Results (MultiFlat, MultiCurve)

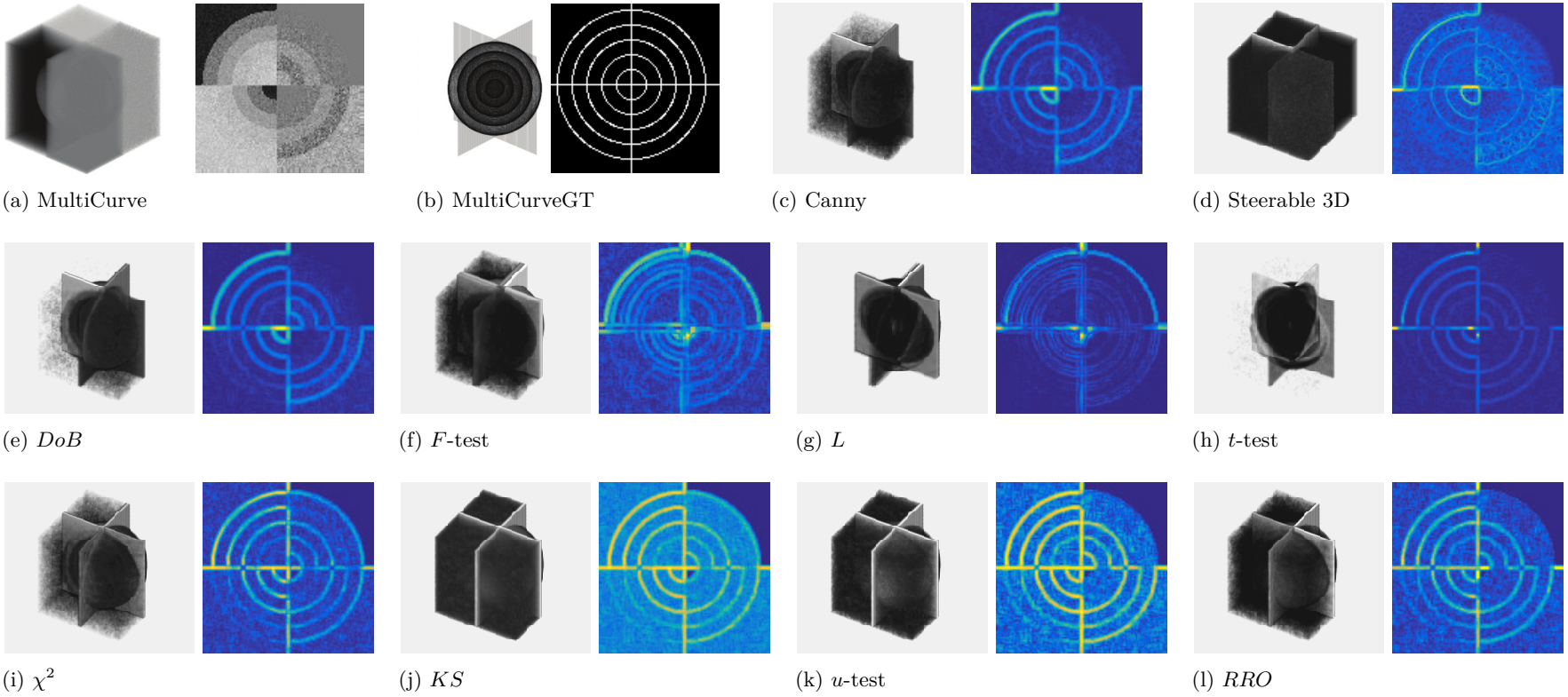


(a) MultiFlat (b) MultiFlatGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) $F$-test (g) $L$ (h) $t$-test

(i) $\chi^2$ (j) $KS$ (k) $u$-test (l) *RRO*

Figure 5.29: Multiple flat interfaces image (MultiFlat) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.



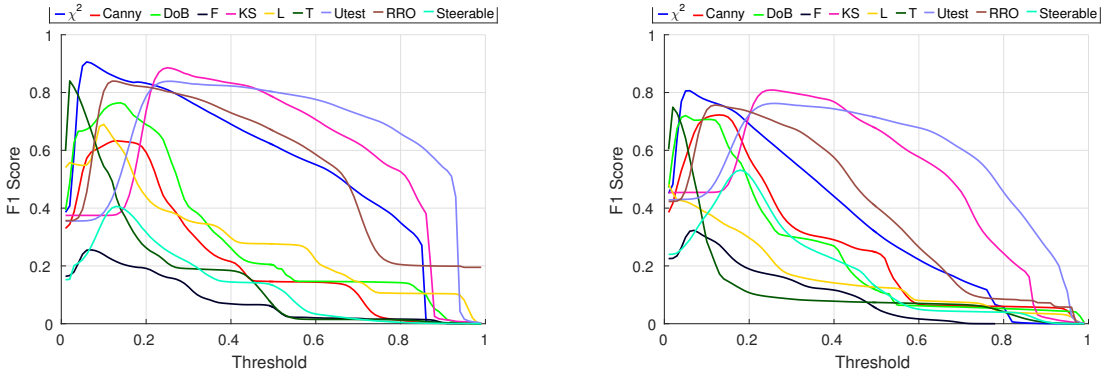
(a) MultiCurve (b) MultiCurveGT (c) Canny (d) Steerable 3D

(e) *DoB* (f) *F*-test (g) *L* (h) *t*-test

(i) $\chi^2$ (j) *KS* (k) *u*-test (l) *RRO*

Figure 5.30: Multiple curved interfaces image (MultiCurve) with corresponding surface map results from each statistical test method and control method. In addition the central layer from the surface map is presented to show a cross-sectional representation of the result.

### 5.7.3 F-Measure Results (MultiFlat, MultiCurve)




(a) MultiFlat

(b) MultiCurve

Figure 5.31: F-Measure analysis of Multiple Interface (MultiFlat, MultiCurve) tests. Results are gathered from a F1-score comparison between a normalised NMS filter result at 100 different hysteresis threshold values (lower threshold at 40% of upper) with that of the reference image.

### 5.7.4 Discussion

The 3-D surface map visualisations indicate that $t$-test is able to achieve good detection with the fewest instances of spurious surface points (Fig. 5.29). However, examination of the 2-D layers reveal that while the $t$-test is a good suppressor of image noise, it does produce a surface map response with a high dynamic range across the different interfaces, this is signified by certain resolved interface types possessing a significantly different magnitude response to others, leading to some of the detected interfaces not being visible. This can lead to a narrow operating window for optimal threshold selection, this narrow effective threshold range is observed Fig 5.31, here it can be seen that the $t$-test was effective only with a very low threshold.

The superior method is the $\chi^2$ method, offering good accurate detection across a wide range of interface types with a moderately low number of spurious responses on both MultiFlat and MultiCurve datasets. Further still, analysis of the individual 2-D layers reveal that the relative magnitude between the resolved surface of the interface and the magnitude of the spurious response is quite significant, allowing for more optimal thresholding. The $\chi^2$ method also has a small dynamic range in its response across the different interface locations. This is due to the nature of the test effectively offering local normalisation and this results in surfaces which have a locally strong magnitude. This makes the $\chi^2$ method not only suitable for resolving many different kinds of surfaces at different interface strengths but is also able to achieve success across multiple interface types concurrently in the same image volume. Visually the $\chi^2$ method is the most

optimal, this is further supported by the quantitative analysis which shows the $\chi^2$ method to offer the best results, followed closely by the $KS$ test (Fig. 5.31).

The $KS$-test performed to a similar standard as the $\chi^2$ method, however the magnitude of the spurious responses relative to the surface interface in this instance is higher. The next strongest performing filters were the $RRO$, $t$-test and $u$-tests. With the $RRO$ and $u$-test methods offering similar performance to the $KS$ method in terms strong accurate connected detection, and the presence of spurious surface points. The $DoB$ method offered a comparable response to the Canny method, but with marginally improved noise suppression. While the poorest performing tests were the Steerable, $L$ and $F$ test methods. Results from the MultiCurve test image offer a comparable findings to the MultiFlat test. The $\chi^2$ method again produced the best visual surface map results indicated by offering good detection of the surface at a lower noise cost when compared with the other non parametric techniques. In addition, the $\chi^2$ method detected the large exterior interface in the upper right quadrant of the image, which all other methods failed to do (Fig. 5.30i). The effectiveness of the $\chi^2$ method was also reflected in the F-Measure scores, however the objective results also indicate that $KS$ method performed strongly, but at a higher threshold level, where more of the noise is reduced (Fig. 5.31b).

### 5.7.5 Summary of Evaluation 3

Evaluation 3 was designed to measure the performance of the filters in the presence of multiple different regions with different relative interface strengths. The best performing measures were the $\chi^2$ method and the $KS$ methods, which significantly outperformed the baseline Canny operator and 3-D steerable methods according to the F-measure scores (Fig 5.31). This provides a strong rationale to use statistical methods in complex data with multiple image regions. Since these methods use non-parametric statistical means to measure dissimilarity, they are able to achieve a degree of local normalisation, whereby a high output response for a local region, results in a high magnitude response in the context of a whole image volume. This is desirable since weaker interfaces, which may be indicative of a type of boundary being targeted, can be resolved without the introduction of significant amounts of spurious responses. The remaining two non parametric tests, $RRO$ and $u$-test methods, also achieve good detection across the MultiFlat and MultiCurve test images.

When preventing against over-detection is most significant, the $t$-test becomes a desirable option since it does not produce as much noise as the other methods tested, however this it at the expense of some undetected boundaries, namely in localised areas such and corners and junctions. The $DoB$ method is most comparable method to the Canny

method, while offering slightly better noise suppression. The Steerable method, $L$ and $F$ test methods did not perform well on the multiple interface volumes, the Steerable method presented with noise and significant over-detection, while the $L$-test produced a strong result for one interface type, and this led to the suppression of weaker interfaces within the volume. The $F$-test detects the boundaries and suppresses noise, but the results produced duplicated and displaced boundaries which violates one of Canny's criteria for optimal detection.

| Method | MuliFlat | MultiCurve |
|---|---|---|
| Canny | Missed surfaces, some noise, good connectivity, | Missed surfaces, good noise suppression, low magnitude surfaces |
| Steerable | Missed surfaces, considerable noise, low magnitude surfaces | Poor detection, considerable noise, weak surfaces |
| $\chi^2$ | Best detection, some noise, good connectivity, strong magnitude surfaces | Best detection, some noise, good connectivity, strong surface magnitudes |
| $DoB$ | Some missed surfaces, low levels of noise, good connectivity | Missed surfaces, good noise suppression, low magnitude surfaces |
| $F$-test | Good detection, but duplicated surfaces, some noise, good connectivity | Good detection, but duplicated surfaces, some noise, good connectivity |
| $KS$ | Good detection, considerable noise, good connectivity, strong magnitude surface | Good detection, considerable noise, good connectivity, strong magnitude detection |
| $L$ | Good detection, but duplicated surfaces, good noise suppression, magnitude surfaces | moderate detection, but duplicated surfaces, moderate noise suppression, strong magnitude surfaces |
| $u$-test | Good detection, moderate noise, good connectivity, strong magnitude surfaces | Good detection, moderate noise, good connectivity, strong magnitude surfaces |
| $RRO$ | Good detection, considerable noise, good connectivity, strong magnitude surfaces | Good detection, low-moderate noise, good connectivity, strong magnitude surfaces |
| $t$-test | Good detection, good noise suppression, good connectivity, mixed magnitude surfaces | Good detection, good noise suppression, good connectivity, low magnitude surfaces |

Table 5.11: Evaluation 3, Summary of Characteristics for MultiFlat and MultiCurve interfaces

## 5.8 Summary

In Chapter 5 an objective evaluation of statistical surface detection was presented using synthetically created data. The development of accurate synthetic data introduces many challenges due to the complexity of real images in terms of surface structure, and the wide array of possible region profiles and textures which can exist in 3-D imagery as noted by Bowyer et al. (2001). To address these challenges a multi-variable approach was undertaken where resolving power, surface topology and multi-region detection were assessed independently.

The objective analysis in Evaluation 1 indicated that the $\chi^2$ was the most reliable detection method, characterised by good detection across the most interface types, notably best on the Com1 and Com2 interface types, with strong connected surface results. The *DoB* and *t*-test performed well where there was a strong intensity component, and the $L$ and non-parametric test methods also performed relatively well where the variance component was significant.

Evaluation 2 introduced topological differences in the interface to assess the impact on surface detection methods. While topological differences were shown to be not as pronounced in the results as the difference in interface types, the effect was not negligible. The initial stage of the Canny surface detection process applies Gaussian filtering to assist in the suppression of noise in the output, this had both positive and negative influences on result based on the topology under assessment. The Gaussian filtering on the IntSp interface improved the uniformity of the curvature, combined with the small scale parameter of the method ($3 \times 3 \times 3$), the Canny surface filtering produced a response which had uniform magnitude, in turn this improved the robustness of the method with respect to hysteresis thresholding and allowed the Canny technique to achieve the best performance score (Fig 5.23a) on the IntSp image. Conversely the Gaussian filtering stage negatively affected the visual performance of the method on the other topologies, this is most apparent on the FSt topologies where the finer details of the interface structure were negated by the Gaussian filtering. This effect is apparent in Fig 5.24o where the 'steps' in the structure were effectively smoothed into a flat surface plane. This was not the case for the *DoB* method which is the statistical method that most closely resembles the Canny method in terms of its functionality. In Fig 5.24q it can be seen that the topological details are preserved to a greater extent than the Canny method. While these artefacts are clear in the visual results, the Canny method maintained a strong performance F1-score on the FSt image type due to proximity of the smoothed surface response to the reference ideal, falling within the $T$-match allowance. Here the statistical methods which had better visual results, did

not improve upon the Canny method performance (Fig 5.28e,g,h), this due to the fact NMS has limitations on corners, which are commonly suppressed, therefore the benefits of the superior filter responses of the statistical methods were not fully realised. This explanation of the discrepancy is confirmed by the results from the CSt image type, where the Canny method was outperformed by statistical methods, because the NMS was not as significant of a limiting factor (Fig 5.28a-d).

Evaluation 3 addressed the issue of multiple strength interfaces present in the same image which is more typical of real imagery. The MultiFlat and MultiCurve images provided the results with the biggest performance gap between the baseline techniques and the statistical techniques. Here statistical methods, particularly the non-parametric tests with their greater sensitivity achieved improved detection, with more interfaces detected, better connectivity of the interfaces and improved signal noise ratio (Fig 5.29 and 5.30). While the non parametric tests did produce more noise in the surface map, the signal to noise ratio was great enough that hysteresis thresholding allowed for successful results. The performance scores indicate that when multiple interfaces are present, it is better to use a sensitive filter such as non-parametric test, which has better detection at the cost of increased noise in order to detect the weaker interfaces in the image, since the noise can be reduced through hysteresis thresholding. This can be observed in Fig 5.31 where the sensitive non-parametric tests achieved the best scores, but in the lower threshold range.

Generally, the non parametric tests achieved the best detection across the evaluations, but at the cost of increased noise in the output. The parametric $DoB$ and $t$-test achieved better detection than the baseline methods, but with some topological limitations for the $t$-test. The remaining $L$ and $F$ test improved over baseline methods on the Tex interface type, however the $L$ and $F$ test did produce duplication artefacts and were not as successful as other methods on the Int, Com1 and Com2 interface types.

While efforts were made to improve the reliability of an analysis using synthetic data, it cannot be automatically assumed that a correlation exists between the results of synthetic data testing and performance on real imagery as noted by Bowyer et al. (2001). Following a confirmation and validation framework, the following chapter introduces a qualitative approach to performance evaluation in order to assess if the filter characteristics are transferable to a real data case.

# Chapter 6

# 3-D Surface Detection in Multi-Model MRI

## 6.1 Introduction

In Chapter 5, a supervised objective analysis using synthetically generated image data was employed to provide an assessment of surface detection performance. However, the complexity of real data contrasts that of synthetically generated images, implying that optimal performance in one domain, may not directly translate to another. Chen and Zhu (2019) states that both subjective and objective assessment are crucial for evaluating the visual performance of algorithms.

To verify whether the synthetic evaluation can be relied up on and has generalisability and transferability, a qualitative approach is undertaken analysing surface detection filtering of real Multi-Model MRI data. This allows the investigation of whether trends found with synthetic data apply to real data. The dataset consists of 5 paediatric patients with Pilocytic Astrocytoma(PA) tumours in the posterior fossa region of the brain. The analysis will examine how the surface detection methods respond to the various interfaces which exist in this region, across 3 independent MRI modalities, T1-weighted, T2-weighted and contrast enhanced T1-weighted data.

This case study will observe the characteristics of the filter results and compare the findings with results generated through synthetic data. The purpose of this analysis is to attempt to visually verify whether or not the characteristics of statistic surface detection methods discovered in Chapter 5 are transferable in a real world application case. The analysis informs which methods provide the best results, but also identify characteristics from each filter type which are likely to present across a wider range of applications. It will also give indication to what conclusions can be reliably be

drawn from synthetic data testing while identifying the potential limitations of such an approach.

## 6.2 Background

The World Health Organisation lists over 30 different classes of brain tumour, in addition brain tumours are listed as the second most common kind of cancer found in children. In the medical field there are a number of 3-D imaging techniques for which surface detection algorithms are applicable, these include scanning techniques such as magnetic resonance imaging (MRI), computed tomography (CT), single photon emission computed tomography (SPECT) and positron emission tomography (PET). Each of these methods provide their own unique utility and possess their own advantages and challenging problem areas. The standard technique for brain tumour diagnosis is to apply magnetic resonance imaging (DeAngelis, 2001; Wen et al., 2010).

A particular issue common to 3-D scanning methods is the ability to locate interfaces in the image where the boundary is not clearly defined by intensity differences alone. Frequently there is poor contrast between the tumour and its surroundings, this can be due to noise (Abdel-Gawad et al., 2020), but the lack of contrast can also occur if pathologies possess tissue properties similar to that of its surroundings.

### 6.2.1 Magnetic Resonance Imaging (MRI)

MRI is a non-invasive imaging technique that is able to produce a 3-D image volume that can reveal various anatomical structures of a patient (Bauer et al., 2013). The MRI process requires the patient to be placed in a strong magnetic field. This realigns randomly orientated hydrogen atomic nuclei in water molecules present in the patient's tissue with that of the static magnetic field. A radio frequency pulse is emitted from the scanner to excite the hydrogen nuclei, forcing some nuclei to be 180° out of alignment with the static magnetic field and in phase with the surrounding hydrogen nuclei. To obtain a signal for imaging, the magnetic field is disrupted through a specific sequence of radio frequency pulses, causing the atomic nuclei to return or relax to their normal state. Through this process of relaxation, the change in the local magnetic field induces an electrical current in the receiver coils of the machine, for which the voltage can be measured over time to produce a signal. This signal is then processed using a Fourier transformation to evaluate the intensity of RF energy in a specific location. A map of the processed RF energy is used to construct an image, bright regions equate to high energy and dark regions correspond to low energy and these structures can correspond to different tissue types within the image.

### 6.2.2 Multi-Modal MRI

Adjustment of the pulse sequences allows for different imaging modalities, namely T1-weighted and T2-weighted formats, this distinction allows for different structures contained within the body to be revealed in a non-invasive fashion (Bauer et al., 2013). Multi-modal techniques (MRI scans using multiple modalities T1, T2 and contrast enhanced modalities) is the standard approach to a tumour diagnosis (Bauer et al., 2013). However, MRI machines are not without their drawbacks, for instance, a patient must remain still in an enclosed machine for an extended period of time which may be a problem for claustrophobic patients (Murphy and Brunberg, 1997; Dewey et al., 2007). As some patients find this experience stressful, the duration required of the patient spent in the scanner is often minimised. As a consequence the patient data acquired from MRI is not always the highest resolution achievable by the MRI scanner (Du et al., 2020). Additionally, multi-modal scanning requires further time for the patient in the scanning machine, therefore single modality analysis is sometimes undertaken. Thus, medical practitioners do not always have access to the most complete data. One aim of the evaluation is to assess the performance of the surface detection filters on individual imaging modalities, in order to make recommendations for each domain.

## 6.3 Evaluation Methods

There are two distinct approaches to evaluating surface detection in an application specific domain, one employs the use of model data, and the other uses real data. The first of these approaches further expands on the objective analysis presented in Chapter 5, with increased development of the synthetically generated images. In this approach the aim is to replicate real data as close as possible for use with a specific application case. This requires the inclusion of more complex topological structures, and the development of more complex textures. The most recognised approach for simulating MRI data is to use MR phantom images, such as the BrainWeb data set (Collins et al., 1998). These simulated models have a number of advantages. Firstly they are created using a mathematically defined model which allows for the creation of repeatable reference images which are valid for 3-D data. Secondly, models and reference images can be created for individual modalities, such as T1-weighted, T2-weighted and T1-contrast enhanced MRI images allowing for objective analysis of algorithms to be performed on single modalities separate from one another. The current BrainWeb models, are considered the benchmark of simulated 3-D MRI data for the analysis of different segmentation methods (Despotović et al., 2015). However, simulated phantom models introduce some issues. Firstly the data is still not from a real application case, thus the reliability of the result is dependent on the accuracy of the model. However, more importantly when

considering Phantom MR for surface detection evaluation, is that this type of dataset is modelled specifically for the segmentation of specific tissue types, such as white-matter, grey-matter and cerebral spinal fluid regions. Simulations of other brain structure or specific pathologies are not available, such as types of brain tumours which have their own spatial and statistical characteristics.

The second approach is to use real data, which has been manually segmented. The current gold standard method for obtaining reference images for real data is the methodology established in the BRATS multimodal brain tumour segmentation challenge (Menze et al., 2015). The validation of reference material is as follows:

Firstly pre-segmentation is applied to remove all material which is not part of the brain, such as skull, nasal and optic regions, and other soft tissues. Then all imaging datasets are segmented manually by one to four 'raters' following an annotation protocol as described by Menze et al. (2015), the BRATS data includes T1-weighted, T2-weighted, T1-weighted contrast enhanced, and T2-weighted FLAIR imagery. The lower resolution images are interpolated to have the same resolution as the T1-c image modality which are then then co-registered ready for segmentation. Manual segmentation of a particular pathology is performed on a single modality, with data obtained from the other modalities used as supplementary resource to assist in the segmentation decision making process. This process requires viewing a combination of T1-weighted, T2-weighted image layers concurrently in order for the radiologist to make a more informed assessment of where object boundaries lie. The segmentation annotations are then approved by experienced radiologists. A huge number of segmentation methods have been objectively evaluated using this methodology with real data, and is therefore recognised framework for evaluating techniques using a real application case (Menze et al., 2018).

While suitable for general segmentation The BRATS challenge methodology is not directly transferable to objectively measuring the performance of edge and surface segmentation. This is due to 3 different factors.

- It is an objective analysis framework for region segmentation and object detection, not surface detection.
- Only 1 reference image is provided for 4 co-registered modalities.
- The reference image is comprised of manually segmented 2-D layers and is not true 3-D.

Segmentation is a category of image processing tasks, of which surface segmentation is considered a sub-category. General segmentation remains one of the most studied

problems in image analysis and computer vision (Fernandez et al., 2015). Wang et al. (2020) states that segmentation algorithms are abundant in the literature and that evaluation frameworks are continuously being proposed to meet these needs. There exists a number of evaluation frameworks to evaluating the effectiveness of segmentation methods (Cardoso and Corte-Real, 2005; Martin et al., 2006; Udupa et al., 2006; Zhang et al., 2008; Menze et al., 2015). However, these approaches are specifically tailored toward objectively analysing 2-D object segmentation, and are not directly applicable to surface segmentation.

The BRATS reference solutions are a labelled image which define only the structure which relates to the tumour interface. Therefore this methodology discriminates against other structures which may be relevant for a different segmentation task. It therefore effectively penalises surface detection algorithms for serving their primary function of detecting surfaces, based on an subjective assumption of which structure is relevant in the image. While this approach to creating reference solutions is suitable for object detection, and for aiding a subjective qualitative analysis, it is not suitable for objectively analysing surface detection for real data.

The second factor is that only a single reference image is provided unique to each case. This is important if the goal is assessing the evaluation of object detection, since the object's position is most accurately described by taking into account all available data. Surface detection algorithms have no prior image understanding; they are not intended to combine outputs from different image modalities in order to generate a single result. Using co-registered reference images does not allow for a valid objective analysis to be performed on each of the individual T1, T2 and contrast enhanced T1-weighted modalities independently. In order to accurately analyse the algorithms performance, a separate reference image for each modality would be required. Then a richer set of recommendations could be made for specific algorithms with respect to the image modality, and guide how parameters should be set and define best use for semi-automated practice.

Thirdly Bomans et al. (1990) showed how structure can exist when considering data in 3-D that would be missed when viewing the data in 2-D, this is illustrated earlier in Figure 2.19 and within the results presented in Chapter 5. An often overlooked issue of using manual segmentation for 3-D data is that manual segmentation is performed on individual 2-D axial layers, and ultimately neglects interfaces which exist between the layers. This causes the reference solutions for 3-D data to be flawed for an algorithm which employs 3-D processing. When manual segmentation annotations are made in 2-D, and then interpolated into 3-D volumes, some structure which exists across the z-plane may be missed (Bomans et al., 1990). Therefore manually segmented reference

images may contain incomplete information and are unsuitable for true 3-D objective examination.

As a consequence these factors also make the BRATS protocol an unsuitable methodology for objective surface detection evaluation.

## 6.4 Case Study: Pilocytic Astrocytoma Tumour and Cyst Interfaces

Surface detection algorithms are exclusively used to process 3-D data, and therefore to evaluate surface detection approaches in a real application, a case study with a 3-D data format is required. The purpose of this study is to provide context of where various surface detection algorithms are effective in different modalities of MRI data.

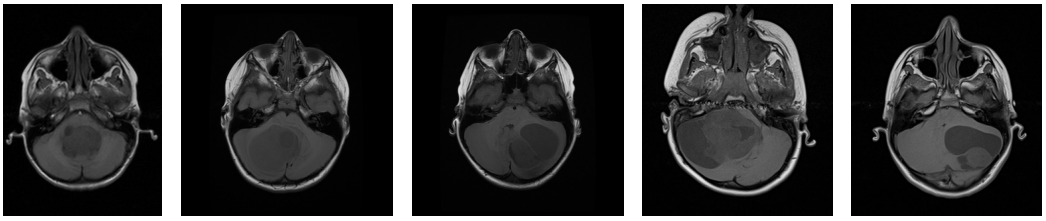
The case study selected for this work is the examining of surface detection methods for their ability to determine Pilocytic Astrocytoma tumour and cyst boundaries in paediatric patients. Unlike the adult population, the majority of childhood tumours present in the posterior fossa region of the brain, this is a sensitive region of the brain where critical and vital brain functions take place, making treatment challenging. Around 55% of childhood brain tumours arise in the posterior fossa, compared with 15% to 20% of adult tumours (Fetit, 2015). Accurate segmentation of these pathological structures is of paramount importance in order to avoid further complications in treatment planning. Analysis of the surface detection algorithms will be based on their utility at detecting tumour and cyst interfaces in 3-D data. In contrast to the methodology employed by the BRATS challenge (Menze et al., 2015), the effectiveness of detection methods on different imaging modalities in MRI (namely T1, T2 and T1 contrast enhanced) is considered. This approach is used as it allows for recommendations to be made of where specific algorithms are preferential.

## 6.5 MRI Datasets

The data provided in this study was made available by Birmingham Children's Hospital and the Children's Brain Tumour Research Group. The data is comprised of 5 Multimodal MRI Patient cases and are of paediatric patients with Pilocytic Astrocytoma brain tumours, and for each case, T1-weighted, T2-weighted and T1-contrast-enhanced MR imaging modalities were available. The complete dataset is provided in Appendices A,B,C.

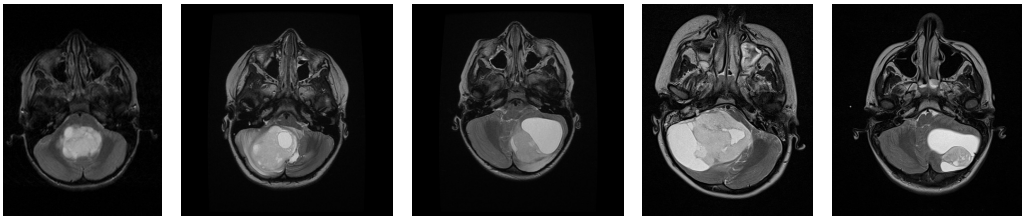
Example images of T1-weighted images are presented in figure 6.1, T2-weighted images in figure 6.2 and T1-weighted contrast enhanced images in 6.3. The images are in

their original formats and have not been interpolated or registered so they remain with the original acquired resolution. Nor has any pre-segmentation been applied, so the entire scan of the head region is available, and is not confined to a pre-segmented brain image, as seen in BRATS data. All data was anonymised by the clinical lead scientist at Birmingham Children's Hospital brain tumour research group for the study, and all data was processed following formal ethical clearance review process.



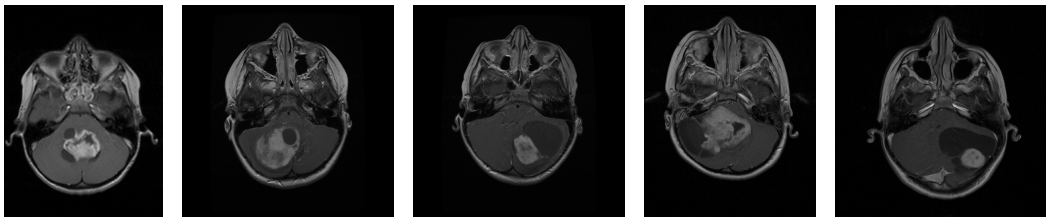
(a) Patient case 1 (b) Patient case 2 (c) Patient case 3 (d) Patient case 4 (e) Patient case 0

Figure 6.1: T1-weighted volumes with Pilocytic Astrocytoma tumours present in the posterior fossa region of the brain



(a) Patient case 1 (b) Patient case 2 (c) Patient case 3 (d) Patient case 4 (e) Patient case 5

Figure 6.2: T2-weighted volumes with Pilocytic Astrocytoma tumours present in the posterior fossa region of the brain



(a) Patient case 1 (b) Patient case 2 (c) Patient case 3 (d) Patient case 4 (e) Patient case 5

Figure 6.3: Contrast enhanced T1-weighted volumes with Pilocytic Astrocytoma tumours present in the posterior fossa region of the brain

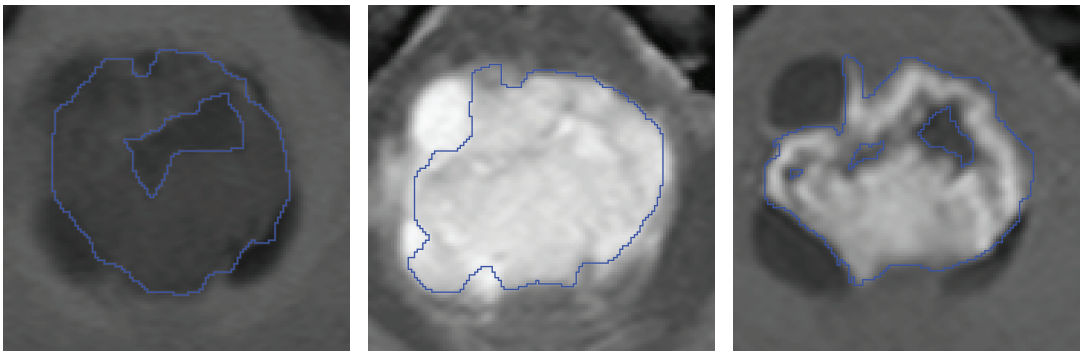
| T1 - weighted | | | |
|---|---|---|---|
| Cases | $x \times y \times z$ dimensions | $xy$-res (mm) | $z$-res (mm) |
| 1 | 384×512×30 | 0.45 | 5 |
| 2 | 512×512×25 | 0.45 | 5 |
| 3 | 512×512×24 | 0.45 | 5 |
| 4 | 288×320×39 | 0.63 | 4 |
| 5 | 288×320×40 | 0.72 | 4 |
| T2 - weighted | | | |
| Cases | $x \times y \times z$ dimensions | $xy$-res (mm) | $z$-res (mm) |
| 1 | 384×512×38 | 0.45 | 3 |
| 2 | 512×512×41 | 0.45 | 3 |
| 3 | 512×512×39 | 0.45 | 3 |
| 4 | 360×448×43 | 0.45 | 3 |
| 5 | 402×448×50 | 0.51 | 3 |
| T1 - weighted with Contrast agent | | | |
| Cases | $x \times y \times z$ dimensions | $xy$-res (mm) | $z$-res (mm) |
| 1 | 384×512×30 | 0.45 | 4 |
| 2 | 512×512×33 | 0.45 | 4 |
| 3 | 512×512×32 | 0.45 | 4 |
| 4 | 260×320×37 | 0.72 | 4 |
| 5 | 320×320×35 | 0.72 | 4 |

Table 6.1: Dataset dimensions and resolution.$x \times y \times z$ are the image dimensions measured in voxels. $x, y$ res is the measurement of the space covered by each voxel in axial plane in mm. $z$ res is voxel thickness also measured in mm. For each image volume the resolution is anisotropic with higher resolution in the $x$ and $y$ dimensions compared with that of the $z$ dimensions.

Due to the limitations of providing a repeatable objective analysis using real data, namely the problem of obtaining valid reference imagery for 3-D data and individual MRI modalities, an objective analysis of real data was not undertaken. Reference images can still be of useful assistance in a qualitative analysis. In a qualitative analysis reference images can act as visual guide to a region of interest and can indicate the location of where some of the tumour interfaces are expected to exist. This can assist in providing context for discussion of the results.

For this data, reference images were created manually in the 2-D axial plane by the clinical lead expert at Birmingham Children's Hospital and validated by the brain tumour research group separately for each of the T1-weighted, T2-weighted and T1-contrast-enhanced MR imaging modalities, allowing for a visual assessment to be made on results for each.

Upon observing the reference data for different modalities, it is apparent that different modalities provide unique image structure, further increasing the necessity of using separate reference images for each modality in an evaluation (Fig 6.4). As this analysis is visual and not objective, the annotations label only the boundary of solid tumour regions omitting regions of cyst from the reference image, these can then serve as a guide for regions of interest .



(a) T1 (b) T2 (c) T1-contrast

Figure 6.4: Reference : Annotations for T1, T2 and T1 contrast MRI volumes. The annotations indicate the interface of the tumour with other regions of the image. The annotations do not include brain regions such as grey matter, white matter, and cerebral spinal fluid, nor do the annotations indicate cyst regions. Notably the annotations are different for each of the modalities on a single patient

In Figure 6.4 the tumour regions are annotated for each of the modalities on a single patient. In Figure 6.4a, the T1 image, cysts are signified by the dark grey oval shaped regions on the perimeter of the tumour, while normal brain tissue is the lighter shaded

grey surrounding this region, the normal tissue has similar visual properties to the tumour, making manual segmentation of these types of interfaces non-trivial (Bauer et al., 2013). In Figure 6.4b, the T2 image, cysts are signified by the bright white regions on the perimeter of the tumour, and the normal brain tissue is signified by the darker grey surrounding region. Here the change in intensity between tumour and normal tissue is greater, however, visually differentiating between tumour and cyst is more challenging since they are both represented by high intensity regions. In Figure 6.4c, the T1-contrast image, cysts are signified by the dark grey oval shaped regions on the perimeter of the tumour, and the normal tissue is the brighter grey surrounding area. For T1-contrast images, the tumour region is higher intensity than both cyst and normal regions. Each independent modality offers a unique visualisation of the physical pathological structures within each of the patient cases, therefore surface detection applied separately to each modality is a valuable tool.

## 6.6 Qualitative Assessment

Subjective methods of evaluation can also be defined as observation methods, and these methodologies are dependent on a human assessor to visually inspect results (Wang et al., 2020). For this case study an observational methodology of analysis is employed, providing a visual assessment of the filter responses from the same 8 statistical surface detection and baseline methods analysed in Chapter 5.

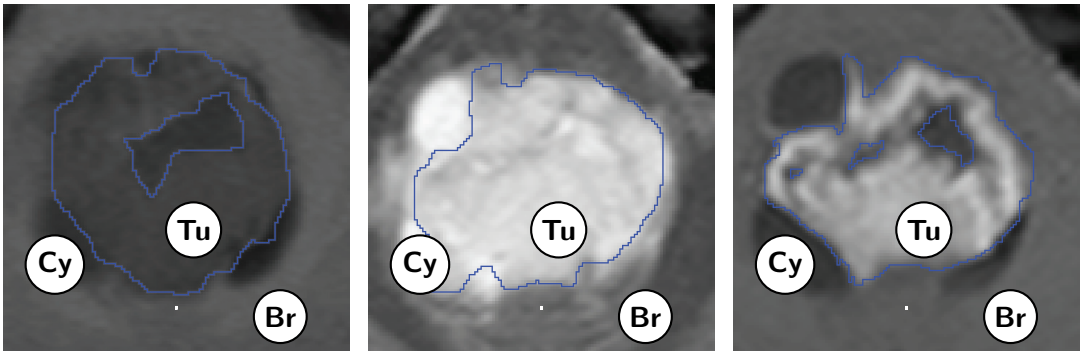
The surface detection methods are applied across 5 volumetric data sets, each of which possess data from three individual imaging modalities (T1-weighted, T2-weighted and T1-weighted contrast images).

In this case study the primary surfaces of interest are illustrated in Fig 6.5 and are which that exist between:

- Tumour (Tu) and Cyst (Cy) - $Tu - Cy$ interface.
- Tumour (Tu) and normal brain tissue (Br) - $Tu - Br$ interface.
- Cyst (Cy) and normal brain tissue (Br) - $Cy - Br$ interface.

The primary advantage of a subjective visual assessment is the ability to accurately describe the features and structure of the responses which is relevant to determining the context in which they can be applied. In this case a number of context specific characteristics shown in Chapter 5 are assessed in the results, namely:

- Detection of a surface interface.
- Strength of the detected surface relative to surroundings.




(a) T1 (b) T2 (c) T1-contrast

Figure 6.5: Tumour(Tu), Cyst(Cy) and Brain (Br) region labels

- Connectivity of the surface.
- The quantity of clutter (duplicated surface information)
- Prevalence of noise

These characteristics can be assessed visually, without the need to use a supervised comparison. These characteristics can also assist in the discussion of recommendations for a particular surface detection method. Attributes such as the connectivity of a detected surface, the amount of surface clutter (duplicated surface information) and the amount of noise in the image affects the suitability of the responses for specific tasks.

In alignment with chapter 5, the following same Vector Magnitude statistical methods are applied at a neighbourhood mask scale of $5 \times 5 \times 5$:

- Difference of boxes
- Fisher test
- Likelihood ratio test
- Student's $t$-test
- $\chi^2$ test
- Kolmogorov Smirnov test
- Mann-Whitney $u$-test
- Robust Rank Order test

All methods, including the baseline 3-D Canny surface detector and 3-D steerable methods were applied without the non maximum suppression and hysteresis thresholding stages. While these are fundamental components of the Canny and Steerable method, the results from Chapter 5 reveal different methods produce their optimal result in different threshold ranges. Within the dataset, as is typical for MRI data, not all

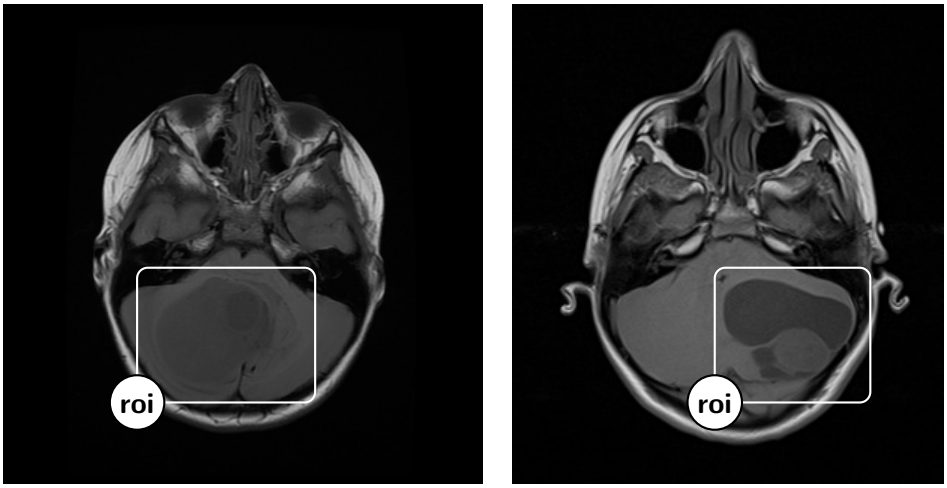
images across modalities have the same spatial dimensions or voxel spacing which leads to images with anisotropic resolution (Table 6.1). Information regarding the scanner settings is available within the DICOM files (MRI Image format) which is utilised by the VM method to adjust the dimensional weighting as defined earlier in Eq 3.13. In all patient cases the images have equal voxel spacing in the $x$ and $y$ dimensions of an axial view, but the $z$-spacing is larger by varying degrees.

Showing the visual results prior to NMS and Hysteresis thresholding stages allows for a direct comparison of the filter response, rather than subjectively selecting the optimal post-filtering stage parameters. Furthermore, the intensity of surfaces is useful in evaluation of subsequent processes and can offer guidance into the strength of the interface, mitigation of post processing techniques for this purpose is in alignment with the work of Williams et al. (2014). Image results cover a single layer from a region of interest situated in the tumour location. Presented with the results are 2-D manually segmented reference images for the tumour interface, this is to be used as a visual guide to indicate where a surface interface is expected.

## 6.7 Evaluation 4: T1-weighted 3-D

T1-weighted images are obtained with a short echo time and short repetition time sequence. In T1-weighted images of the brain, the white matter presents with bright regions, while conversely the grey matter presents as dark regions. Other structures such as cerebral spinal fluid (CSF) appear as dark regions, while the cortex (grey matter) appears grey, and fat contained within bone marrow appears bright (Fetit, 2015). PAs commonly present with one or more large cystic components (Fig 6.6b) which are generally composed of air, fluid and semi-solid material. In T1-weighted images, the tumour present as mildly hypo-intense(darker) regions with respect to surrounding white matter tissue, typically at a similar level to the CSF intensity while the cystic components present as hypo-intense regions (Fetit, 2015). This similarity in brightness to the surrounding tissue also makes differentiation between tumour pathology and healthy regions non-trivial when using gradient style methods. The similar intensity levels between the regions of the $Tu - Br$ interface share characteristics of the Tex type interface while the $Cy - Br$ and $Tu - Cy$ interfaces are characteristically similar to the Com2 interface type used in evaluations 1 and 2 in chapter 5. Topologically, the tumours interfaces are complex in shape, with small details and curvature, and are typically not uniformly flat. Therefore characteristics observed in evaluations 2.2 and 2.3 are explored in the real data. Additionally, the images are complex with several regions including those which are not part of the region of interest (ROI), as these regions are unique, the relative magnitude of the surface interfaces will be different

throughout the image. Evaluation 3 demonstrated a trade-off in performance between sensitivity and noise suppression when multiple different strength interfaces are present, with the non-parametric tests and the $t$-test achieving the best results (Fig 5.31). It is therefore expected that the trade-off observed in evaluation 3 would be present or elevated for real T1 imagery.




(a) (b)

Figure 6.6: Pilocytic Astrocytoma pathologies in T1 imagery

### 6.7.1 Results

Figures for the results are provided on the following page.

(a) Volume 1 (b) Reference 1 (c) Canny (d) Steerable (e) *DoB* (f) $F$-test

(g) $L$ (h) $t$-test (i) $\chi^2$ (j) $KS$ (k) *RRO* (l) $u$-test

(m) Volume 2 (n) Reference 2 (o) Canny (p) Steerable (q) *DoB* (r) $F$-test

(s) $L$ (t) $t$-test (u) $\chi^2$ (v) $KS$ (w) *RRO* (x) $u$-test

Figure 6.7: T1-weighted surface maps patient 1 and 2

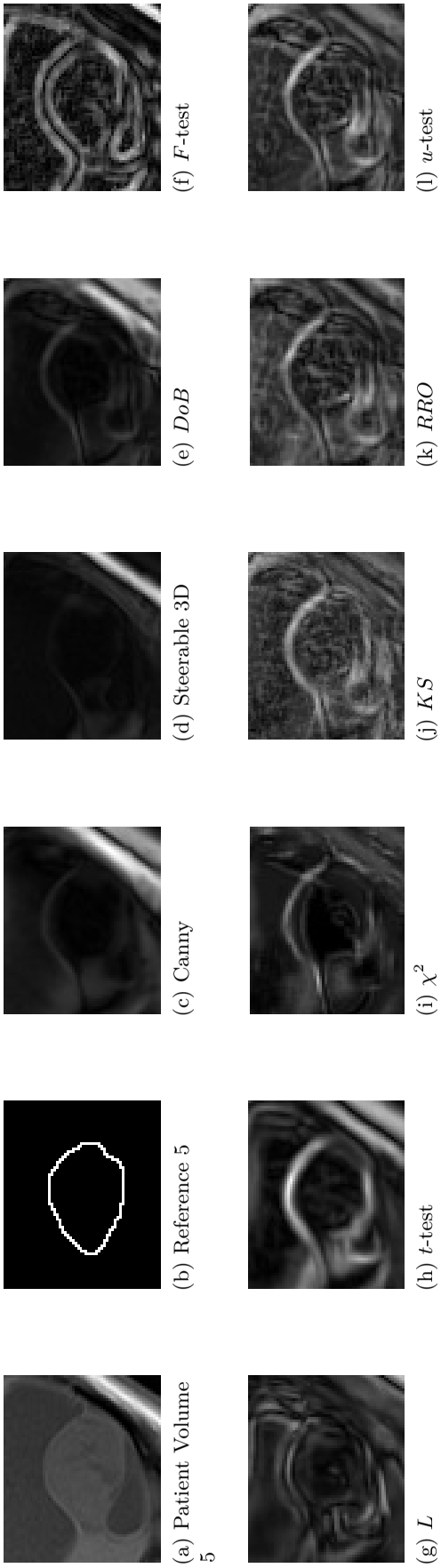
(a) Volume 3 (b) Reference 3 (c) Canny (d) Steerable (e) *DoB* (f) *F*-test

(g) $L$ (h) $t$-test (i) $\chi^2$ (j) $KS$ (k) $RRO$ (l) $u$-test

(m) Volume 4 (n) Reference 4 (o) Canny (p) Steerable (q) *DoB* (r) *F*-test

(s) $L$ (t) $t$-test (u) $\chi^2$ (v) $KS$ (w) $RRO$ (x) $u$-test

Figure 6.8: T1-weighted surface maps patient 3 and 4



(a) Patient Volume 5 (b) Reference 5 (c) Canny (d) Steerable 3D (e) *DoB* (f) $F$-test

(g) $L$ (h) $t$-test (i) $\chi^2$ (j) $KS$ (k) *RRO* (l) $u$-test

Figure 6.9: T1-weighted surface maps patient 5

### 6.7.2 Discussion

For the T1-weighted images, all methods detected interfaces surplus to those revealed in the reference image. All methods produced what can be considered spurious responses and clutter (connected spurious responses). However the best performing method was the Student's $t$-test, which achieved near complete accurate visual detection of the target interfaces in most cases. The $t$-test response was characterised by connected high magnitude surfaces across PA tumour cases, but this was most prominent with the case presented in Fig 6.7h, here the high magnitude interface is signified by a solid connected, bright surface interface. The $t$-test method on some occasions also resolves some internal structural details of the tumour (Fig 6.8h), which may be considered clutter, thus undesirable for various segmentation techniques. However, this can be mitigated with hysteresis thresholding.

The $\chi^2$ method also produced accurate detection with connected responses (Figures 6.7i 6.8u). The responses were visually less cluttered than the results from the other non-parametric tests ($KS$, $u$-test, $RRO$), while maintaining a similar level of detection of the tumour interface. Generally the $\chi^2$ method produced connected tumour interface surfaces, however there were some missing surface details as can be seen in Fig 6.7u between cyst and tumour as well as some localisation error incurred due to a displacement of the surface (Fig 6.9i). The cause of the displacement is due to surface interfaces which exist between adjacent image layers.

The Canny surface detector, Steerable filters and the statistical $DoB$ method are designed to produce high value responses to shifts in image intensity, T1-weighted images do not always produce a large intensity differentiation between solid tumour material and the surrounding white matter brain tissue, therefore these methods are not optimal with much of the tumour interface not always being detected (Figures 6.7o,p,q 6.8c,d,e,o,p,q). However, while the Steerable filter failed to resolve the tumour interface, the Canny and $DoB$ method can offer value since they typically produced a lower noise output, mostly absent of cluttered surfaces, and under some conditions produced good detection of the tumour interface (Fig 6.7c,).

The non-parametric $u$-test, $RRO$ and $KS$ tests offer stronger magnitude boundaries across most image volumes, with good differentiation between tumour and cyst regions. This characteristic is due to the local normalising effects of ranking the data. However a consequence of the added sensitivity is a significant increase of clutter and image noise (Figures6.7, 6.8j,k,l,v,w,x, 6.9j,k,l). The variance based $F$ and $L$ test methods both produced duplicated surfaces along the interface, and both presented responses

which contained clutter. However, the $F$-test typically produced a stronger magnitude and connected surface at the interface of the tumour (Fig 6.7f,g,r,s and 6.9f,g).

### 6.7.3 Key Findings

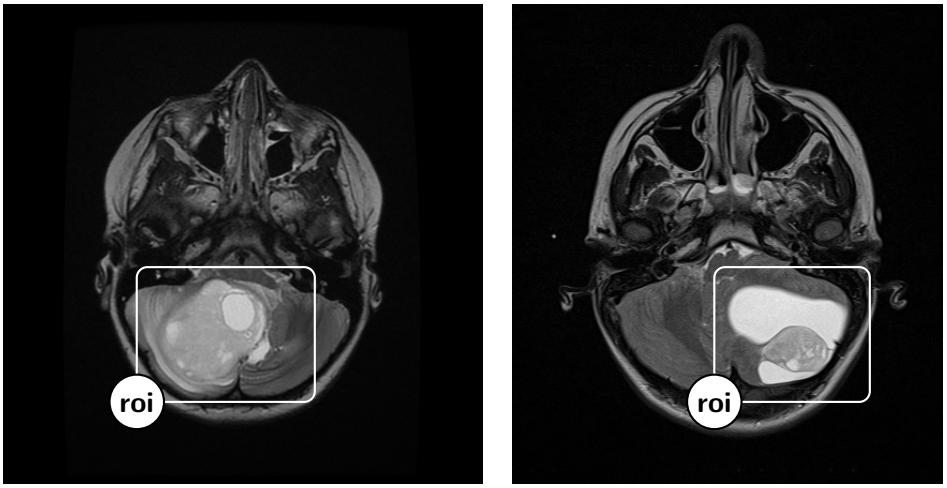
Table of the key findings is provided on the following page.

| Method | $Tu - Br$ | $Cy - Br$ | $Tu - Cy$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| $DoB$ | ✓ | ✓ | partial | low | Good detection, but weak interfaces |
| $F$ | ✓- duplicate | ✓- duplicate | ✓- duplicate | High | Duplicated surfaces |
| $L$ | × | × | × | mixed | Poor detection |
| $t$-test | ✓ | ✓ | ✓ | mixed | Good connectivity, strong surfaces |
| $\chi^2$ test | ✓ | ✓ | ✓ | Low | Good connectivity and noise suppression |
| $KS$ | ✓ | ✓ | ✓ | high | Good detection, but weak surfaces |
| $u$-test | ✓ | ✓ | ✓- partial | high | good detection but significant clutter |
| $RRO$ | ✓ | ✓ | × | high | partial detection, but cluttered |
| Canny | ✓-mixed | ✓-mixed | ✓-mixed | low | good detection but weak interface |
| Steerable | × | × | × | low | poor detection of axial interfaces |

Table 6.2: Surface filter characteristics on T1-weighted MRI scans. (✓) indicates detection of surface. Tu-Br is brain and tumour interface, Cy-Br is brain and cyst interface and Tu-Cy is cyst and tumour interface

## 6.8 Evaluation 5: T2-weighted 3-D

T2-weighted images are obtained using a relatively long echo time and repetition time sequence. For T2-weighted bran imagery, structures are revealed with different intensity profiles when compared with T1-weighted images. White matter presents as dark grey, and the cortex appears light grey. In T2-weighted images fluids appear as very bright regions including the CSF, while fat contained in bone presents as a light region(Fetit, 2015). PA solid mass tumours present as hyper-intense regions with better differentiation between solid tumour and white matter (Fetit, 2015). However, the large cystic components which are commonly associated with PA tumours also present as bright hyper-intense regions, making differentiation between solid tumour and cyst a non trivial problem for gradient methods of surface detection (Fig 6.10b). Characteristically, the Cy-Br interface has a strong intensity component similar to that of the Int type interface utilised in chapter 5, while the Tu-Cy has similar attributes to the Com1 interface type.




(a) (b)

Figure 6.10: Pilocytic Astrocytoma pathologies in T2 imagery

### 6.8.1 Results

Figures for the results are provided on the following page.

(a) Volume 1 (b) Reference 1 (c) Canny (d) Steerable (e) *DoB* (f) $F$-test

(g) $L$ (h) $t$-test (i) $\chi^2$ (j) $KS$ (k) $RRO$ (l) $u$-test

(m) Volume 2 (n) Reference 2 (o) Canny (p) Steerable (q) *DoB* (r) $F$-test

(s) $L$ (t) $t$-test (u) $\chi^2$ (v) $KS$ (w) $RRO$ (x) $u$-test

Figure 6.11: T2-weighted surface maps patient 1 and 2

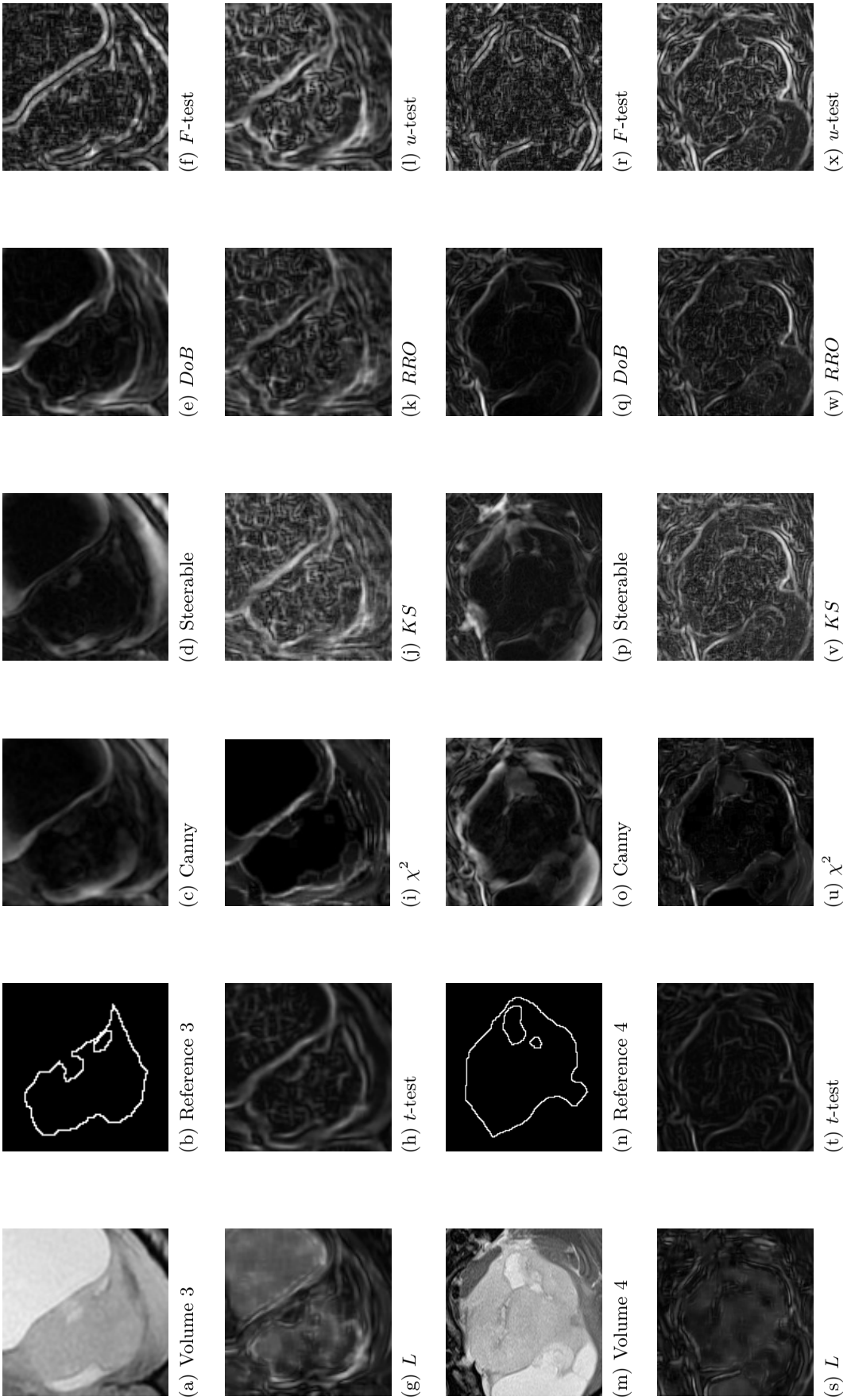
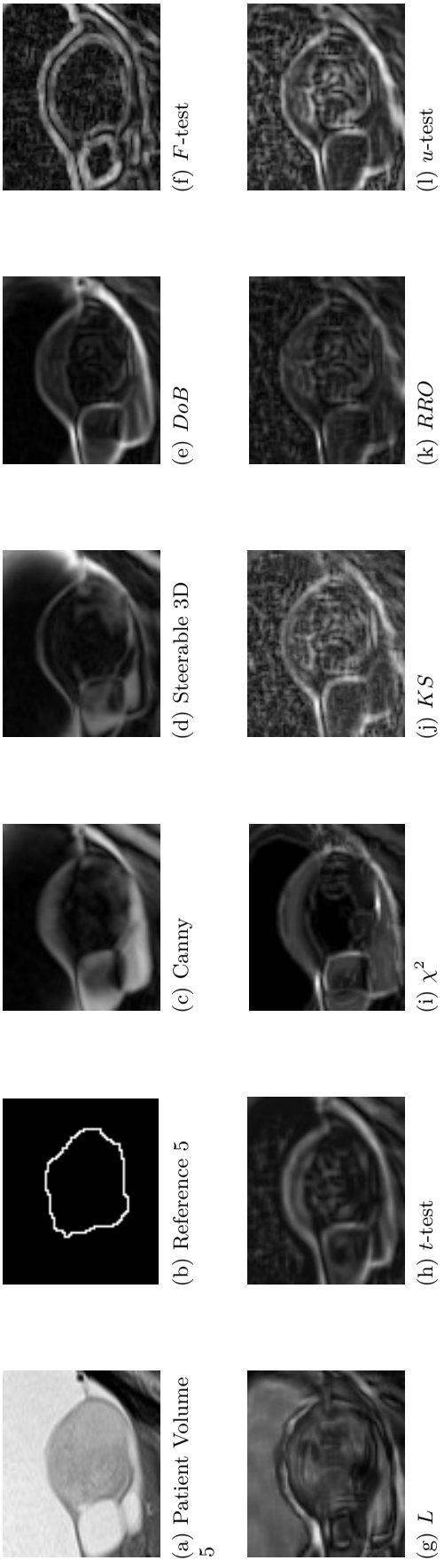



Figure 6.12: T2-weighted surface maps patient 3 and 4



(a) Patient Volume 5 (b) Reference 5 (c) Canny (d) Steerable 3D (e) *DoB* (f) $F$-test

(g) $L$ (h) $t$-test (i) $\chi^2$ (j) $KS$ (k) *RRO* (l) $u$-test

Figure 6.13: T2-weighted surface maps patient 5

### 6.8.2 Discussion

Overall, across the T2-weighted datasets the strongest performing method was the $t$-test. This was established by accurate, connected surfaces with a high magnitude response (Figures 6.11h,t, 6.12h,t and 6.13h). The $t$-test successfully resolved interfaces between each of the three target interface types, solid tumour and cyst, solid tumour white matter, as well as between cyst and white matter. Some clutter is also present, however this can be minimised by thresholding the responses.

The $\chi^2$ method also successfully resolved many of the tumour boundaries while maintaining relatively low levels of noise and clutter (Figures 6.11i,u, 6.12i,u and 6.13i). The resolved surfaces were of a strong magnitude for both interfaces between white matter and solid tumour as well as solid tumour and cysts. The interfaces presented with less fragmentation. The $DoB$ method produced clear responses in terms of image noise and clutter, and produced accurate detection of each of three interface types with good connectivity (Figures 6.11e,q, 6.12e,6.13e).

However, the $DoB$ method produced low intensity surfaces between some of the cyst and solid tumour interfaces (Fig 6.12q). The $RRO$, $u$-test and $KS$ test methods provided good detection of the tumour and cyst boundaries, however these responses were heavily cluttered due to the high sensitivity of the operators, with the $KS$ test being the slightly more cluttered than the $RRO$ and $u$-test methods (Figures 6.11, 6.12j,k,l,v,w,x, 6.13,j,k,l). The $F$ and $L$ test methods produced relatively cluttered surfaces interfaces, as well as duplicated responses (Figures 6.11, 6.12f,g,r,s, 6.13f,g).

The Steerable filter and Canny methods remained relatively noise and clutter free, and provided good detection of the cysts with good connectivity in some cases. Traditional methods remain a valid approach to T2-weighted images, but they are still outperformed by statistical approaches, and due to the fact they are impeded by resolution distortion this effect is clearly apparent in Figure 6.11c,d,e. Despite the Canny, Steerable and $DoB$ approaches all being techniques suited for intensity shifts, the $DoB$ method has much clearer defined surfaces. Additionally the baseline approaches are comparatively weak at detecting the tumour and brain interface when compared to the non-parametric statistical methods.

### 6.8.3 Key Findings

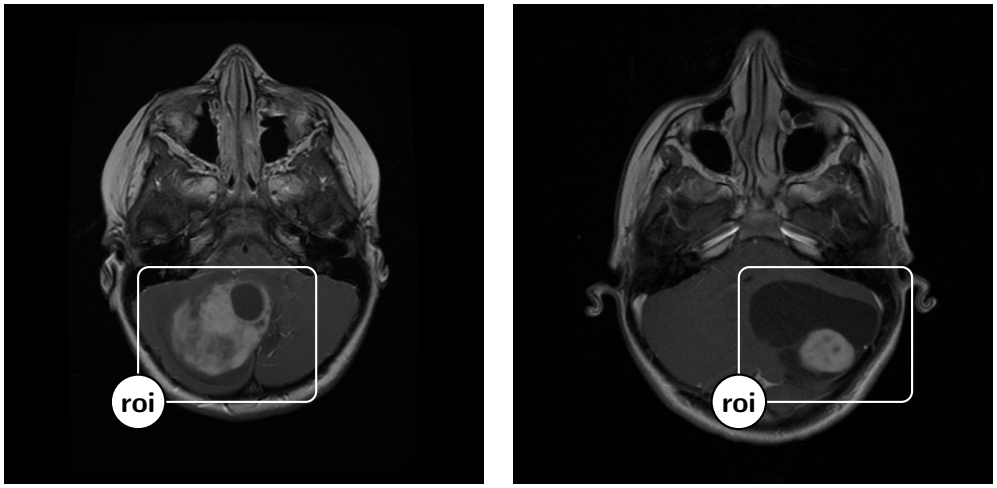
Table of the key findings is provided on the following page.

| Method | $Tu - Br$ | $Cy - Br$ | $Tu - Cy$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| $DoB$ | ✓ | ✓ | ✓ | low | Low noise, good detection, |
| $F$-test | ✓-duplicate | ✓-duplicate | ✓- duplicate | high | Detects duplicated surfaces, but noisy |
| $L$ | ✓-weak | × | ✓-weak | mixed | shaded regions due to noise |
| $t$-test | ✓ | ✓ | ✓ | medium | connected boundaries, with clutter |
| $\chi^2$ | ✓ | ✓ | ✓ | low | connected surfaces |
| $KS$ | ✓ | ✓ | ✓ | high | Sensitive detector, reveals noise |
| $u$-test | ✓ | ✓ | ✓ | high | sensitive, reveals noise |
| $RRO$ | ✓ | ✓ | ✓ | high | low magnitude surfaces |
| Canny | ✓ | ✓-weak | ✓ | low | weak interfaces in $x, y$-plane |
| Steerable | × | ✓-weak | ✓ | medium | weak interfaces in $x, y$-plane |

Table 6.3: Surface filter characteristics on T2-weighted MRI scans. (✓) indicates detection of surface. $Tu - Br$ is brain and tumour interface, $Cy - Br$ is brain and cyst interface and $Tu - Cy$ is cyst and tumour interface

## 6.9 Evaluation 6: T1-weighted with Contrast 3-D

In addition to adjusting the MRI sequencing to reveal different structures, a Gadolinium contrast agent can be administered to the patient prior to the scan. The contrast agent produces a bright response, and it concentrates in the solid tumour region of the patient. For T1-weighted imagery, this can be used in order to increase the contrast between tumour and white matter regions, resulting in the solid tumour material presenting as a hyper-intense region with respect to the reference white matter of the surrounding tissue. Since the contrast agent concentrates in the solid tumour material, and not the cystic components, differentiation between solid tumour and cysts is also improved (Fig 6.14). The increase in contrast allows for improved differentiation between solid tumour and white matter, as well as improved differentiation between solid tumour and cysts. Assisting in improving the performance of surface detection in these areas. This leads to a $Tu$-$Cy$ interface which shares similarities with the Int type interface employed in Chapter 5, while the $Tu - Br$ and $Cy - Br$ interface more closely resemble the Com1 interface types but at different intensity levels.




(a) (b)

Figure 6.14: Pilocytic Astrocytoma pathologies in T1 imagery with contrast agent

### 6.9.1 Results

Figures for the results are provided on the following page.

(a) Volume 1 (b) Reference 1 (c) Canny (d) Steerable (e) *DoB* (f) $F$-test

(g) $L$ (h) $t$-test (i) $\chi^2$ (j) $KS$ (k) *RRO* (l) $u$-test

(m) Volume 2 (n) Reference 2 (o) Canny (p) Steerable (q) *DoB* (r) $F$-test

(s) $L$ (t) $t$-test (u) $\chi^2$ (v) $KS$ (w) *RRO* (x) $u$-test

Figure 6.15: T1c-weighted surface maps patient 1 and 2

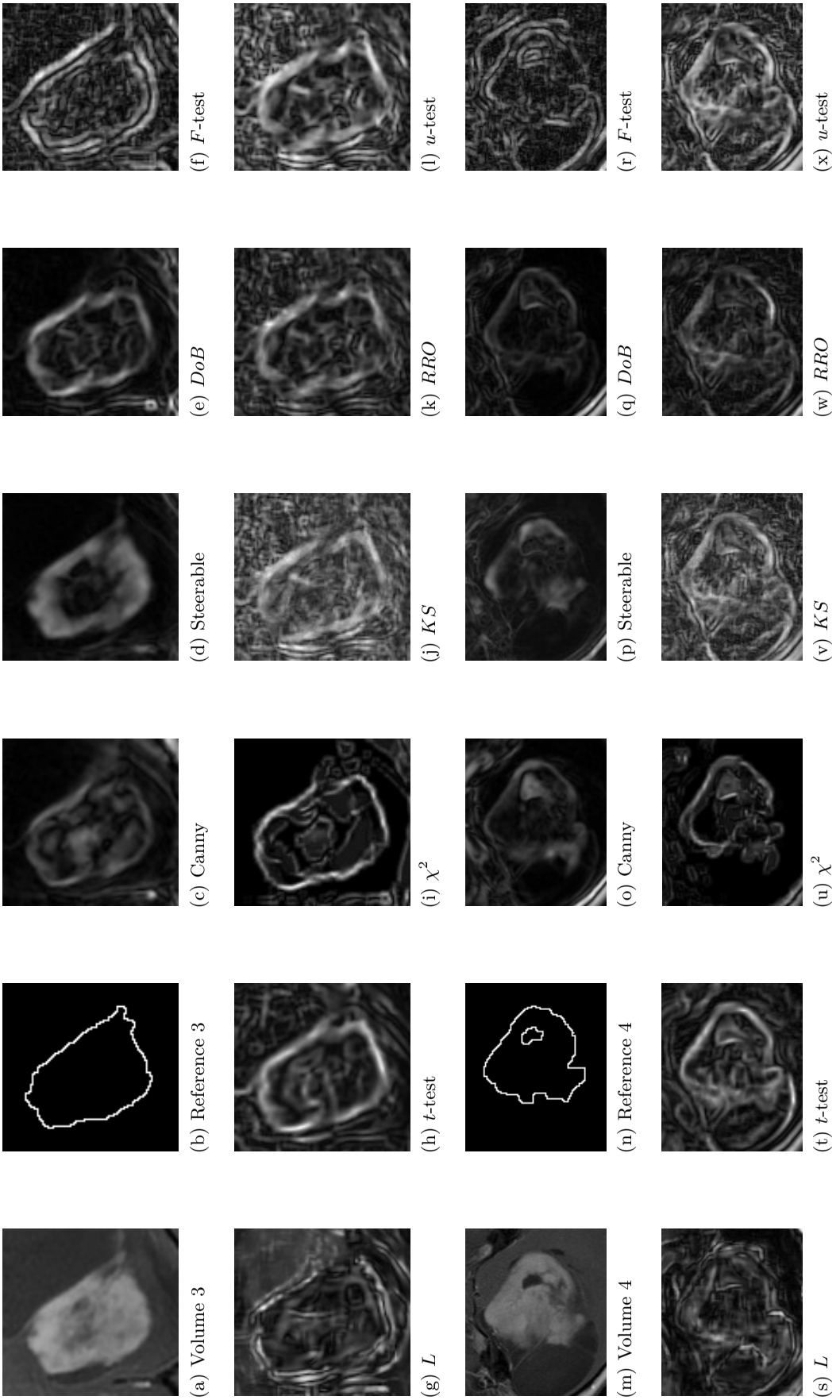
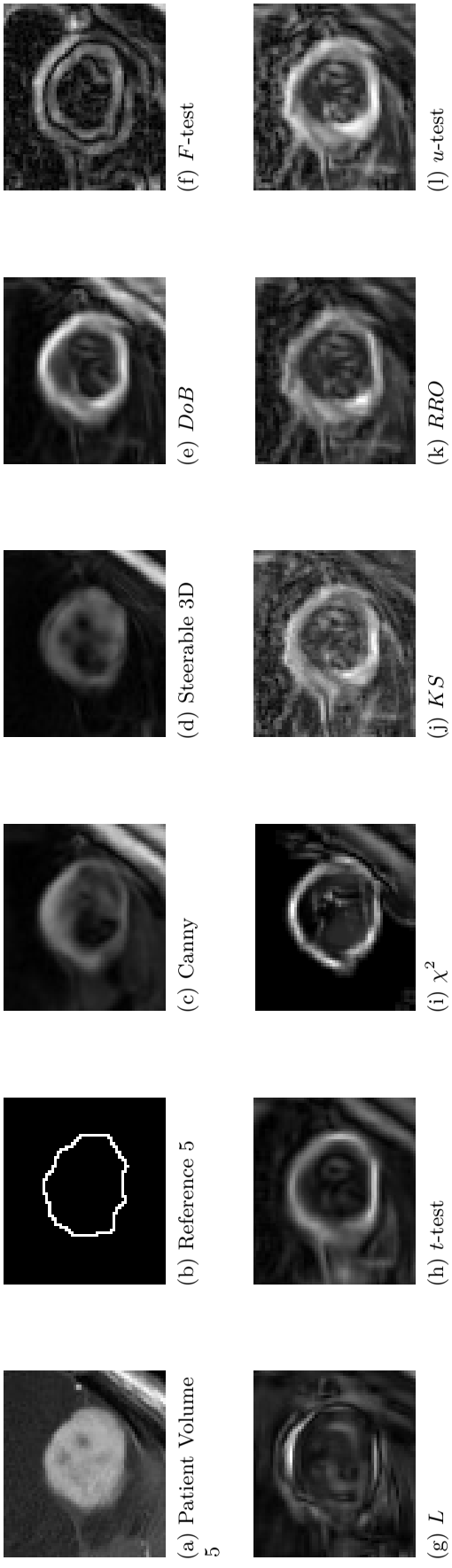



Figure 6.16: T1c-weighted surface maps patient 3 and 4



(a) Patient Volume 5 (b) Reference 5 (c) Canny (d) Steerable 3D (e) *DoB* (f) *F*-test

(g) *L* (h) *t*-test (i) $\chi^2$ (j) *KS* (k) *RRO* (l) *u*-test

Figure 6.17: T1c-weighted surface maps patient 5

### 6.9.2 Discussion

The T1-weighted image with a contrast agent visually provides the greatest differentiation between regions of white matter tissue, solid tumour pathologies and cyst material. The best method for detecting the interface between cyst and solid tumour in addition to the interface between solid tumour and white matter was the $\chi^2$ method, which produced high magnitude, accurate surfaces, with good connectivity. The $\chi^2$ responses can also be categorised as having a low noise profile with only minor clutter. While the $\chi^2$ method produced high magnitude tumour surface points, it also suppressed interfaces between cyst and white matter.

The parametric $DoB$ and $t$-tests produced good visual detection, exhibiting strong magnitude surfaces with good connectivity. In contrast to the $\chi^2$ method, interfaces between cyst and white matter regions were resolved with a high magnitude response. The Canny method offered some detection of the surfaces, however the anisotropic spatial resolution of the MRI volumes was present in these cases again, this gives the appearance of strong interfaces as large filled regions viewed in a 2-D axial view. When viewed in 3-D, this appears as a large uniform truncated surface. However, the Canny method response is uncluttered and low noise compared with some of the non-parametric statistical responses. The Steerable filter also produced low noise results, however, not all of interfaces of the tumours are resolved. Similarly to the $\chi^2$ method, the Steerable filter also suppresses the cyst boundaries with white matter tissue.

The non parametric $KS$, $u$-test and $RRO$ tests again provided similar surface map responses, with the $RRO$ method slightly outperforming the $u$-test and to a greater extent the $KS$ method in terms of noise and clutter. However, both filter responses were reasonably cluttered relative to the $\chi^2$, parametric and gradient methods.

The surfaces resolved by the non parametric statistical filters were of a high magnitude and accurately positioned, however the surfaces possessed some areas of fragmentation, and missed boundaries. The $F$ test method produced strong magnitude surfaces across all three interface types. |However, as with the T1-weighted and T2-weighted imaging modalities, those surfaces were duplicated.

### 6.9.3 Key Findings

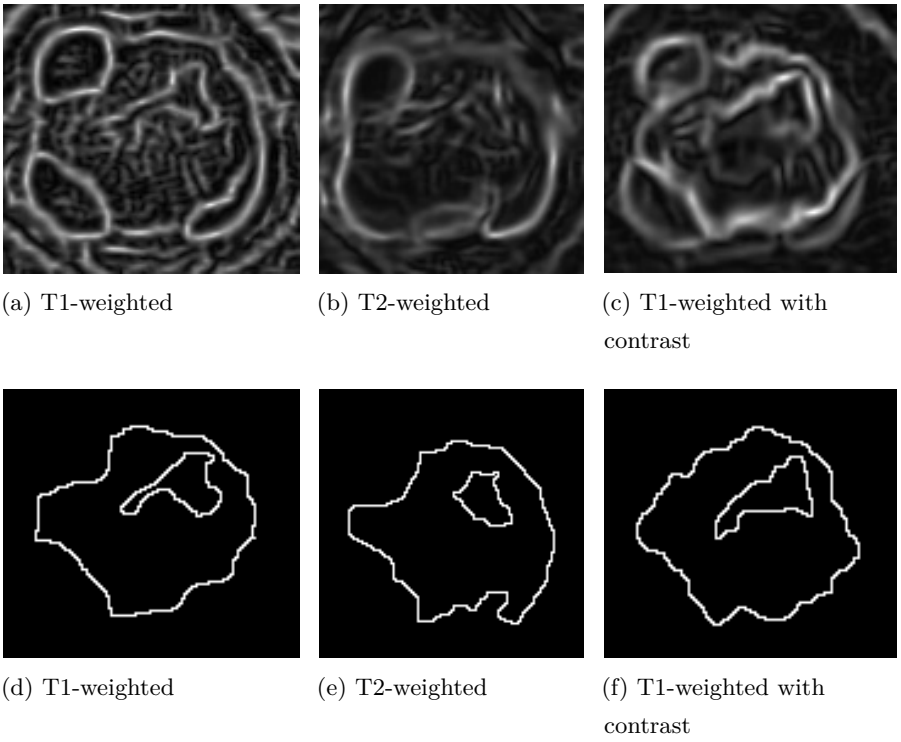
Table of the key findings are provided on the following page.

| Method | $Tu - Br$ | $Cy - Br$ | $Tu - Cy$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| $DoB$ | ✓ | ✓ | ✓ | low | Good detection, connectivity and magnitude |
| $F$-test | ✓-duplicate | ✓-duplicate | ✓- duplicate | high | Detects duplicated surfaces, but noisy |
| $L$ | ✓-duplicate | × | ✓-weak | mixed | shaded regions due to noise |
| $t$-test | ✓ | ✓ | ✓ | medium | high magnitude connected boundaries, with clutter |
| $\chi^2$ | ✓ | × | ✓ | low | high magnitude and connected |
| $KS$ | ✓ | ✓ | ✓ | high | Sensitive to noise,low magnitude surfaces |
| $u$-test | ✓ | ✓ | ✓ | high | Sensitive to noise,low magnitude surfaces |
| $RRO$ | ✓ | ✓ | ✓ | high | Sensitive to noise,low magnitude surfaces |
| Canny | ✓ | ✓-weak | ✓ | low | weak interfaces in $x, y$-plane relative to $z$ |
| Steerable | × | ✓-weak | ✓ | medium | weak interfaces in $x, y$-plane |

Table 6.4: Surface filter characteristics on T1-weighted MRI scans with contrast agent. (✓) indicates detection of surface. *Tu-Br* is brain and tumour interface, *Cy-Br* is brain and cyst interface and *Tu-Cy* is cyst and tumour interface

## 6.10 Summary of Results

Generally the best performing surface detection method was the Student's $t$-test, providing high magnitude surfaces with good connectivity. Across the different patient cases, the $t$-test provided best surface map results for the T1-weighted and T2-weighted pre contrast image modalities. Providing good delineation between solid tumour and cyst, between grey matter tissue and solid tumour as well as between cyst and grey matter tissue. The $t$-test also produced the most optimal results for the tumour pathology cases across the different modalities, particularly in the case of the T1-weighted modality, where other methods struggled to produce an optimal result (Fig 6.18).



(a) T1-weighted (b) T2-weighted (c) T1-weighted with contrast

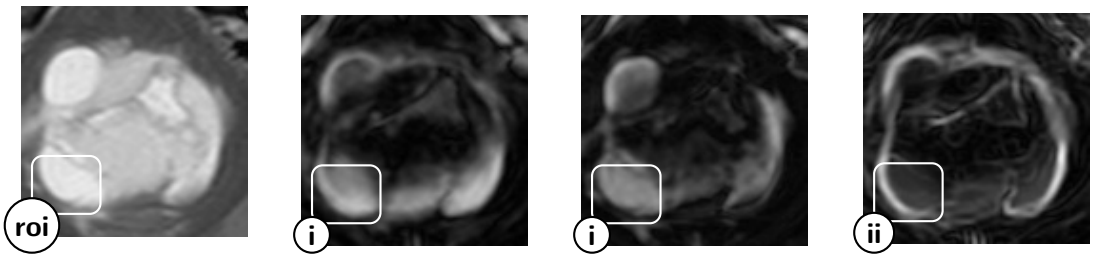
(d) T1-weighted (e) T2-weighted (f) T1-weighted with contrast

Figure 6.18: Cross section of $t$-test results from a single patient across the three modalities. The results show good detection of the surface interface on each of the modalities characterised by high magnitude surfaces with good connectivity.

A characteristic of the $t$-test results were surface maps which were not overly cluttered. In addition, because there is no local normalisation of the neighbourhood mask region, image noise remained low on a global scale. There are instances where internal structure within the solid tumour region is resolved, depending on context this could be deemed as clutter. However the magnitude of these surfaces is lower than that of the solid tumour interface, thus post processing steps such as hysteresis thresholding and morphological operations could be employed to assist in mitigating some of the effects of this if needed.

The second best performing method on pre contrast modalities, and the best method with the use of a contrast agent was the $\chi^2$ method. This method overall produced surface maps which could be characterised by containing very little clutter and noise, while at the same time producing accurate high magnitude responses with good connectivity. However, some localisation inaccuracies and missed surfaces on T1-weighted pre contrast images, led the $t$-test to be deemed the preferential method (Fig 6.8h,i). The low noise profile of the response, and very little clutter makes the $\chi^2$ test method quite suitable for a number of segmentation techniques, such as region growing algorithms which are optimised for low noise surface maps with strong magnitude connected boundaries (Militello et al., 2015). By manipulating the number of bins employed by the $\chi^2$ method, the types of interfaces present in the image can be selected to be resolved. For example, using 4 bins for T1-weighted images with contrast produces surface maps where the cyst-grey matter interface is omitted. Producing a very clear surface map with just the solid tumour interface present.

The $DoB$ method produced good delineation across the different interface types, however, this method was not as consistent as the $t$-test and $\chi^2$ methods on T1-weighted modalities, signified by some success (Fig 6.9e) and some failure (Fig 6.8e) across different PA pathologies. However, for T2-weighted and T1-contrast modalities, the $DoB$ method produced good results. As the vector magnitude method is adaptable to image volumes which are anisotropic in resolution, the $DoB$ method produced markedly better results than the 3-D Canny method, even though they are both designed to produce high outputs to strong differences in image brightness (Fig 6.19).




(a) T2-Patient Volume (b) Canny (c) Steerable (d) DoB

Figure 6.19: Baseline methods produce weak $x$ and $y$ component surfaces relative to $z$ component, signified by high magnitude shaded regions within the internal tumour structure(i), however Vector Magnitude techniques account for anisotropic image resolution thus are not as greatly affected by poor $z$ resolution (ii).

The $KS$ method was able to resolve the tumour surface as well as the surface interface between cyst and tumour across the different modalities. Surfaces presented typically with good connectivity. However, as the $KS$ used ranked data points instead of pixel

values, it has the effect of locally normalising the output. This effect decreases the signal to noise ratio of the surface interface. This effect presents itself with a cluttered surface map, signified by spurious responses throughout the image, which can be seen in Fig 6.12i,v. Consequently, the presence of this effect can make post processing steps such as hysteresis thresholding more challenging, and limiting the use of the $KS$ method use to a narrower context. The $RRO$ and $u$ test methods also exhibit this local normalisation property, and the results are comparable with similar all round performance characteristics.

In a similar fashion to the synthetic tests, the variance based $F$ and $L$ method produced outputs with duplicated surfaces, this effect is most clearly visible in Fig 6.12f and 6.16g, signified by the production of a additional surface boundary either side of the correct surface interface.

The analysis undertaken indicates that the application case results produce comparable findings to the synthetic results across the 3 different MRI modalities. For T2-Weighted imagery, tumour interfaces typically present with a step change in image brightness. In the evaluations presented in chapter 5 the Int and Com1 type interfaces types are comparable to interfaces present in T2-weighted imagery, these interface types were shown to be the least challenging for all but the $L$ and $t$-test filter methods, with the gradient based Canny method well suited for this type of interface. However, the anisotropic resolution of the real data impeded the Canny method's ability to resolve boundaries in the $x$ and $y$ plane to the same extent as interfaces with a strong $z$ dimension component, resulting in strong over detection of surfaces which exist between layers in the axial plane. This resulted in solid filled in regions (Fig 6.19a). The 3-D statistical $DoB$ and $t$-tests are also methods which exploit changes in image brightness, but the Vector Magnitude method filter architecture allows for weighted vector components based on image resolution, mitigating some of the excessive $z$ component surfaces, producing good results (Fig 6.19c).

In addition, the statistical methods which performed well on the intensity based synthetic interfaces also performed strongly on T2-weighted images on grey matter - tumour interfaces. Notably the $\chi^2$ test. The T1-weighted results also produced comparable results to the synthetic data, here some of the interfaces from this modality are comparable to the synthetic combination interfaces types in terms of region profiles. Results of the synthetic tests strongly indicate the Student's $t$-test and $\chi^2$ test to be the best performing methods, which was reflected in the results of the application case data.

# Chapter 7

# Discussion

## 7.1 Introduction

Chapter 5 presented an objective analysis of statistical surface detection, while chapter 6 presents a qualitative analysis of statistical surface detection across independent MRI modalities. In this chapter, the characteristics of the results are explored to identify trends between the evaluations in order to discuss the generalisability and transferability of the results. Primarily the advantages and disadvantages of the different filter methods are discussed, including any performance benefit of utilising a Vector Magnitude approach to surface detection. Additionally, parameters such as neighbourhood scale are explored in order to see if the trends observed in 2-D statistical edge detection are applicable in 3-D.

## 7.2 3-D Statistical Surface Detection approaches

In this work, two 3-D configurations of Statistical surface detection were developed, The Maximum Response configuration, and the Vector Magnitude configuration. 3-D detection is a requirement for accurate detection in 3-D data in order to identify surface features which exist between the layers within a 3-D image volume (Monga et al., 1991). 2-D edge detection approaches are often used for 3-D data, as is typically applied to the 2-D layers individually and sequentially, however this approach is not optimal as $z$-component surfaces are not identified. 3-D detection also uses 26-voxel connectivity neighbourhoods instead instead of 8-connectivity. Therefore for a given scale, the 3-D approach has greater noise suppression and resolving power.

The Maximum Response method is a modified implementation of the Statistical Edge techniques evaluated by Williams et al. (2014) for 3-D data. In 3-D, the approach requires the neighbourhood mask to be orientated through 13 unique positions. While the Vector Magnitude method is a novel approach developed to increase efficiency by

using 3 mask positions oriented orthogonally to the $x, y$ and $z$ image volume dimensions. This allows for a L2-Norm calculation to be applied to determine a statistical difference vector gradient in place of an outright maximal orientation and corresponding statistical difference magnitude. A number of benefits arise from this configuration:

- Improved computational efficiency
- Improved scaleability
- $z$-spacing correction

Firstly the method is more efficient, requiring the processing of 3 mask orientations in place of 13 which is a considerable reduction in processing time (Table 3.1). Additionally this allowed for parametric tests to be implemented using a convolution based approach as described in Section 3.3.3.2, this allows for very fast processing times for parametric tests, at speeds comparable to traditional gradient techniques.

With the maximum response technique, a constrained approach is used when scaling the size of the neighbourhood mask, fixing the number at 26-voxel directions for each mask size. However, when increasing the size of the neighbourhood mask, new potential orientations become possible, this increase is also exponential. Therefore at larger mask sizes the maximum response method is not fully optimised in detecting the true surface gradient direction. Contrastingly, the vector magnitude approach is scaleable, in the sense that as the neighbourhood size is increased, no further orientations are required to accurately compute the L2-Norm statistical vector.

One of the primary advantages of the vector magnitude method which was not apparent in evaluations 1-3, but was strongly visible once the VM technique was applied on real data in evaluations 4-6, was the mitigation of the negative effects of $z$-space biasing. This remained very apparent for both baseline Canny and Steerable techniques, but not the VM method (Fig 6.19). As is typical in real patient cases, the MRI data did not have isotropic resolution (Tab 6.1), but because the VM approach applies weighting to the $z$-component of the statistical vector relative to the $z$-spacing resolution, this leads to a reduction in the strength of surfaces in the $z$ direction, for a clearer and more representative surface map.

## 7.3 Parametric Tests Characteristics

The following section provides details of the characteristics and recommendations for the different parametric statistical methods implemented using the VM configuration.

### 7.3.1 Difference of Boxes (*DoB*)

#### 7.3.1.1 Surface Response

The *DoB* method is a computationally efficient parametric statistical measure. A 2-D Implementation of the *DoB* method indicates performance similar to that of the Canny method, strong where there is a substantial intensity component to the edge, and weaker where the variance component is large (Williams et al., 2014). The results from chapters 5 and 6 support this observation with an overall summary of each evaluation presented in Table 7.1. The *DoB* measure compares the mean value of each mask region to produce a response, and as such behaves in a similar fashion to traditional gradient methods such as the 3-D Canny method, by responding to shifts in image brightness. Characteristically the *DoB* method was therefore effective on intensity based interfaces. (Figures 5.9e, 5.13a)

Conversely, since high variance regions, such as those in the texture (Fig. 5.2b)) interfaces, are comprised of shifts in image intensity, just on a smaller spatial scale, the *DoB* method produces spurious responses in the high variance regions as illustrated in Fig 5.10e. Thus it was largely unsuccessful at resolving surfaces for the Tex type interface volumes. The *DoB* method objectively performed well on the combination interfaces. However, in the high variance regions the *DoB* method produced over-detection artefacts.

#### 7.3.1.2 Observations

One advantage of the *DoB* method over other intensity based statistical measures, such as the Student's *t*-test is that the *DoB* method is not as susceptible to potential fail points in areas such as junctions and corners, this can be seen in the cuboid and staircase images in Fig 7.1. Therefore if 3-D image structure is known to contain flat surfaces and regular corners a priori, the *DoB* is preferential over the *t*-test method.

On T1-weighted MRI case data the *DoB* method produced good delineation across the different tissue types, although not as reliable as the *t*-test signified by some successes (Fig 6.9e) and some failures characterised by missed responses such as the example in Fig 6.8e. For T-1 weighted modality the *DoB* method offers similar performance to the baseline 3-D Canny method.

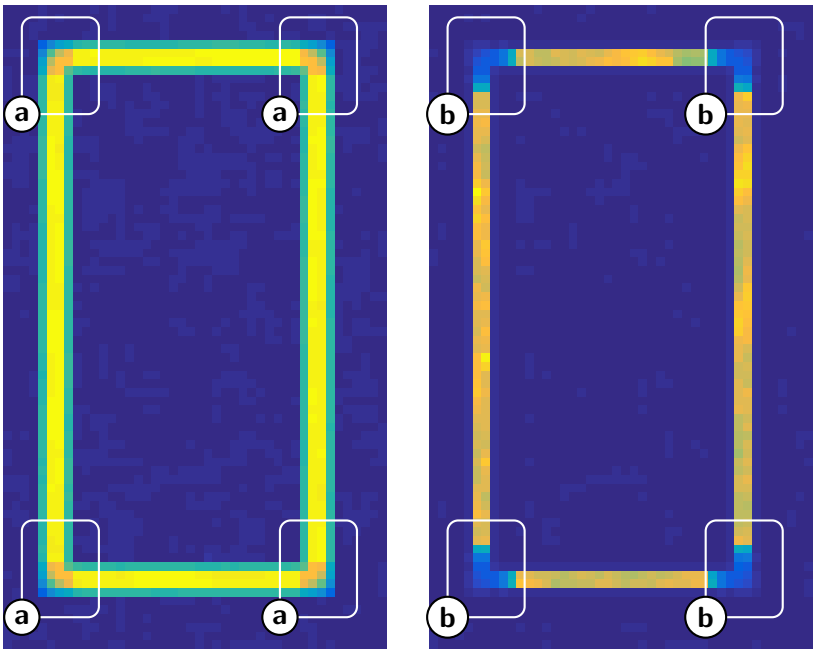
For the T2-weighted and T1-weighted (contrast enhanced) images, the *DoB* method performed significantly better than the baseline Canny and Steerable methods. This was due to the anisotropic resolution of the data, which unlike the VM approach, is not accounted for. The effect can clearly be seen in Fig 6.19 signified by the lack of defined interfaces from the baseline methods compared with the *DoB* method in the

| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, no over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some localised over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, no over-detection artefacts | No surface, over-detection artefact | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, connected surface, no over-detection artefacts | No surface resolved, some over-detection artefacts | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface, no over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some regional over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Accurate detection, connected surface, very few spurious surfaces | No surface resolved, over-detection artefacts | Surface resolved and connected with some loss of detail, regional over-detection artefacts | Surface resolved with finer detail lost some regional over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Some missed surfaces, low levels of noise, good connectivity | Missed surfaces, good noise suppression, low magnitude surfaces |

| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Partial detection | Low | Good detection, but weak interfaces |
| 5 (T2), Section 6.8 | Detected | Detected | Detected | Low | Low noise, good detection |
| 6 (T1c), Section 6.9 | Detected | Detected | Detected | Low | Good detection, connectivity and magnitude |

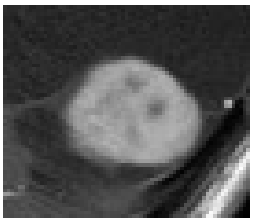
Table 7.1: Difference of Boxes filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.



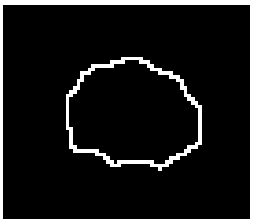

(a) DoB IntCu (b) $t$-test IntCu

Figure 7.1: Difference of boxes was shown to have good performance in corners (a) compared to the mean based parametric Student's $t$-test (b).

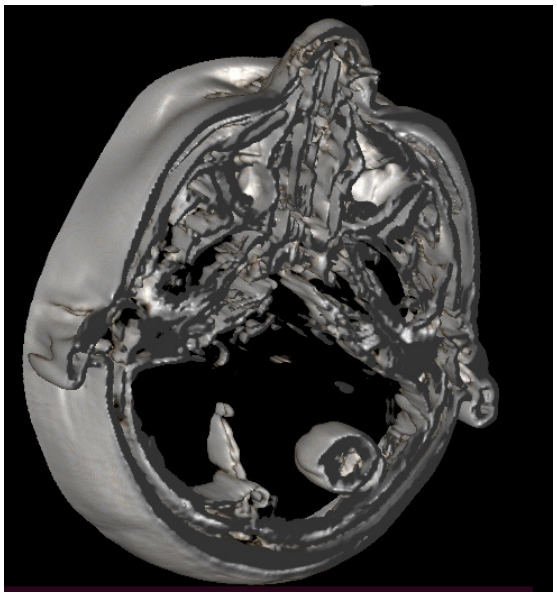
highlighted region of interest. The Canny method has a closed shaded region whereas the statistical methods with a weighted z-component resolves the tumour border with greater clarity. The *DoB* method was well suited to both T1 contrast enhanced and T2-weighted images this resulted in a strong detection of the tumour surface (Tab 7.1).



(a) 2-D layer



(b) 2-D Reference



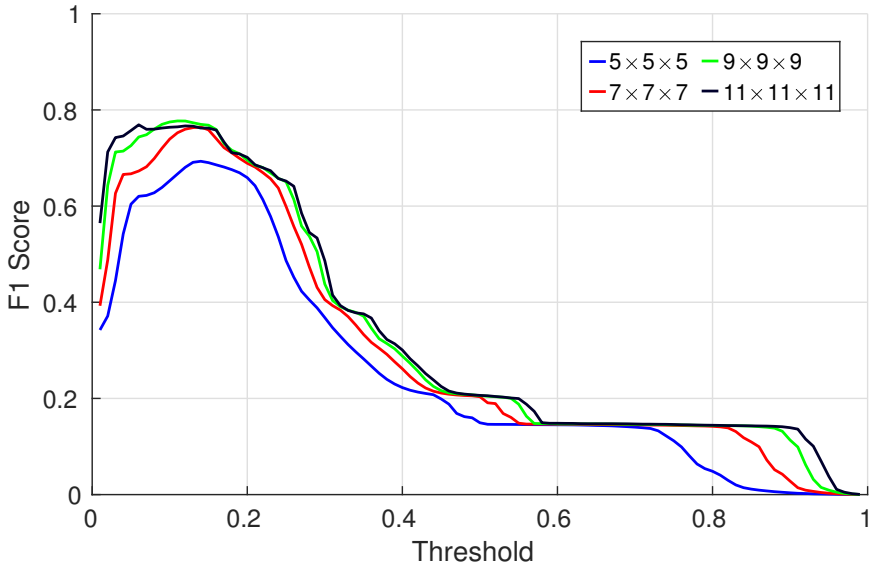
(c) 3-D

Figure 7.2: 3-D visual *DoB* result on T1-Weighted MRI image (case 5). The Cyst interface is strongly resolved by the method with relatively low noise. As a surface detection method, and not a segmentation technique, other internal structures are also revealed in the outer portions of the image.

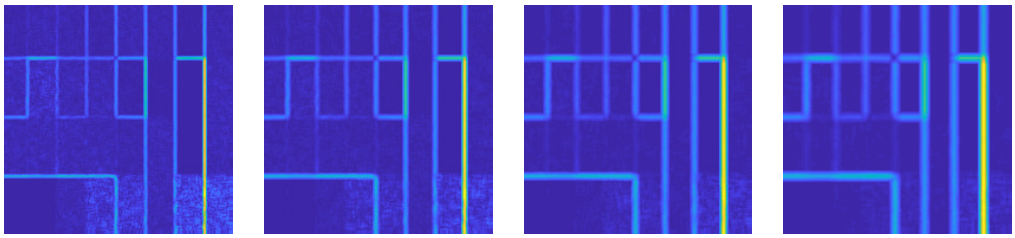
These results indicate that the *DoB* method is most effective on strong intensity boundaries, where the baseline gradient methods are also most effective. However, as the *DoB* was shown to be more reliable across different interface types, therefore statistical *DoB* method is preferential over gradient based methods, the technique also has the benefit of being more computationally efficient than the other parametric tests (Table 3.1).

#### 7.3.1.3 Scale Response

When different neighbourhood scales are used, there is a trade-off between noise suppression and uncertainty in the surface position. Larger neighbourhood masks have greater noise suppression at the cost of uncertainty in the position of the surface. Fig 7.3 illustrates this, as the scale is increased, the number of spurious responses is fewer, indicated by the dark blue background, while the surface width is greater, indicated by the wider light blue and yellow regions. From an objective analysis perspective, the greater noise suppression at large mask sizes achieved greater F1-score performance on this image type.




(a) F1-Score



(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.3: Comparison of neighbourhood scales for *DoB* method. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages.

### 7.3.2 Student's $t$-test ($t$-test)

#### 7.3.2.1 Surface Response

The Student's $t$-test determines if the mean of two samples are significantly different in normally distributed data. It is another parametric test, and as such is a computationally fast performing algorithm (Table 3.1).

Similar to the *DoB* method, and traditional derivative based methods, the $t$-test method is able to accurately detect intensity interfaces or where a difference in the mean exists between the regions. The results from Chapters 5 and 6 support this observation and an overall summary of each evaluation presented in Table 7.2

Typically the $t$-test produced surfaces with a high signal to noise ratio, in images where surfaces exist between regions of uniform brightness (very low variance), this results in a high surface magnitude. However, this can cause problems where surfaces of different strengths exist in the same volume, and was most notable with the IntSp and IntCSt

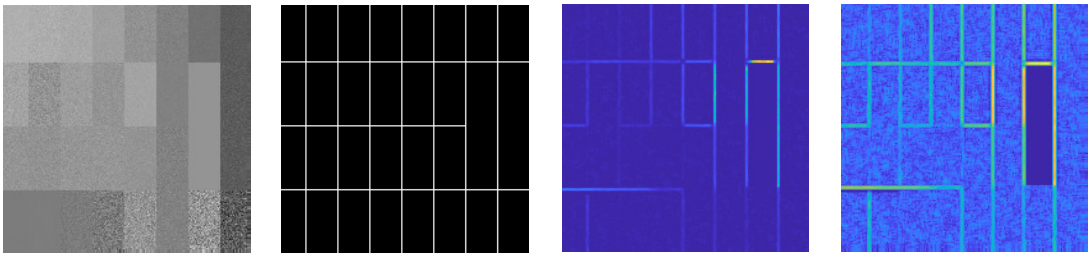
| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, no over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection detection, weak corner resolution, no over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts, weaker corners and edges | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, unconnected surface, no over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, some varying of surface strength, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface, no over-detection artefacts, weaker strength corners | No surface resolved, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts, variable surface strength | Accurate detec-tion,connected surface. some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Accurate detection, connected surface, localised artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, varied surface strength, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Good detection, good noise suppression, good connectivity, mixed magnitude surfaces | Good detection, good noise suppression, good connectivity, low magnitude surfaces |

| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Detected | Mixed | Good connectivity, strong surfaces |
| 5 (T2), Section 6.8 | Detected | Detected | Detected | Medium | connected boundaries, with clutter |
| 6 (T1c), Section 6.9 | Detected | Detected | Detected | High | Sensitive to noise,low magnitude surfaces |

Table 7.2: Students $t$-test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

images (Figures 5.19h and 5.24h). Fig 7.4 provides a logarithmic output from the $t$-test, revealing weaker magnitude boundaries that have also been detected, while still maintaining reasonable suppression of noise. These weaker boundaries have absolute values which would not typically be considered weak, however surfaces positioned between a uniform region can produce a very high magnitude response relative to the other surfaces, this masks the weaker boundaries when normalised for an on-screen visual display. It is therefore recommended that if the $t$-test is applied to images which contain regions of uniform brightness, and the target surfaces are not positioned on the boundaries of the uniform regions, then global thresholding methods should be avoided and replaced with local thresholding methods or avoid any normalisation of output prior to thresholding in order to maximise the performance of the method.

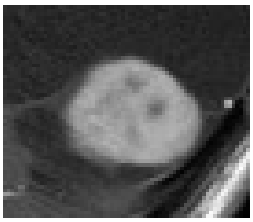


(a) Test Image (b) Reference (c) 5×5×5 (d) 5×5×5 log

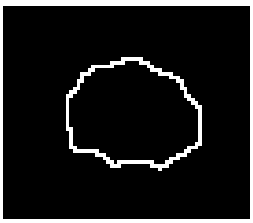
Figure 7.4: Logarithmic output shows that $t$-test detects surfaces, but they can be masked by very strong responses to uniform image regions

#### 7.3.2.2 Observations

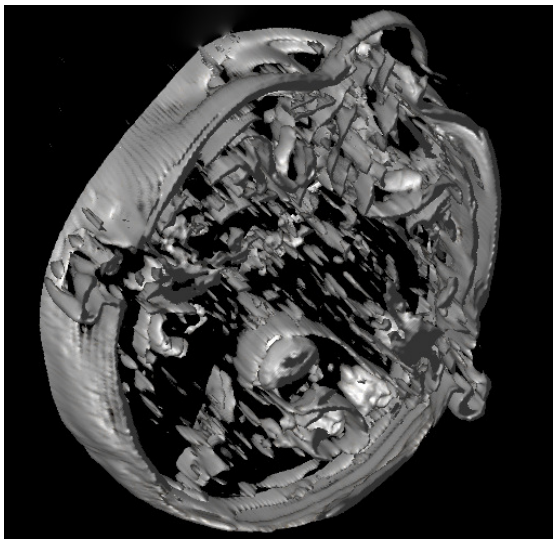
A notable finding was that the $t$-test method proved to be one of the best surface detection methods for tumour interface detection in MRI data, this was most notable on T1-weighted images with no contrast agent, where it outperformed the other methods in terms of resolving the PA tumour interfaces between both white matter tissue, and the cystic components (Figures 6.7h and 6.9h,7.5).



(a) 2-D layer



(b) 2-D Reference



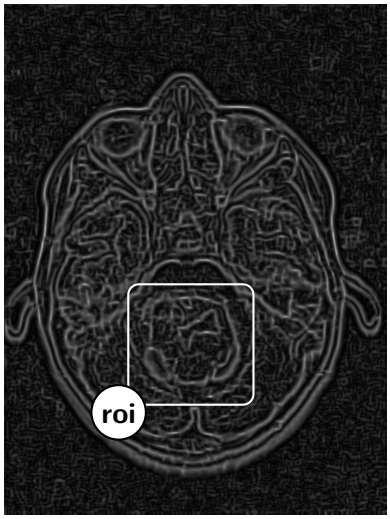
(c) 3-D

Figure 7.5: 3-D visual $t$-test result on T1-Weighted MRI image (case 5). The $t$-test has a strong response to the tumour interface, but other structures are also revealed

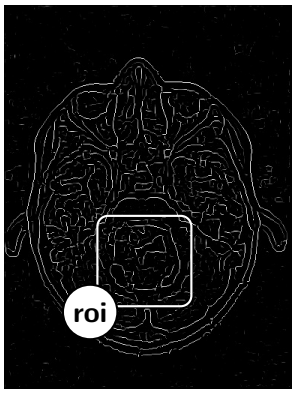
While shown to be one of the most effective methods at resolving intensity based interfaces with strong detection, Chapter 5 evaluations 2 and 3 does reveal that the $t$-test was not optimally suited to resolving corner interfaces when compared against the other mean based parametric test, the *DoB* method (Fig 7.1). Williams et al. (2014) found the $t$-test to be effective at detecting mean based interfaces, while providing good connectivity and good noise suppression, these findings were found to be consistent with both the 3-D synthetic imagery, including the real MRI data. Since regular corners and junctions are not typical in human anatomy, and the method offers efficient detection of mean based interfaces superior to that of the baseline methods, the $t$-test methods is recommended for use with MRI data.

Some internal tumour structure was also resolved, which is undesirable in this context, however this can largely be mitigated by using non maximum suppression and hysteresis thresholding (Fig 7.6).

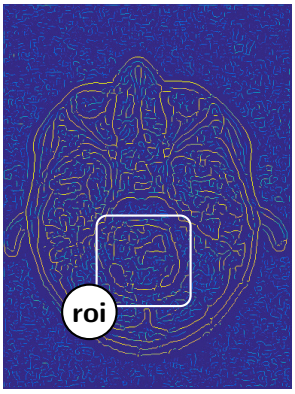

(a) T1-weighted MRI



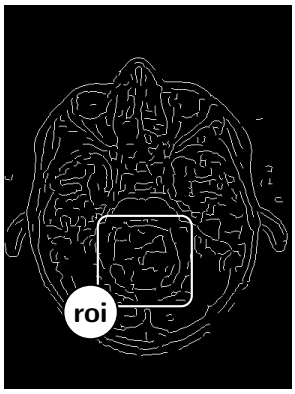

(b) Surface Map




(c) NMS




(d) Surface range




(e) Hysteresis thresholding

Figure 7.6: Removing $t$-test clutter from T1-weighted MRI surface map response: (a) Shows the test image with a region of interest (roi), while (b) shows the cluttered surface map result. (c) With Non Maximum Suppression applied. (d) Shows the relative magnitudes of the surfaces in the NMS image, weaker surfaces are those with a light blue colour, while strong magnitude surfaces are yellow. (e) then shows the removal of the light blue clutter surfaces through hysteresis thresholding

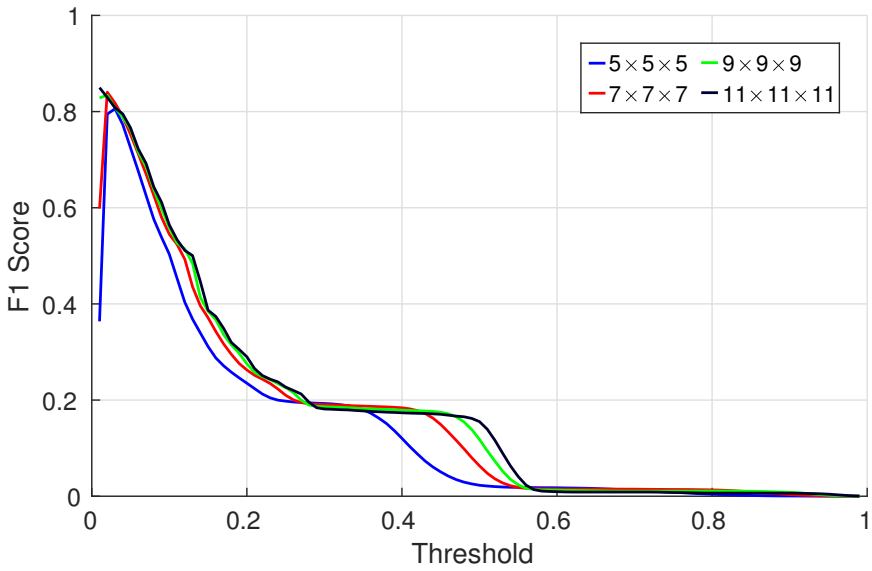
#### 7.3.2.3 Scale Response

Fig 7.7 reveals the narrow threshold window for an optimal result when global thresholding methods are applied however, the $t$-test still achieves the highest peak score out of the baseline methods and parametric measures.

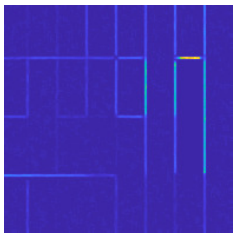
### 7.3.3 Fisher Test ($F$-test)

#### 7.3.3.1 Surface Response

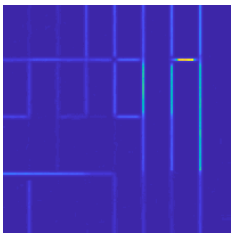
As a variance based statistical measure, the $F$-test produced accurate results for texture based interfaces, but was less successful on other interface types. The results from



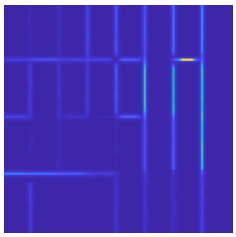

(a) F1-Score



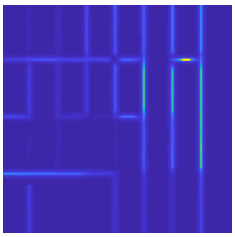
(b) 5×5×5



(c) 7×7×7



(d) 9×9×9



(e) 11×11×11

Figure 7.7: Comparison of neighbourhood scales for *t*-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages.

| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | duplicated surfaces, some additional over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Inaccurate surface, duplication, some over-detection | Inaccurate, connected surface, duplicated surface and additional over-detection artefacts |
| 2.1 (Cu), Section 5.5 | Missed surface, duplicated boundary either side, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Inaccurate surface position, duplicated surface, no general over-detection artefacts | Displaced graduated surface |
| 2.2 (Sp), Section 5.5 | Duplicated surface either side of interface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Duplication of boundary, connected surface, some over-detection artefacts | Displacement of boundary, some over-detection artefacts |
| 2.3 (CSt), Section 5.5 | Duplicated surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Incomplete detection, duplicate responses, over-detection artefacts | Disconnected surface surface with single voxel displacement, some over-detection artefacts |
| 2.3 (FSt), Section 5.5 | Duplicate inaccurate surfaces, some over-detection artefacts | Surface resolved, localised artefacts, some additional over-detection artefacts | Inaccurate duplicated surfaces, some over-detection artefacts | Surface accurately resolved with small shift, some over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.5 | Good detection, but duplicated surfaces, some noise, good connectivity | Good detection, but duplicated surfaces, some noise, good connectivity |

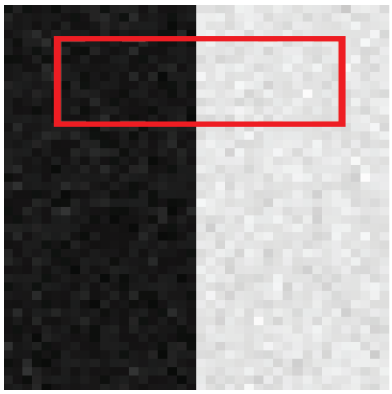
| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 5.5 | Detected - duplicate | Detected - duplicate | Detected - duplicate | High | Duplicated surfaces |
| 5 (T2), Section 5.5 | Detected- duplicate | Detected -duplicate | Detected - duplicate | High | Detects duplicated surfaces, but noisy |
| 6 (T1c), Section 5.5 | Detected -duplicate | Detected -duplicate | Detected - duplicate | High | Detects duplicated surfaces, but noisy |

Table 7.3: Fisher test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

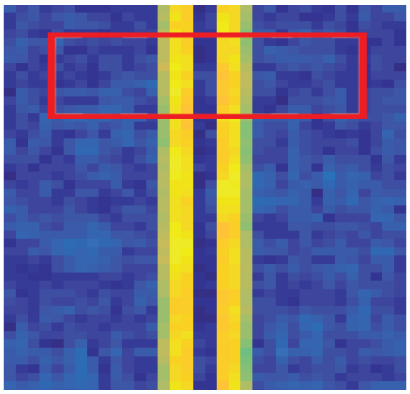
chapters 5 and 6 support this observation and an overall summary of each evaluation presented in Table 7.3. The effectiveness on texture based interfaces can be observed clearly in Fig 5.10f where the method produces the joint best subjective result with the Likelihood method. Objective results presented in Fig 5.13b support the subjective results with Fisher methods having the widest optimal threshold window. In the work of Williams et al. (2014) the 2-D implementation of this technique also performed strongly on variance based interfaces, however it was observed that the technique produced a shift in the resulting edges which is proportional to the size of the neighbourhood mask. While performing well on texture based interfaces, the shift observed in 2-D also occurs in 3-D data, including the production of a double surface boundary. The creation of an erroneous double boundary occurs on opposing sides of the intensity interface and can be seen in Fig. 5.9f. Figures 5.14f, 5.19f, 5.24f,r, reveal this issue extents to different topological interfaces.

#### 7.3.3.2 Observations

The duplicated boundary artefact present with synthetic data also extends to real imagery, on T1-weighted MRI the Fisher method also produced duplicated surfaces along the interface (Fig. 7.8h-j). The cause of the double boundary is due the fact that as the neighbourhood mask approaches the location of an intensity interface, the surface boundary will reside in one of the dual regions of the neighbourhood mask. The presence of the interface leads to a variance difference between each of the dual regions of the mask, thus producing an off-centred surface response. As the filter mask moves into the correct surface boundary location, the variance difference between the regions equalises, resulting in no surface in the correct location, and as the neighbourhood mask passes over the interface location, the off-centred surface response is repeated. To further illustrate the problem Fig. 7.8a-g shows the effect of a Fisher mask traversing a mean based interface, leading to a double interface. For real data, the trends found with synthetic data were also apparent. The duplicated boundary is a characteristic that applies to all intensity based boundaries. In order to fully utilise the strong texture surface resolving potential the Fisher method provides, morphological operations should be applied to convert double boundaries to single surfaces

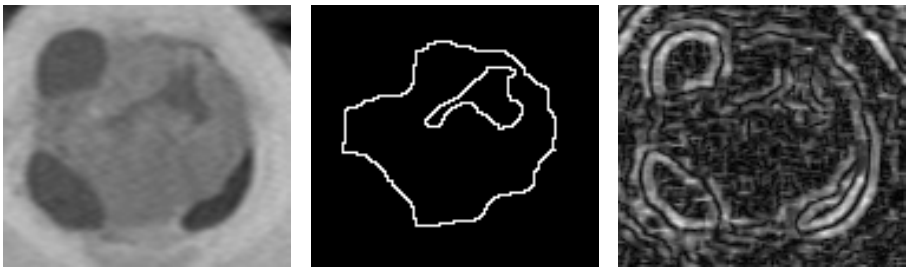


(a) Filter mask in different positions across a surface



(b) Synthetic image response

(c) Low Δ (d) High Δ (e) Low Δ (f) High Δ (g) Low Δ

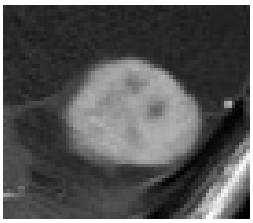


(h) MRI layer (i) Interface location (j) Fisher test surface map

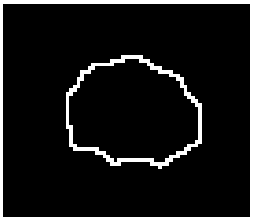
Figure 7.8: 2D representation of the double surface artefact associated with Fisher test and likelihood methods. (a) Synthetic mean interface with path of mask locations. (b) Duplicated surface response on synthetic image. Mask positions 2 (c) and 4 (d) have the large variance differential between regions, thus these two locations produce surfaces in each of the locations, Mask positions (e,f,g) possess small variance differentials, thus do not produce a surface. (h-j) Fisher test response to MRI layer, produces duplicated surface at tumour interface.

Results from Evaluation 1 with the combinational (Com1Fl,Com2Fl) images show that under these conditions the duplication effect is not as pronounced (Figures 5.11f and 5.12f) due to the fact the surface is detected at the correct location in addition to the duplication either side. However, the F-measure analysis indicates that the $F$-test was unsuccessful on this interface type (Fig 5.13c,d) due to the displacement of the surface which occurs after non-maximum suppression and thresholding.

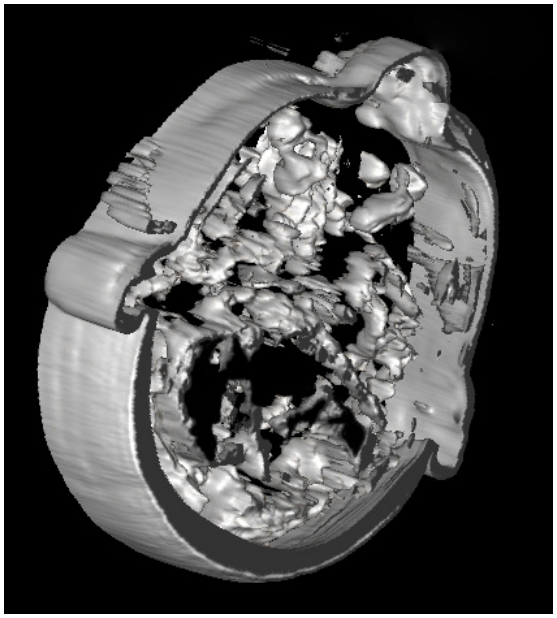
In addition to synthetic and T1-weighted imagery, these problems also arose across T2-weighted and contrast T1-weighted results, it is therefore recommended that the $F$-test only be used when prior information is known about the image structure, and should not be the first choice when the target surfaces are primarily mean based interfaces unless post processing morphological operations are available prior to any further stage, such as segmentation.



(a) 2-D layer



(b) 2-D Reference

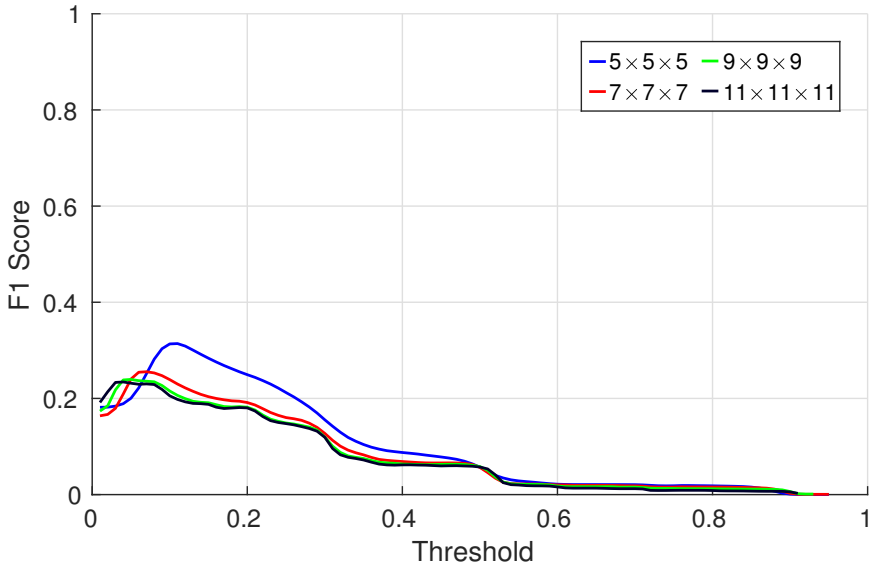


(c) 3-D

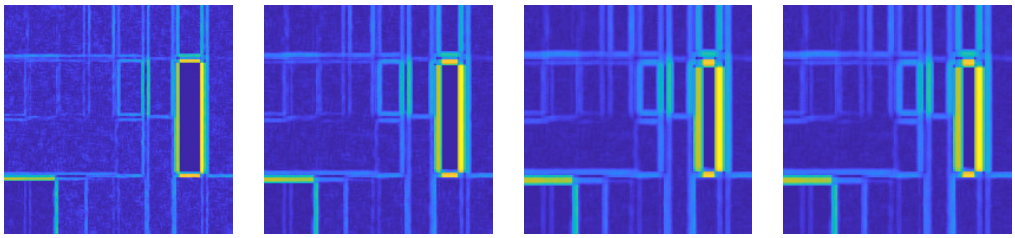
Figure 7.9: 3-D visual $F$-test result on T1-Weighted MRI image (case 5). The $F$-test has a strong response to the tumour interface, but other structures are also revealed

#### 7.3.3.3 Scale Response

At larger scales the F1-score performance metric decreases (Fig 7.10), this is due to the wider surfaces amplifying the double boundary error. Visual results show greater suppression of unwanted texture at larger mask sizes, while producing high magnitude connected boundaries, but morphological post processing operations may be required to mitigate duplication.




(a) F1-Score



(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.10: Comparison of neighbourhood scales for $F$-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages. F1- score is low due to surface duplications in the results.

### 7.3.4 Log-Likelihood Ratio Test ($L$-test)

#### 7.3.4.1 Surface Response

The $L$-test is a parametric test which when used in this context exploits shifts in variance between regions, and therefore should be effective on the Tex interface type. An overall summary of each evaluation presented in Table 7.4 indicates this to be the case, but also the results from Evaluation 1 reveal that the $L$-test method was able to resolve all the interface types tested. However, with the combinational (Com1, Com2) images, the $L$-test is not the optimal choice. The subjective visual results indicate that the method detects the surface (figures 5.11g and 5.12g), but the surface map reveals that there are some duplicated surface points running parallel to the interface. In 2-D the work of Williams et al. (2014) also found that the $L$-test was able to detect both intensity and texture based interfaces.

| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | connected surface, no over-detection artefacts | Accurate detection, connected surface, very few over-detection artefacts | Accurate detection, duplicate surfaces are present, some localised over-detection artefacts | connected surface, duplicated surfaces, some over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, some localised corner and edge artefacts, localised over-detection near surface | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, duplicated surface, no general over-detection artefacts | Displaced surface response, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Duplicated surface, artefacts near interface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Displacement of boundary, connected surface, some over-detection artefacts | Displacement of boundary, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Incomplete detection, duplicate responses, over-detection artefacts | Disconnected surface surface with single voxel displacement, some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Duplicate inaccurate surfaces, very few spurious surfaces | Surface resolved, localised artefacts, some additional over-detection artefacts | Surface resolved, some displacement, very few spurious surfaces | Surface resolves with displacement, very few spurious responses |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Good detection, but duplicated surfaces, good noise suppression, magnitude surfaces | moderate detection, but duplicated surfaces, moderate noise suppression, strong magnitude surfaces |

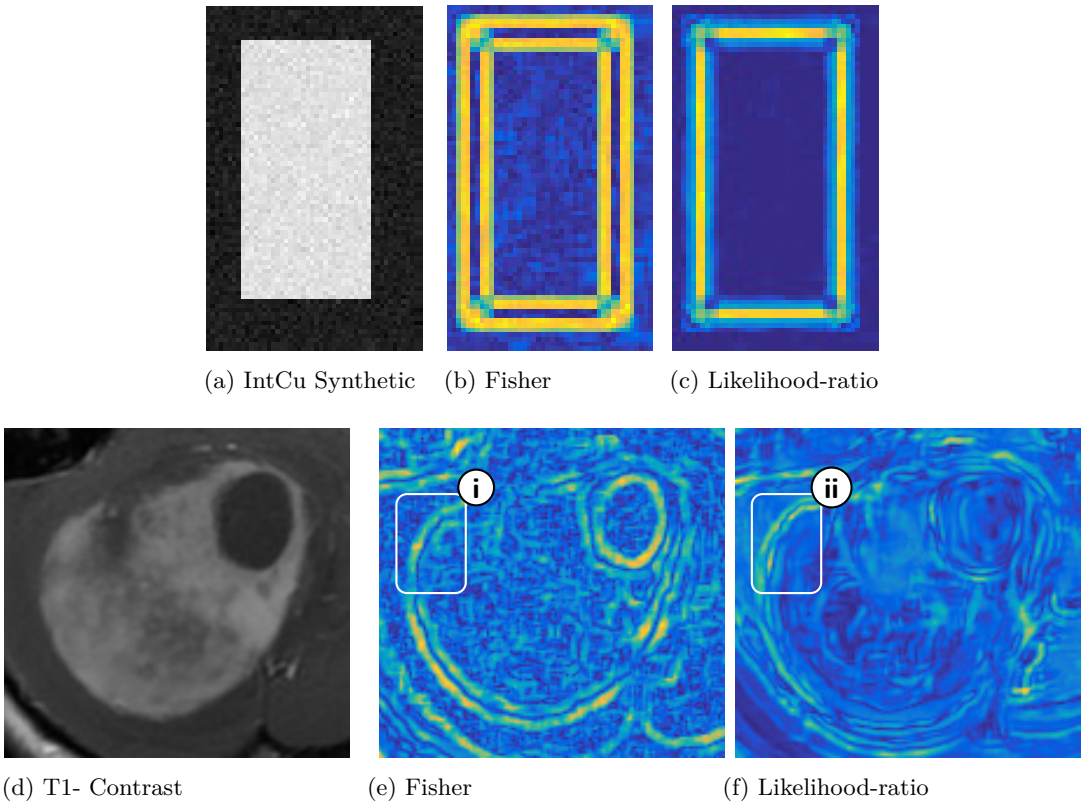
| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Not detected | Not detected | Not detected | mixed | Poor detection |
| 5 (T2), Section 6.8 | Detected -weak | Not detected | Detected -weak | Mixed | Shaded regions due to noise |
| 6 (T1c), Section 6.9 | Detected -duplicate | Not detected | detected -weak | Mixed | Shaded regions due to noise |

Table 7.4: Log Likelihood ratio test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

Objective results from Evaluation 1 reveal that after post processing stages are applied, the Likelihood method also produced optimal results across a wide threshold range superior to that of the baseline Canny and Steerable methods (Fig 5.13c,d). One characteristic of the $L$-test method was the suppression of noise compared to the other methods on synthetic imagery, notably on images with a large variance component (figures 5.15, 5.20, 5.25), but the suppression was consistent across many different interface types as can be illustrated in Fig 5.29g.

#### 7.3.4.2 Observations

While Evaluation 3 presented in Chapter 5 revealed a strong suppression of texture, it does highlight a limitation of the $L$-test, the production of duplicated boundaries. While the Fisher-test produced surfaces which occurred either side of the interface and not in the correct location, the $L$-test boundaries differ, with surfaces occurring both in the correct position and the duplicated positions parallel to the interface, producing what visually appears as a smeared surface interface (Fig 7.11).



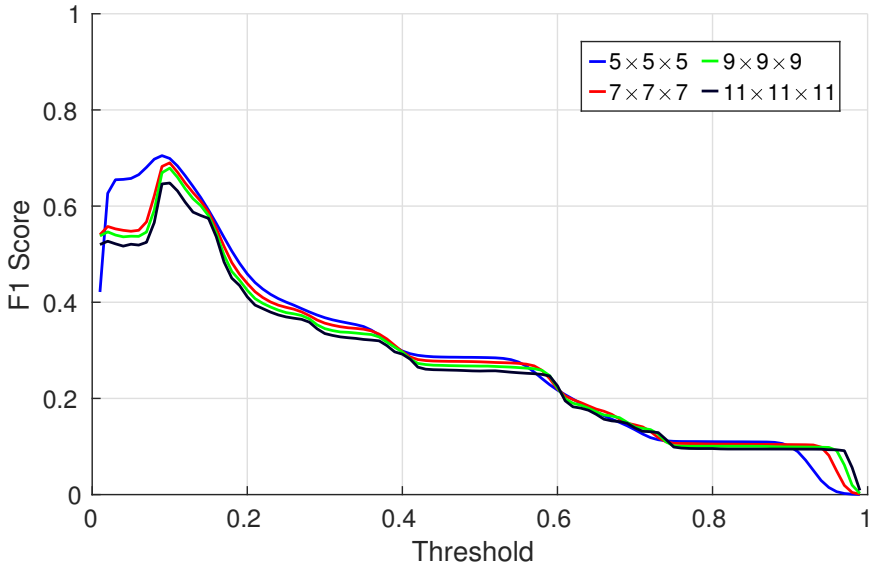

(a) IntCu Synthetic (b) Fisher (c) Likelihood-ratio

(d) T1- Contrast (e) Fisher (f) Likelihood-ratio

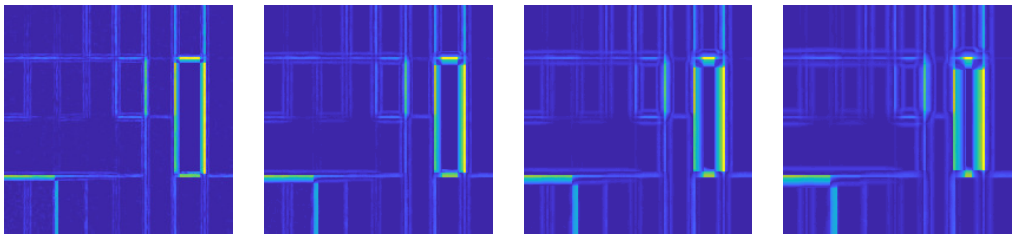
Figure 7.11: Duplicated boundaries; Fisher vs Likelihood ratio test. The Fisher test presents with duplicated boundaries running parallel to the reference interface on synthetic data (i), while the Likelihood method resolves the reference boundary in addition to duplicated boundary producing a "smeared" surface. This is reflected in the real imagery, manifesting as a triple interface (ii), with a stronger central boundary and weaker duplicated parallel boundaries

#### 7.3.4.3 Scale Response

At increased scales, regions with very low to zero variance can cause artefacts, these manifest as wide responses at the surface location, this kind of artefact can be viewed in Fig 7.12 illustrated by the wider yellow and light blue section at larger mask sizes.



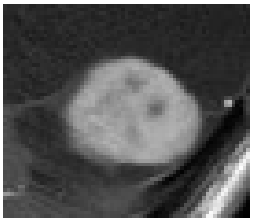

(a) F1-Score



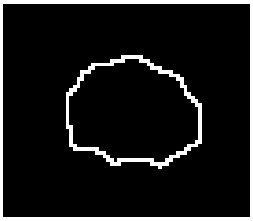
(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.12: Comparison of neighbourhood scales for $L$-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages.

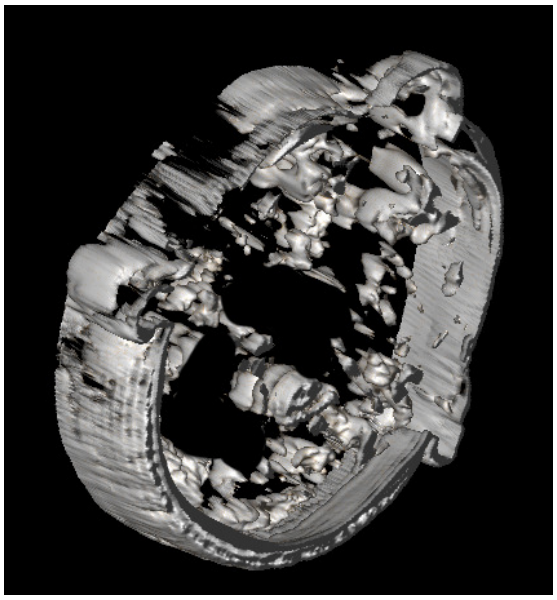
Williams et al. (2014) observed that the Likelihood-ratio test was able to resolve edges in 2-D histological images, however for 3-D MRI data, the $L$-test was shown to not be particularly effective for T1, T2 and T1c modalities. The $L$-test method was able to detect some pathology interfaces, but it did also introduce duplicated surfaces. On real MRI data, visually the $L$-test did not obtain the same level of noise or texture suppression as was the case with 2-D histological data or the 3-D synthetic data, this occurred across T1, T1-contrast and T2 weighted modalities. The $L$-test would therefore not be a first choice recommendation for MRI data.



(a) 2-D layer



(b) 2-D Reference



(c) 3-D

Figure 7.13: 3-D visual $L$ result on T1-Weighted MRI image

## 7.4 Non-parametric Tests Characteristics

The following section provides details of the characteristics and recommendations for the different non-parametric statistical methods implemented using the VM configuration.

### 7.4.1 The $\chi^2$ Test ($\chi^2$)

#### 7.4.1.1 Surface Response

The $\chi^2$ test is a non-parametric rank based test which checks for the independence of two sorted data sets. The $\chi^2$ method sorts each mask region sample into bins based on voxel intensity, the observed number of points in each bin should be similar if the two data samples come from common distributions. This allows for regions to be checked for similarity based on the distribution of the voxels. For the task of surface detection, the number of bins is a specified user parameter which controls the sensitivity of the filter. The greater the number of bins the more sensitive the method becomes, allowing for weaker structures to be resolved. For analysis with synthetic data, the bin parameter remained fixed across all interface types.

Objectively, the $\chi^2$-test was the best performing method on synthetic data, consistently scoring well across all interface topologies and types, as well as performing strongly on the synthetic multiple interface volumes. The effectiveness of the method can be observed across each evaluation summary presented in Table 7.5. Characteristically, the $\chi^2$ method offered good detection and accurate surfaces, which were connected and

| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, some discontinuity, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, no over-detection artefacts | Accurate detection detection, some discontinuities, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts, weaker corners | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, some discontinuities in surface strength, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface some over-detection artefacts | Accurate detection, some discontinuities, some over-detection artefacts | Accurate detection, connected surface some over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts, weaker strength corners |
| 2.3 (FSt), Section 5.6.4 | Accurate detection, connected surface, some over-detection artefacts | Surface detected with some inaccuracies, surface discontinuities, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Best detection, some noise, good connectivity, strong magnitude surfaces | Best detection, some noise, good connectivity, strong surface magnitudes |

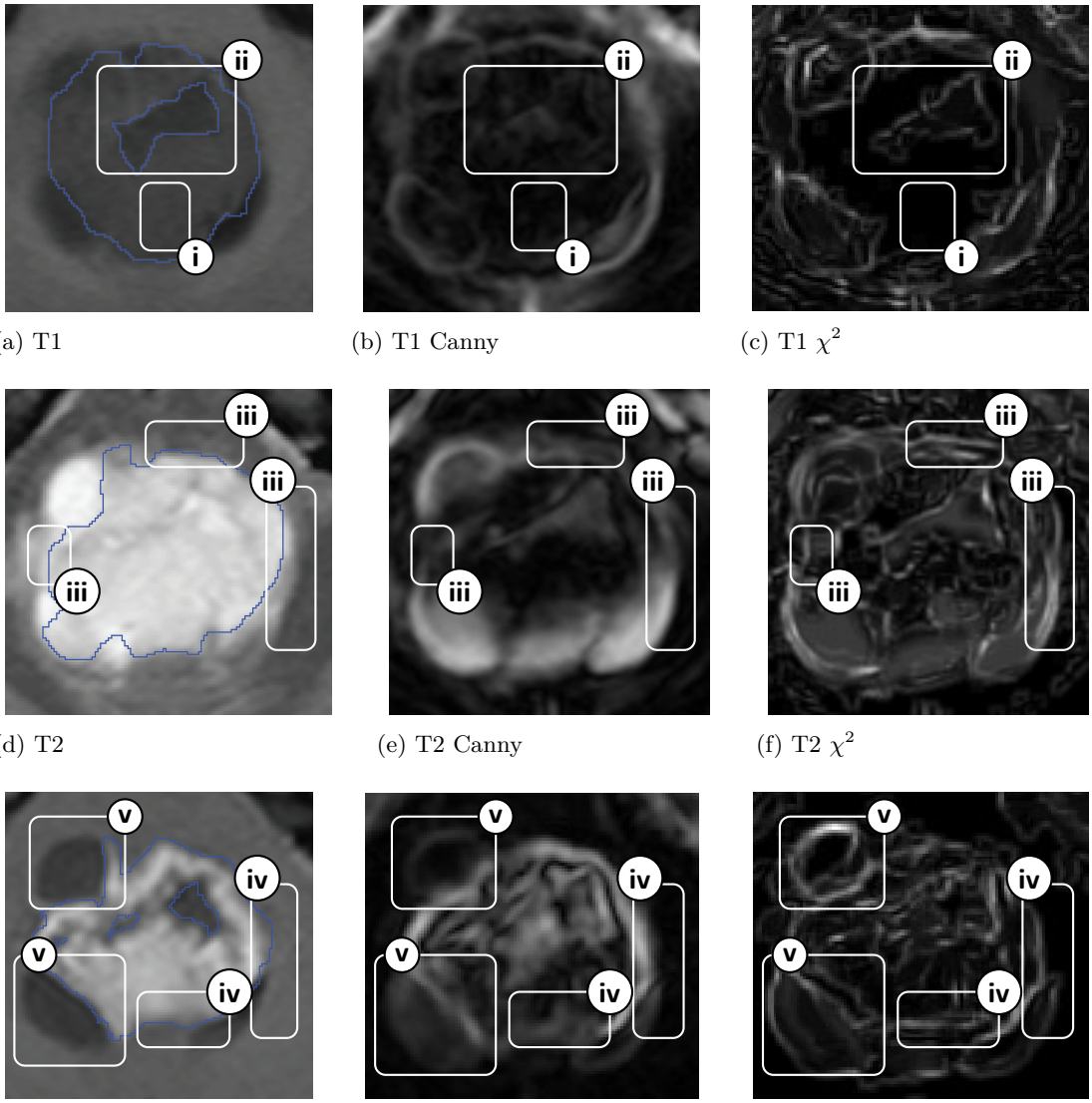
| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Detected | Low | Good connectivity and noise suppression |
| 5 (T2), Section 6.8 | Detected | Detected | Detected | Low | connected surfaces |
| 6 (T1c), Section 6.9 | Detected | Not detected | Detected | Low | high magnitude and connected |

Table 7.5: $\chi^2$ test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

high magnitude. In addition, the $\chi^2$ method is able to produce surface map responses which are low in spurious responses when compared against the other non parametric methods. The $\chi^2$ method achieved good performance across all topologies and was the best performing method on both of the Evaluation 3 image types (MultiFlat and MultiCurve). (Figures. 5.29i, 5.30i, 5.31).

#### 7.4.1.2 Observations

Observing the MRI results, the $\chi^2$ method produced good results across a range of interface types. When T1-weighted imagery is observed the $\chi^2$ method produced results comparable to the Canny method, but with fewer spurious responses, stronger surface magnitude and improved surface connectivity, and in some instances better detection of structures internal to the tumour pathology (Fig 7.14a-c). T2-weighted imagery produces interfaces characterised by greater luminosity differentials, in which derivative based techniques should be the optimal choice. Performance between the baseline methods and the $\chi^2$ technique show comparable results, however the $\chi^2$ method does produce surfaces with a greater magnitude and improved connectivity this is illustrated in Figure 7.14d-f, and also be seen in Chapter 6 in Figures 6.11u and 6.12i,u.




(a) T1 (b) T1 Canny (c) T1 $\chi^2$

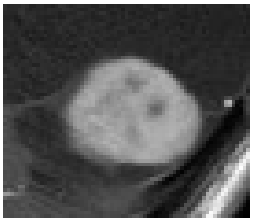
(d) T2 (e) T2 Canny (f) T2 $\chi^2$

(g) T1-contrast (h) T1-contrast Canny (i) T1-contast $\chi^2$

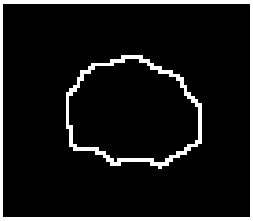
Figure 7.14: Comparison of $\chi^2$ vs baseline on real data. For T1 imagery, the $\chi^2$ method produced fewer spurious responses, signified by the empty black region (i). While also more clearly resolving internal structure (ii). On T2 imagery, the $\chi_2$ method produces stronger magnitude surfaces with improved connectivity (iii). On T1-Contrast imagery, the $\chi^2$ method produce better detection and connectivity (iv), including improved detection of the cyst interfaces with the tumour and normal brain tissue (v).

The $\chi^2$ was best suited to the T1-weighted contrast images, here the $\chi^2$ method consistently produced the best results across all examples. The method produced fewer spurious response and improved detection of the cyst interfaces in terms of connectivity and surface strength when compared with the baseline Canny method, as illustrated in

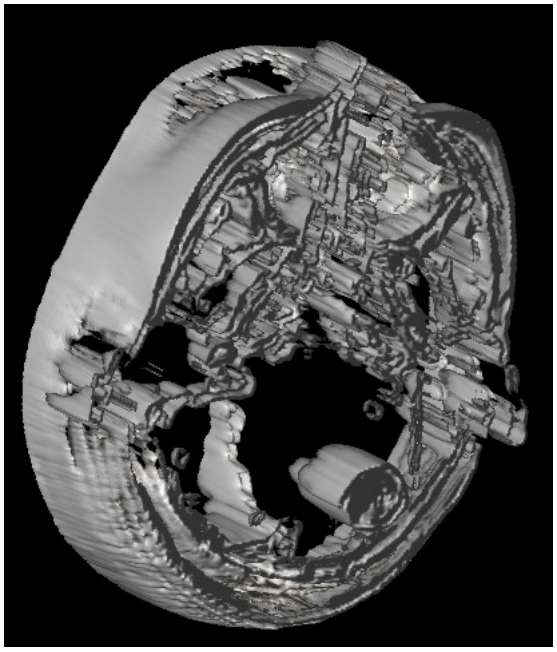
Figure 7.14g-i. In addition the $\chi^2$ method also produce good results on the T1-contrast images compared with the other statistical methods (Figures 6.15i,u, 6.16i,u and 6.17i).



(a) 2-D layer



(b) 2-D Reference

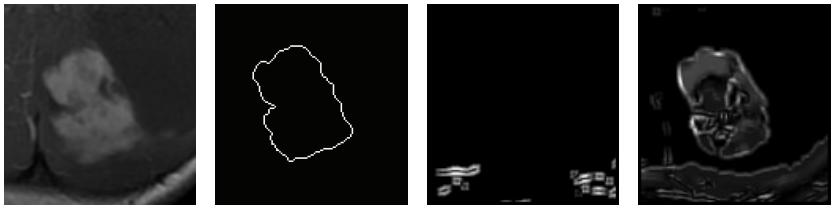


(c) 3-D

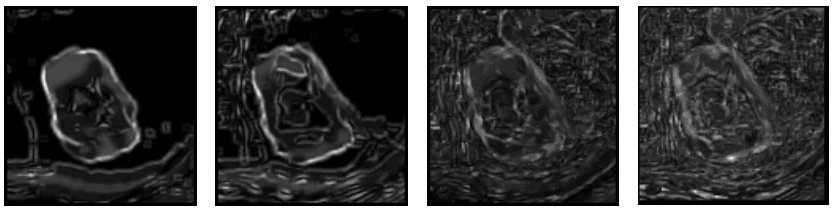
Figure 7.15: 3-D visual $\chi^2$ result on T1-Weighted MRI image

One feature of the $\chi^2$ test is that it requires an additional parameter which determines the number of bins used. Establishing the optimal number of bins is non trivial and is dependent on the image properties. Typically fewer bins result is fewer detected surfaces, but the magnitude of the surfaces increases, allowing for greater differentiation between regions. This is demonstrated in Fig 7.16, here different surface structure is resolved as the number of bins is increased. Fig 7.16b lacks a sufficient number of bins to differentiate between most regions in the image. While Fig 7.16h produced a response typified by excessive spurious responses which are not ideal. Fig 7.16,d and e presents a more suitable range due to the low noise, high magnitude, connected surfaces. The ability to control the sensitivity of the filter without adjustment of the scale parameter can produce more optimised results. The principal advantage of this is when the image modality in known a priori, manipulation of the parameter can better target specific interfaces. Such as with T1-weighted images with a contrast agent, the interface between white matter and cyst components can be suppressed (Fig 7.16c,d) producing a surface map response better suited for segmentation of the solid tumour material. However, the parameter needs to be specified manually, thus increasing the degrees of freedom, impacting the simplicity of use. The results of the $\chi^2$ method in

Chapters 5 and 6 were all obtained using a bin size of 4, which offered the best trade-off between detection and noise suppression.



(a) Original (b) Reference (c) Bin 2 (d) Bin 4



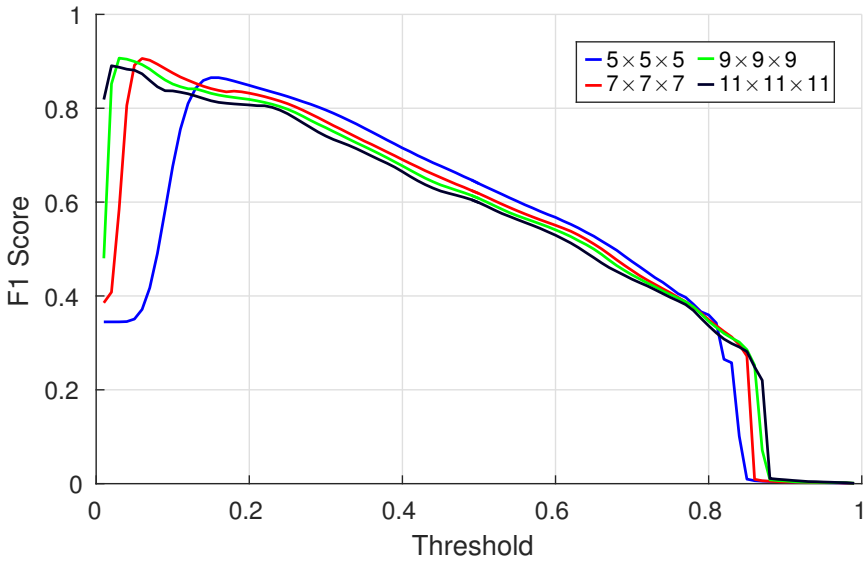
(e) Bin 8 (f) Bin 16 (g) Bin 32 (h) Bin 64

Figure 7.16: The sensitivity of the filter can be adjusted to reveal different image structure by modifying the number of intensity bins used by the method. (c-h) show the surface map response from the $\chi^2$ method with increasing number of bins

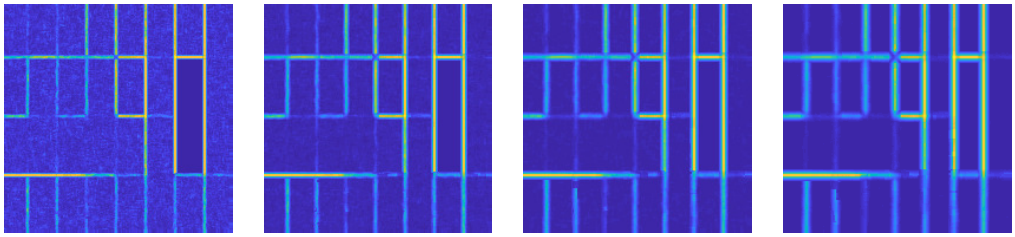
#### 7.4.1.3 Scale Response

In the work of Williams et al. (2014), the $\chi^2$-test was effective at larger scales for 2-D imagery. However, for 3-D it was observed that the $\chi^2$-test was effective across smaller scales (Fig 7.17), due to the increased signal to noise factor of a 3-D neighbourhood geometry.

When applied as a statistical edge detection measure, the method offers good detection of mean, mean-variance and variance interface types at larger mask sizes, but was less successful with a small mask size (Williams, 2008). However in 3-D it was found that the $\chi^2$-test was successful across a wide range of mask sizes, and no significant improvement above a neighbourhood size of $7 \times 7 \times 7$ (Fig 7.17). Monga et al. (1991) proposed that 3-D methods offer improved performance due to improved immunity to noise due to smoothing along the z- dimension, thus in 3-D a smaller mask size can be used to achieve the same level of noise immunity.




(a) F1-Score



(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.17: Comparison of neighbourhood scales for $\chi^2$-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages. Higher surface signal to noise ratio at larger scale due to suppressed noise signified by darker regions between surfaces. However, greater uncertainty in surface location, signified by a 'smeared' surface interface

Non-parametric statistical methods, which evaluate differences in region distributions, were shown to be robust to changes in interface types (Fig 7.17b,c,d,e), this was perhaps most notable for the $\chi^2$ method, in comparison to both baseline Canny and Steerable methods in addition to the statistical parametric tests the $\chi^2$ method produced the most complete detection while not over saturating the result with spurious surfaces. Therefore, when the type of interface is not known a priori, it can be inferred that non parametric tests such as the $\chi^2$-test should be more reliable at resolving different types of surfaces within the image volume.

The $\chi^2$ method produced surfaces which were characteristically accurate, strong and connected, with low levels of noise, these trends were also observed using real data. The $\chi^2$ method performed strongly across all 3 modalities, but was particularly effective

on T1-contrast images, producing a strong connected surface boundary with very little noise. It is therefore recommended that when surfaces detection is required on contrast enhanced T1-weighted imagery, the $\chi^2$ with a bin size of 4 should be the preferred approach.

### 7.4.2 Kolmogorov-Smirnov Test ($KS$)

#### 7.4.2.1 Surface Response

The $KS$-test is a non-parametric statistical test which compares the Empirical Cumulative Distribution Function (ECDF) of sorted ascending data points from two sample distributions. The $KS$ statistic is defined as the maximum difference between the ECDFs for each mask sample region. By calculating the $KS$ statistic the similarity of the distributions can be evaluated, in the context of surface detection, this local ranking of the data simulates the effect of local normalisation of the output, such that the strength of detected surfaces are relative the local neighbourhood region and not the global image. Using ranking values instead of absolute values, the $KS$ statistic will always fall between the range of 0 to 1. Consequentially surfaces which are of weak magnitude relative to the global image, but strong locally can be resolved. This means that the $KS$ statistical test is a sensitive method which has the ability to resolve surfaces with statistically similar regions. Across evaluations this can be observed as good detection but a noisy response (Table 7.6). This has advantages and disadvantages, the method can resolve different interface types, including those which are weak where other methods are not able to resolve the interface, for example figure 5.29j) reveals that the $KS$ method can resolve multiple surfaces of different strengths in the same image, including weak boundaries, consequently the output can also be saturated with weak surfaces, which may or may not be considered spurious responses depending on the context. This effect is more pronounced in regions of uniformity where no surfaces should be expected. Evaluation 1 illustrates this with detection of the different boundary types, but also the production of false positive surface points (Figures 5.9, 5.10, 5.11 and 5.12).

The results illustrate good detection, but with spurious surface points, $F_1$ scores reveal that the $KS$ method can achieve optimal detection results. However, optimal results are achievable with a narrower threshold window compared to some of the other statistical measures such as the $\chi_2$ and $L$-test methods (Fig 5.13a,c,d). The advantage of the sensitivity of this method is best illustrated in Fig 7.18, where after non-maximum suppression has been applied, the $KS$ method clearly offers one the best detection results in terms of resolving the most interfaces, this is supported by Fig 5.31a in which

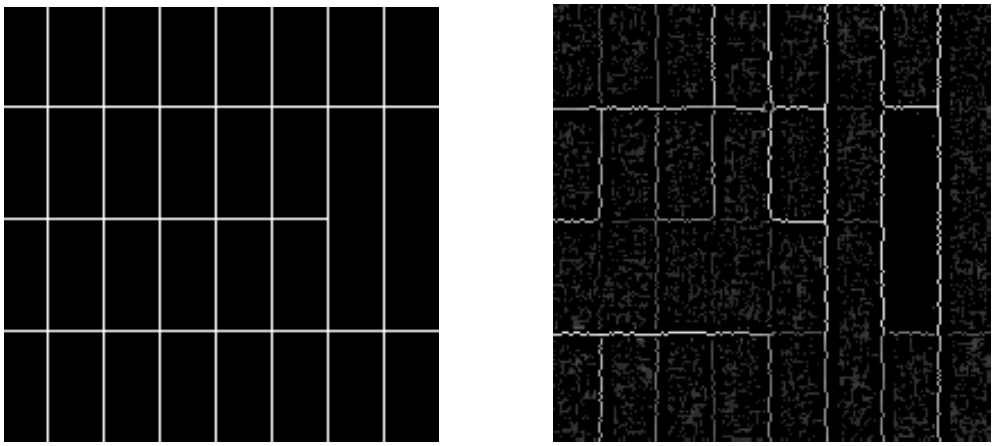
| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Accurate detection, connected surface, some over-detection artefacts | Surface detected with some inaccuracies, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Good detection, considerable noise, good connectivity, strong magnitude surface | Good detection, considerable noise, good connectivity, strong magnitude detection |

| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Detected | High | Good detection, but weak surfaces |
| 5 (T2), Section 6.8 | Detected | Detected | Detected | High | Sensitive detector, reveals noise |
| 6 (T1c), Section 6.9 | Detected | Detected | Detected | High | Sensitive to noise, low magnitude surfaces |

Table 7.6: Kolmogorov-Smirnov test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

the $KS$ method obtains the second best $F_1$-score after the $\chi_2$ output. When surfaces exist at different strengths the $KS$ test can be very effective (Fig 7.21).



(a) MultiFlatGT (b) $KS$ NMS Result MultiFlat

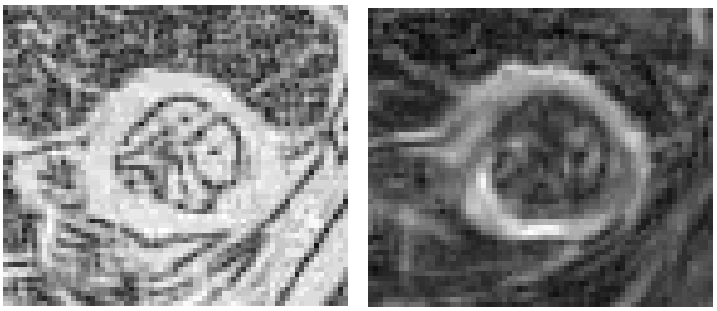
Figure 7.18: $KS$ cross sectional result from MultiFlat image after NMS is applied. This shows that despite a noisy response, the $KS$ method achieves good detection, with many of the interfaces detected.

#### 7.4.2.2 Observations

The high sensitivity characteristics of the $KS$ method were also observed with the real MRI data. The method was able to resolve solid tumour interfaces between white matter tissue and cyst material for each of the MRI modalities, however T1-weighted images produces more over-detection artefacts compared to the other modalities relative to the $\chi^2$, $RRO$ and $u$-test non-parametric methods which also provide a normalised output. On T2-weighted and T1 contrast enhanced imagery, the $KS$ would not be considered the optimal method, due to the sensitivity of the method and over-detection artefacts.

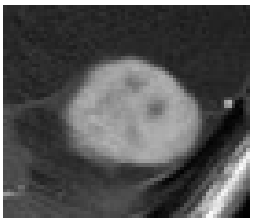
The effectiveness of the non-parametric tests correlates with the findings in the work of Williams et al. (2014). In that study an improved performance was observed when a larger scale parameter was applied. This is due to increases in the number of pixels in each sample region of the mask as the scale parameter is increased. However the primary drawback of a larger mask size in 2-D is the introduction of localisation error. By using a 3-D statistical filter the number of voxels processed in the mask increases by a factor of $v^2$ in comparison to a 2-D filter with the same scale parameter. Therefore 3-D statistical filtering methods provide increased resolving power at a comparable scale, negating the requirement for always applying a large neighbourhood mask. This effect can be observed in Fig 7.19, here the interface of the tumour boundary is resolved

by both 2-D and 3-D non parametric statistical filters. The signal to noise ratio of the detected surfaces is improved with 3-D filtering, signified by greater delineation of the tumour boundary from the background in 3-D compared with 2-D at the same scale.

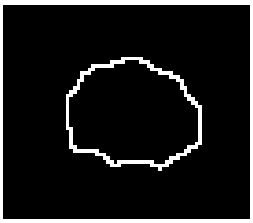


(a) 2D KS (b) 3D KS

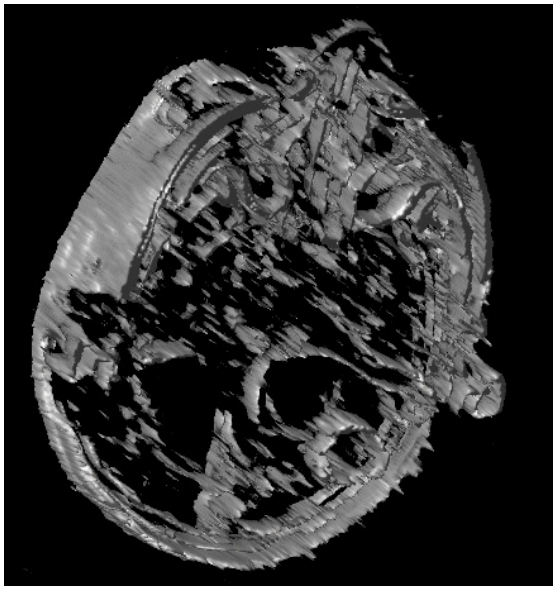
Figure 7.19: Comparison between 2-D and 3-D KS test at equivalent scale parameter. a) -2D 5x5 and b) 3-D 5x5x5



(a) 2-D layer



(b) 2-D Reference

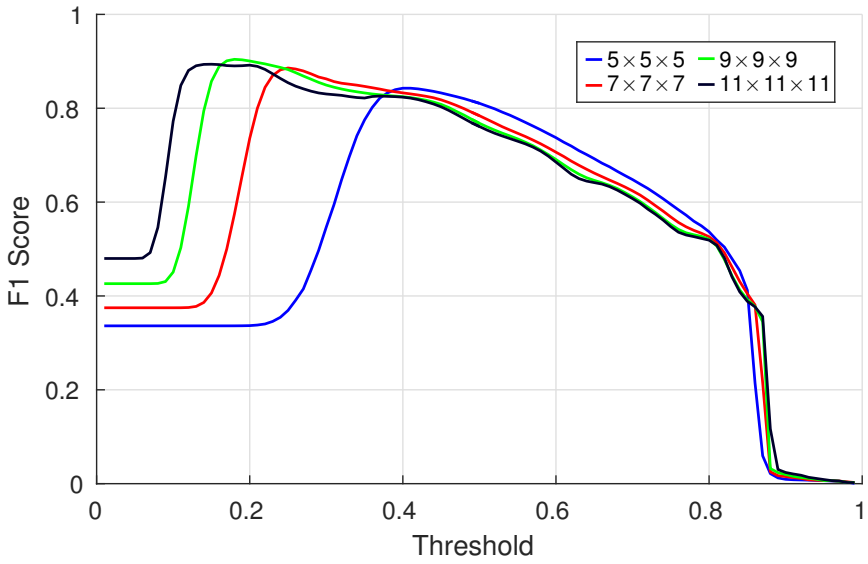


(c) 3-D

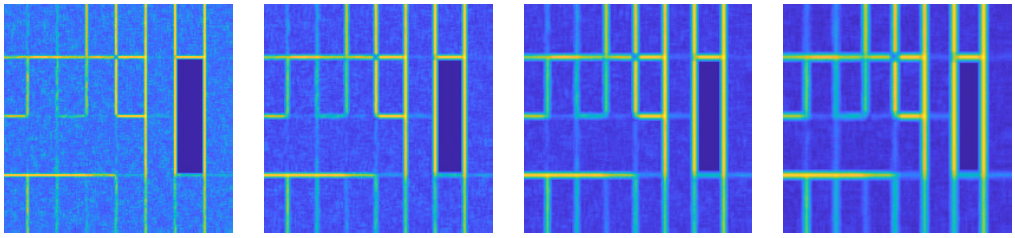
Figure 7.20: 3-D visual $KS$ result on T1-Weighted MRI image

#### 7.4.2.3 Scale Response

In the work of Williams et al. (2014), the observed characteristics of the 2-D KS method were also present in 3-D. The $KS$ method is a very sensitive detector that achieved good detection but with many spurious edges being resolved. It was shown that increasing the scale parameter of the method suppresses the noise of the algorithm, this also applies in 3-D (Fig 7.21).




(a) F1-Score



(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.21: Comparison of neighbourhood scales for $KS$-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages. Higher surface signal to noise ratio at larger scale due to suppressed noise signified by darker regions between surfaces. However, greater uncertainty in surface location, signified by a 'smeared' surface interface

For real MRI data, the $KS$ method can be used to resolve the solid tumour interfaces, the cyst interfaces and the normal tissues interfaces, but it also detects many more structures which are not detected by the other methods. This was characteristic across the different modalities (Figures 6.9, 6.13, 6.17l). This can be disadvantageous for normal segmentation tasks, as this could be considered noise or spurious responses. However, the increased sensitivity reveals structure about the tumour which could be of use, therefore the KS test should not be excluded as a potential option.

| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, some over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts | Accurate detection,connected surface. some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved and connected with some loss of detail, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Good detection, moderate noise, good connectivity, strong magnitude surfaces | Good detection, moderate noise, good connectivity, strong magnitude surfaces |

| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Detected - partial | High | Good detection but significant clutter |
| 5 (T2), Section 6.8 | Detected | Detected | Detected | High | sensitive, reveals noise |
| 6 (T1c), Section 6.9 | Detected | Detected | Detected | High | Sensitive to noise, low magnitude surfaces |

Table 7.7: Wilcoxon Mann-Whitney test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

### 7.4.3 Wilcoxon Mann-Whitney Test $u$-test

#### 7.4.3.1 Surface Response

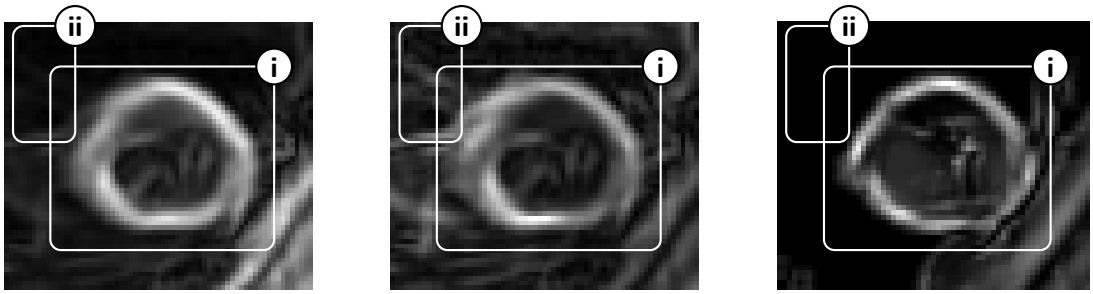
The non-parametric Wilcoxon Mann–Whitney test is commonly used for detecting differences in central tendency between two samples. As the method is evaluating differences in central tendency between distributions, it is expected that the method would be successful for mean based surface interfaces. The work of Williams et al. (2014) found the 2-D $u$-test to be particularly effective for mean based edges.

The Mann Whitney $u$-test performed strongly on the intensity interface type and the combination images (Figures 5.9k, 5.11k and 5.9k ). However, the $u$-test does not resolve surfaces of between textured regions (Fig 5.10k). Similar to the $KS$ method, the ranking nature of the technique locally normalises the output. This results in a sensitive detection method with a propensity to create over-detection artefacts. This high sensitivity reduces the signal to noise ratio with respect to the detection of the surface interface. These characteristics were observed across evaluations as shown in Table 7.7, signified by good detection across interface types, but with high amounts of noise.

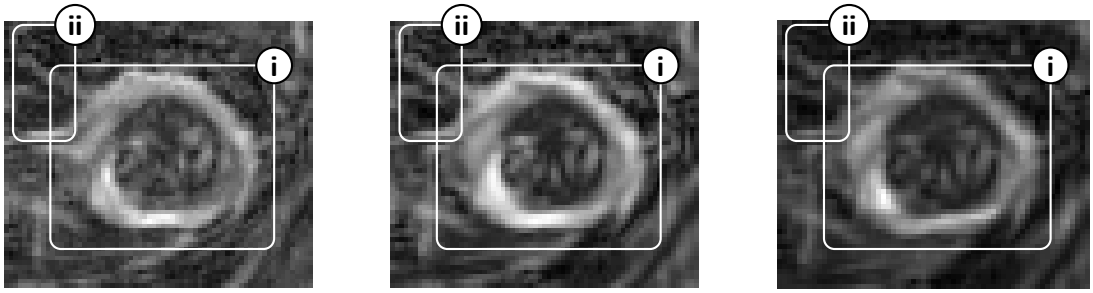
However, on the multiple statistic images, the $u$-test proved to be an effective method outperforming the many of the statistical methods and significantly outperforming the baseline Canny and Steerable methods (Fig 5.31). The objective results indicate that while the $u$-test does not achieve the highest peak score on the MultiFlat and MultiCurve image volumes it can operate over a wider threshold range than the other methods (Fig 5.31). This reliability makes the $u$-test suitable when the image characteristics are not known a-priori.

#### 7.4.3.2 Observations

Subjectively, the $u$-test method was third best for T1-weighted MRI modalities behind the $\chi^2$ and $t$-test methods as the signal to noise ratio of the surfaces is higher than that of the $KS$ and $RRO$ methods, while achieving better detection than the baseline and other statistical methods(Fig 6.7). For T2-weighted imagery, the $u$-test does produce over-detection artefacts due to the high sensitivity of the method, however, it does produce surfaces with a higher magnitude than the clutter (Fig 6.11) and is greatly outperformed by the parametric $DoB$ and $t$-test methods which are characteristically less noisy in their outputs. For contrast enhanced T1-weighted imagery, the $u$-test produces strong magnitude outputs of the tumour structure, however there are still more noise artefacts produced when compared against the $\chi^2$, $t$-test and $DoB$ methods (Fig 7.22).

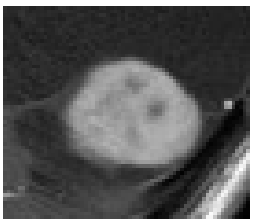



(a) *DoB* (b) $t$-test (c) $\chi^2$

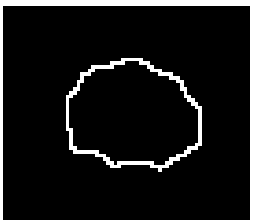



(d) *KS* (e) $u$-test (f) *RRO*

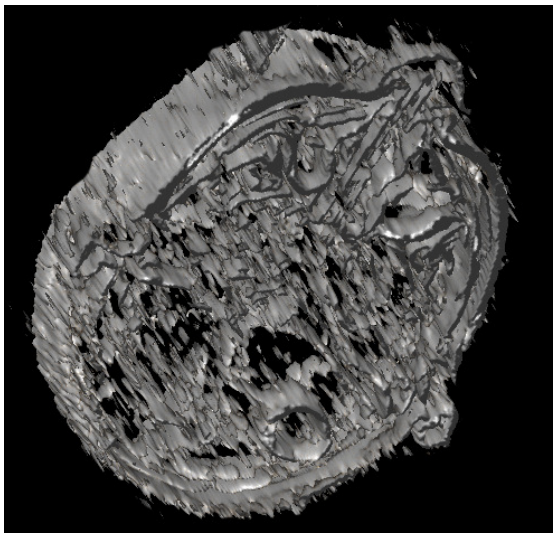
Figure 7.22: Characteristics of statistical filters with successful responses on T1-contrast images. $u$-test method, *KS*, *RRO* have similar strong surface responses (i), however compared against other methods such as the *DoB*, $t$-test and $\chi^2$-test there is increased sensitivity in the response, signified by clutter in region (ii)



(a) 2-D layer



(b) 2-D Reference

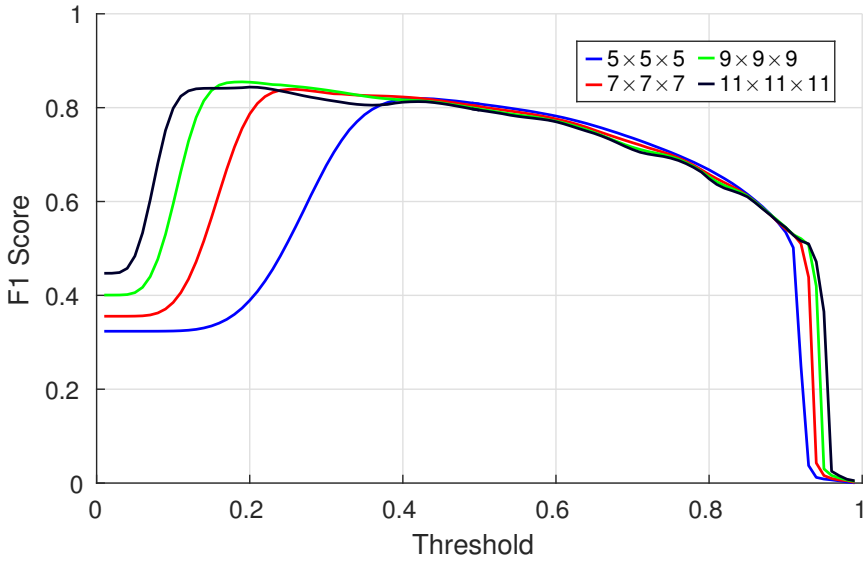


(c) 3-D

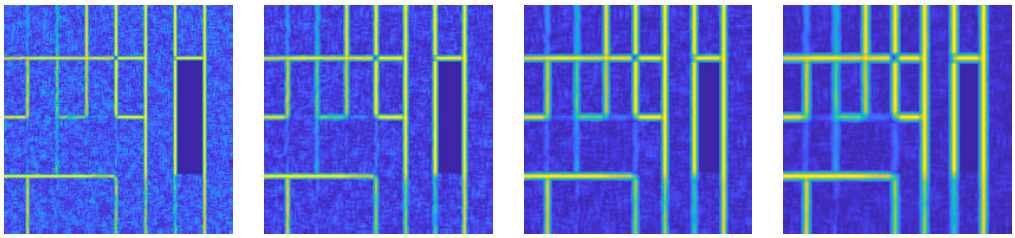
Figure 7.23: 3-D visual $u$-test result on T1-Weighted MRI image

#### 7.4.3.3 Scale Response

Increasing the scale of the neighbourhood mask increases the signal to noise ratio of the interface (Fig 7.24), and suppresses to the noise to a greater extent than smaller mask sizes. However, the larger mask sizes introduce greater uncertainty in the location of the detected surfaces.




(a) F1-Score



(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.24: Comparison of neighbourhood scales for $u$-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages. Higher surface signal to noise ratio at larger scale due to suppressed noise signified by darker regions between surfaces. However, greater uncertainty in surface location, signified by a 'smeared' surface interface

Overall the $u$-test performed similarly to the $KS$ and $RRO$ methods, with the $u$-test providing marginally better results on the MRI data. When weaker surfaces require detection, the $u$-test is a good option due to its high sensitivity, offering a slightly improved signal to noise ratio than the $KS$ and $RRO$ methods.

### 7.4.4 Robust Rank Order Test (*RRO*)

#### 7.4.4.1 Surface Response

The Robust rank order is a statistic similar to $u$-test method which was developed by Fligner and Policello (1981), it was later adapted as an edge detection method by Lim (2006). The method does indeed produce visually comparable and objectively comparable results to the $u$-test, but also the $KS$ method. An overview of the each evaluation is presented in Table 7.8, and the general characteristics reveal a sensitive filter with good detection, but generally some noise in the response.

On the intensity (Int) based interfaces the *RRO* method performed strongly (Figures 5.9,5.14,5.19). However the method failed to resolve many of the Tex type interfaces in evaluations 1 and 2 (Figures 5.10,5.15,5.20,5.25). A sensitive detector with a tendency to produce spurious responses, the *RRO* method performed similarly to the other non-parametric tests. On images with multiple statistical regions the *RRO* method was the 3rd highest scoring method behind the $\chi^2$ and KS methods (Fig 5.31).

#### 7.4.4.2 Observations

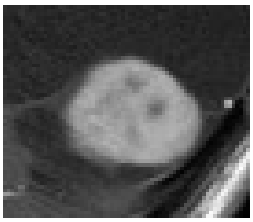
On T1-weighted imagery, the *RRO* method produced results comparable to the $KS$ method, characterised by poor surface signal to noise ratio due to the sensitivity of the method (Fig 6.8l,k). For T2-weighted imagery the *RRO* method has improved surface signal to noise ratio, however it does not perform as well as other statistical methods, such as the $DoB$, $t$-test and $\chi^2$ methods (Fig 6.12e,h,i,k), which produce more accurate connected surfaces with fewer spurious responses. For contrast enhanced T1-weighted MRI data, the *RRO* method can detect some surfaces, but again the surface signal to noise ratio is poorer than the $DoB$, $t$-test and $\chi^2$ methods (Fig 6.17e,h,i,k).

| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, some over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts, weaker corners | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some over-detection artefacts, weaker corner strength | Accurate detection,connected surface. some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Surface resolved, connected surface, some inaccuracies, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved and connected with some loss of detail, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |

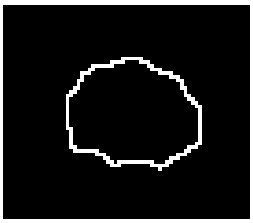
| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Good detection, considerable noise, good connectivity, strong magnitude surfaces | Good detection, low-moderate noise, good connectivity, strong magnitude surfaces |

| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Not detected | High | Partial detection, but cluttered |
| 5 (T2), Section 6.8 | Detected | Detected | Detected | High | Low magnitude surfaces |
| 6 (T1c), Section 6.9 | Detected | Detected | Detected | High | Sensitive to noise,low magnitude surfaces |

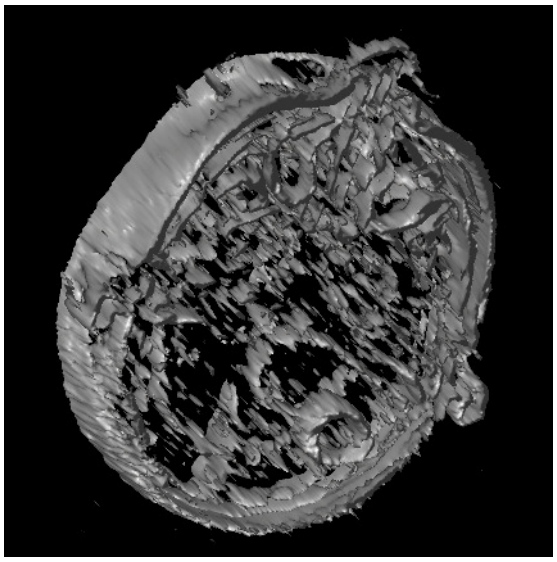
Table 7.8: Robust Rank Order test filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.



(a) 2-D layer



(b) 2-D Reference

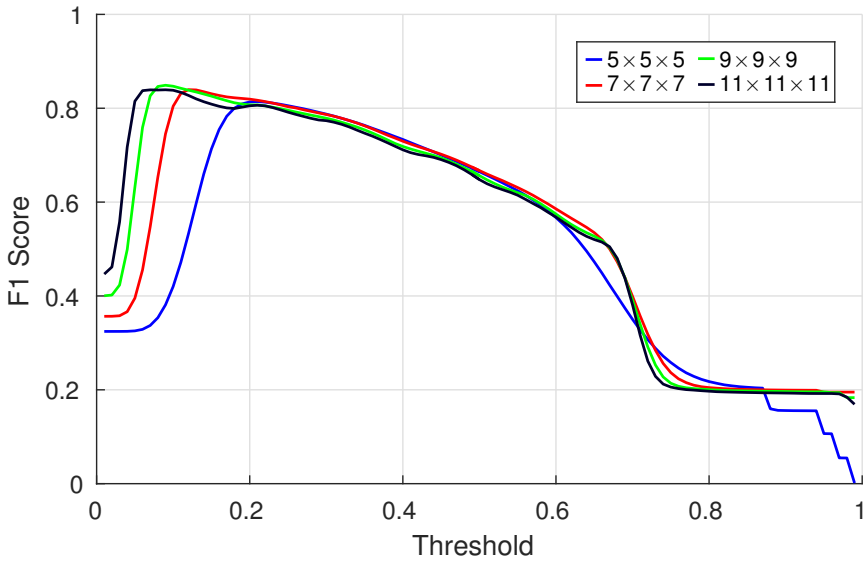


(c) 3-D

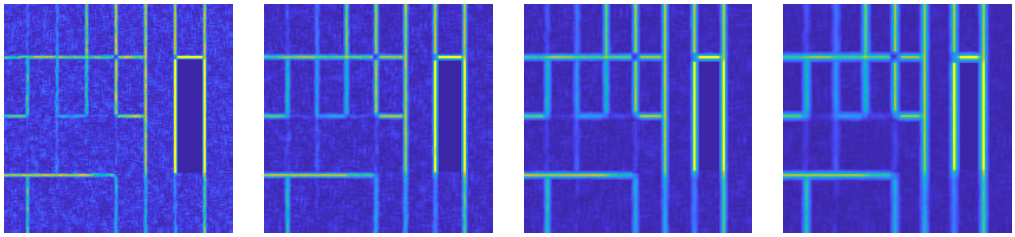
Figure 7.25: 3-D visual *RRO* result on T1-Weighted MRI image

#### 7.4.4.3 Scale Response

At larger scales the *RRO* method is significantly more computationally expensive than the other non parametric methods, however the method saw no improvements in performance beyond mask sizes larger than $7 \times 7 \times 7$ for the multiple statistical region image (Fig 7.26a).




(a) F1-Score



(b) 5×5×5 (c) 7×7×7 (d) 9×9×9 (e) 11×11×11

Figure 7.26: Comparison of neighbourhood scales for *RRO*-test. a) F-1 score performance metric. b,c,d,e) Shows central 2-D layer taken from 3-D surface map volume using different scale neighbourhoods, prior to post processing stages. Higher surface signal to noise ratio at larger scale due to suppressed noise signified by darker regions between surfaces. However, greater uncertainty in surface location, signified by a 'smeared' surface interface

While Lim (2006) found the *RRO* edge detector outperformed the $u$-test, $t$-test and the Canny Edge detection methods, when used in the context of a 3-D dual region filter for surface detection, in this evaluation the $t$-test proved to be an overall more effective method. Instead the *RRO* was found to perform similarly to the $u$-test method but with a smaller signal to noise ratio, this can be observed in the objective results where the effective threshold range is wider (Figures 7.24a and 7.26a). Additionally the *RRO* method comes with a much higher computational cost. Due to the computational cost off the *RRO* method, for MRI data the $u$-test would be the preferential option if *RRO* surface map characteristics are desired (Table 3.1).

## 7.5 Baseline Methods

In this work, the 3-D statistical methods of surface detection were compared against 2 baseline methods. The 3D canny method, adapted from the 2-D Canny edge filter (Canny, 1986), and the 3-D steerable filter from the work of Aguet et al. (2005).

### 7.5.1 3-D Canny

#### 7.5.1.1 Surface Response

The Canny method is known for being an effective method at resolving interfaces where there is a step change in brightness. Table 7.9 reveals that across the evaluations the method is indeed effective at intensity based interfaces. However it was also shown that statistical methods perform equally well on these kinds of interface, notably the $DoB$, $t$-test, and $\chi^2$ methods. When there is not a strong intensity component or, when the there is a strong variance component to one or more of the regions, the Canny method loses its effectiveness, and achieves inferior performances relative to many of the statistical approaches.

The Canny method produced accurate detection on Com1 interface types (Fig 5.11), however again the $DoB$ and $t$-test methods also achieved accurate detection but with improved noise suppression. On the Com2Fl interface type the Canny method was able to achieve optimal detection, but in a much narrower threshold range (Fig 5.13c,d, and was significantly outperformed by a number of statistical methods including $DoB$, $L$, $\chi 2$, $t$ and $u$-test methods. Predictably the Canny method performed poorly on the variance type interfaces (Fig 5.13b). This is because as variance within a region presents as abrupt changes in brightness contained within a region, these abrupt changes are themselves interpreted to be a surface, instead of the interface between regions of different variance. This leads to the production of clutter and noise within the high variance region, whereas some statistical methods are not impacted by this to the same extent.

#### 7.5.1.2 Observations

The performance of the 3-D Canny methods on the real images indicates that it does produce accurate surface maps when there is a strong delineation between solid tumour and grey matter in the image volume in terms of brightness between regions. While this may be the case in T2-weighted images, this is not typical of T1-weighted images, which is the modality where the Canny method was outperformed to a greater extent by some of the statistical methods, namely the $t$-test and $\chi^2$ methods. But even in the case of T2-weighted imagery, where the Canny is more effective compared to T1, the 3-D Canny method was still outperformed by the $\chi^2$, $DoB$ and $t$-test methods (Figures 6.11, 6.12

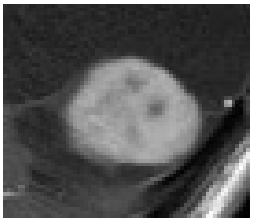
| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Accurate detection, connected surface, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, localised over-detection artefacts | Accurate detection, connected surface, over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Accurate detection, connected surface, no over-detection artefacts | No surface, over-detection artefacts | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, no over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Accurate detection, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface some regional over-detection artefacts | Accurate detec-tion,connected surface. some over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Surface resolved, connected surface, some inaccuracies, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts | Surface resolved with finer detail lost, some over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Missed surfaces, some noise, good connectivity, | Missed surfaces, good noise suppression, low magnitude surfaces |

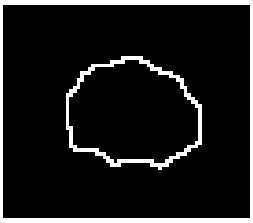
| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Detected | Detected | Detected | Low | Good detection but weak interface |
| 5 (T2), Section 6.8 | Detected | Detected -weak | Detected | Low | Weak interfaces in $x, y$-plane |
| 6 (T1c), Section 6.9 | Detected | Detected -weak | Detected | low | weak interfaces in $x, y$-plane relative to $z$ |

Table 7.9: 3-D Canny Filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

6.13). On contrast enhanced T1-weighted imagery the Canny method was again less successful than the $\chi 2$, $DoB$ and $t$-test methods (Figures 6.15, 6.16 6.17) but it was able to resolve some of the boundary information as illustrated in the 3-D result in Fig 7.27.



(a) 2-D layer



(b) 2-D Reference

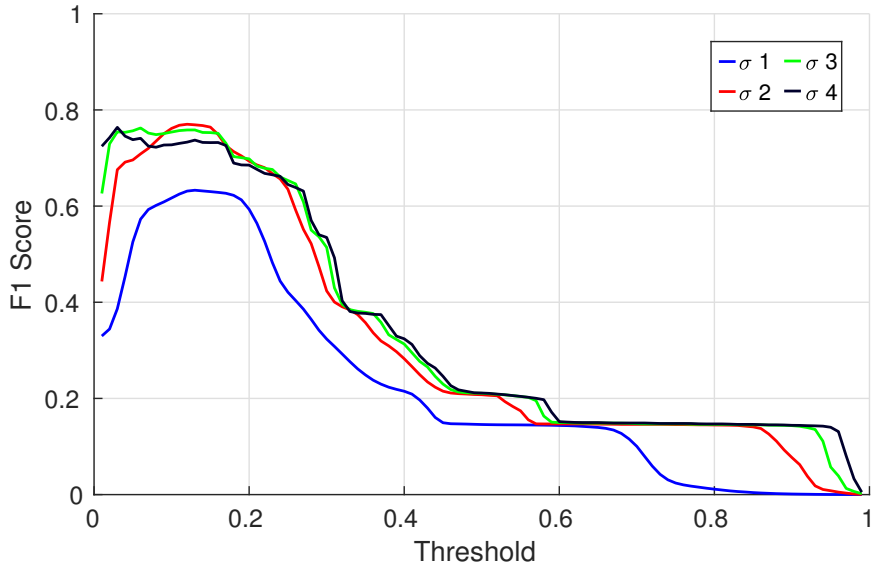
(c) 3-D

Figure 7.27: 3-D visual 3-D Canny result on T1-Weighted MRI image

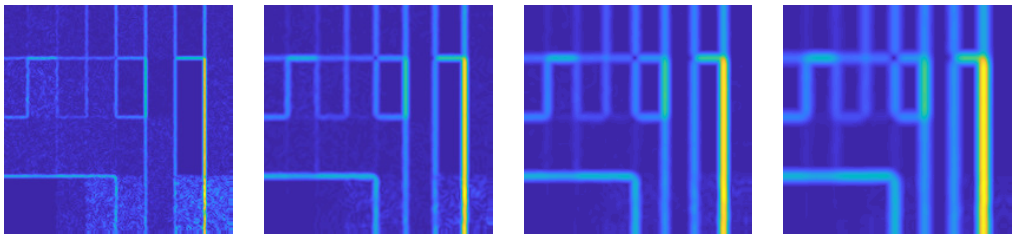
A further drawback of the 3-D Canny method against the 3-D statistical methods in reference to resolving interfaces based on brightness, is the lack of option to weight the mask orientations, or adjust the mask dimensions based on image resolution. This had pronounced negative impact on evaluations 4-6 and the problem is illustrated in Fig 6.19. We make recommendations to adapt the 3-D Canny method to account for this in line with the work of Brejl and Sonka (2000), particularly at larger $\sigma$ values, where the neighbourhood mask is larger and the distortion is exacerbated. Or to only use 3-D implementations of these algorithms on either isotropic or interpolated 3-D data.

#### 7.5.1.3 Scale Response

At increased Gaussian filter kernel sizes the Canny method exhibit the similar characteristics of the increased neighbourhood size for the statistical measures. Such as improved noise suppression offset by greater uncertainty in the surface accuracy (Figures 7.28)




(a) F1-Score



(b) $\sigma 1$ (c) $\sigma 2$ (d) $\sigma 3$ (e) $\sigma 4$

Figure 7.28: Comparison of different sized Gaussian filter kernels with 3-D Canny method. Higher surface signal to noise ratio at larger scale due to suppressed noise signified by darker regions between surfaces. However, greater uncertainty in surface location, signified by a 'smeared' surface interface

### 7.5.2 3-D Steerable Filter

#### 7.5.2.1 Surface Response

The 3-D Steerable filter method is optimised resolving interfaces where there is a step change in brightness. Table 7.10 reveals that across the evaluations the method is most effective at resolving intensity based interfaces. When there is not a strong intensity component or, when the there is a strong variance component to one or more of the regions, the Steerable method loses its effectiveness, tending to produce noisy and cluttered surface maps which results in inferior performances relative to many of the statistical approaches in addition to the 3-D approach.

#### 7.5.2.2 Observations

The steerable method is optimised for detecting changes in brightness and did produce good visual results in Exp.1 IntFL (Fig. 5.9), however the method conversely produced

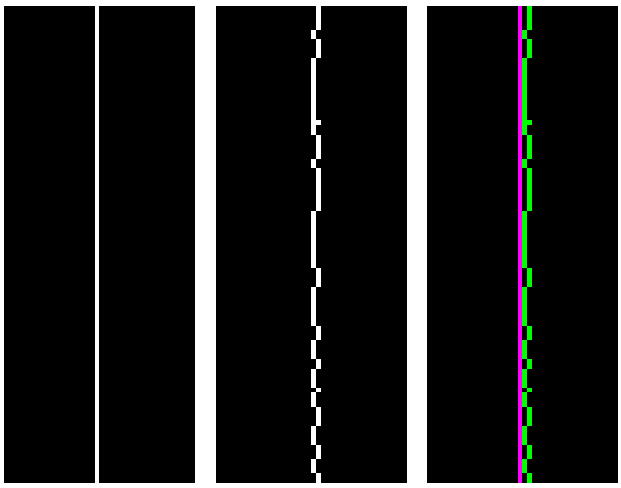
| Evaluation | Int | Tex | Com1 | Com2 |
|---|---|---|---|---|
| 1 (Fl), Section 5.5 | Connected surface, displaced, some over-detection artefacts | Did not resolve surface, over-detection artefacts | Accurate detection, connected surface, some over-detection artefacts | Connected surface, displaced, severe over-detection artefacts |
| 2.1 (Cu), Section 5.6.2 | Displaced surface, connected surface, localised inaccuracies at corners, some over-detection artefacts | No surface, over-detection artefacts | Displaced surface, connected surface, some over-detection artefacts | No resolved surface, over-detection artefacts |
| 2.2 (Sp), Section 5.6.3 | Small displacement, connected surface, some localised artefacts near surface and some over-detection artefacts | No surface resolved, over-detection artefacts | Accurate detection, connected surface, poor surface resolution, over-detection artefacts | poor differentiation between surface and over-detection artefacts |
| 2.3 (CSt), Section 5.6.4 | Displaced surface, connected surface, some over-detection artefacts | No surface resolved, over-detection artefacts | connected surface surface,but some duplicates and positional inaccuracies, over-detection artefacts | Inaccurate surface, connected surface, poor differentiation from over-detection artefacts |
| 2.3 (FSt), Section 5.6.4 | Surface resolved, connected surface, some inaccuracies, some over-detection artefacts | No surface resolved, over-detection artefacts | Surface resolved, inaccurate location, duplicated surface response, over-detection artefacts | Surface resolved with finer detail lost and slight displacement, over-detection artefacts |

| Evaluation | Flat | Curve |
|---|---|---|
| 3 (Multi), Section 5.7 | Missed surfaces, considerable noise, low magnitude surfaces | Poor detection, considerable noise, weak surfaces |

| Evaluation | $T-B$ | $C-B$ | $T-C$ | Noise or Clutter | Characteristics |
|---|---|---|---|---|---|
| 4 (T1), Section 6.7 | Not detected | Not detected | Not detected | low | poor detection of axial interfaces |
| 5 (T2), Section 6.8 | Not detected | Detected -weak | Detected | medium | weak interfaces in $x, y$-plane |
| 6 (T1c), Section 6.9 | Not detected | Detected -weak | Detected | medium | weak interfaces in $x, y$-plane |

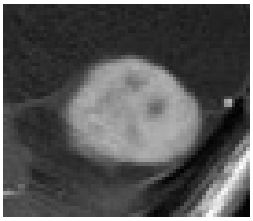
Table 7.10: 3-D Steerable Filter characteristics cross comparison of evaluation summaries through Evaluations 1-6.

a poor objective score (Fig 5.13a). The quantitative metric score indicated some failure of the surface result which was not clearly visually apparent. Here the discrepancy lies with the localisation of the surface. The steerable filter produced a surface with 2 voxel displacement (Fig. 7.29). Often a connected but displaced surface is desirable over a fragmented but accurate surface, thus this displacement is not a significant drawback as the performance results may suggest, with a greater $T$-match zone applied within the performance metric, improved scores would be expected. However overall the Steerable method was still generally outperformed by the statistical methods, frequently being unable to resolve the surface of interest, as illustrated in the 3-D representation in Fig 7.30.

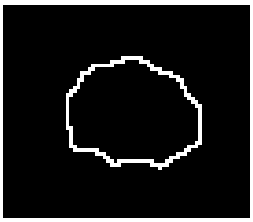


(a) Reference (b) Steerable Result (c) Comparison

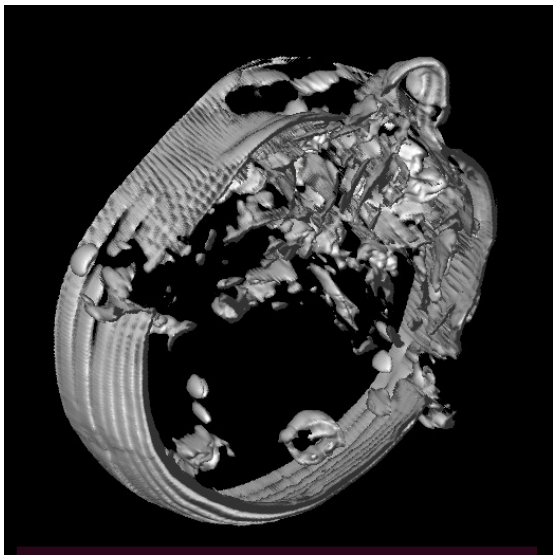
Figure 7.29: 2D comparison of centre layer from steerable filter output. Comparison against Reference indicates small displacement of the surface which contributes to the negative performance metric score.



(a) 2-D layer



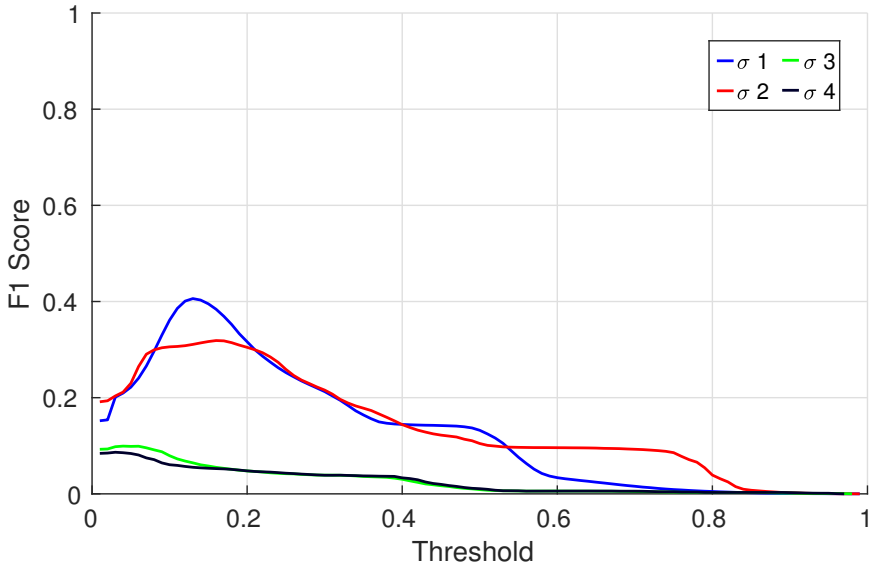
(b) 2-D Reference



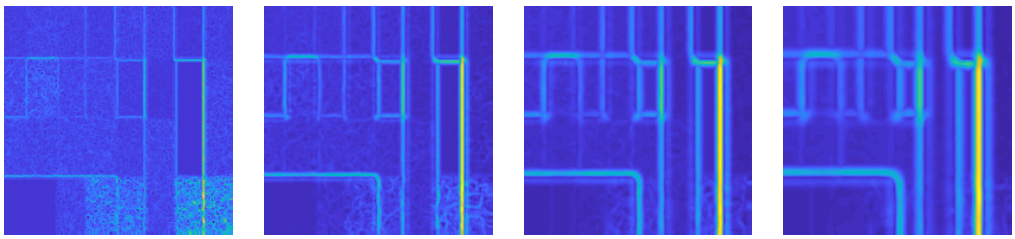
(c) 3-D

Figure 7.30: 3-D visual Steerable result on T1-Weighted MRI image

#### 7.5.2.3 Scale Response




(a) F1-Score



(b) $\sigma 1$ (c) $\sigma 2$ (d) $\sigma 3$ (e) $\sigma 4$

Figure 7.31: Comparison of different sized Gaussian filter kernels with 3-D Steerable filter method. Higher surface signal to noise ratio at larger scale due to suppressed noise signified by darker regions between surfaces. However, greater uncertainty in surface location, signified by a 'smeared' surface interface, in addition to weak performance on corners and T-junctions which lose their regularity. Larger scales also introduce greater displacement of surface

At increased Gaussian filter kernel sizes the Steerable filter method exhibited the similar characteristics of the increased neighbourhood size for the statistical measures. Such as improved noise suppression offset by greater uncertainty in the surface accuracy (Fig 7.31. However the Steerable filter exhibited negative artefacts with increased Gaussian smoothing, particularly in the detection of corners and T-junctions.

## 7.6 Summary

In this discussion, the generalisable trends found in Chapters 5 and 6 are established. For each filter method, the surface response, scale response and observations of the

filter characteristics are presented. The key findings from the objective and qualitative evaluations were extrapolated into summary tables which allow for ease of comparison between evaluations. Through examination of the trends, the generalisability and transferability of the evaluation results were identified. The findings support a strong link between the characteristics of the synthetic results with that of the MRI data, indicating good transferability, therefore establishing that the objective analysis can reliably offer insight to the likely performance of each statistical filter in a real application domain.

In general it was shown that statistical methods, notably the $t$-test and $\chi^2$ test either match or outperform both baseline Canny and Steerable methods across all evaluations, while nearly all statistical methods improve upon the baseline techniques in high variance images. This can be observed by the performance of the baseline methods on all texture type interfaces (TexFl, TexCu, TexSp,TexCSt, TexFSt) relative to the statistical methods. This finding is also present in the Com2 interface results and the MultiFlat and MultiCurve assessments in evaluation 3.

In the discussion, the effects of the scale parameter are also observed. Generally across all filtering methods, as the neighbourhood scale is increased, the level of noise suppression is also improved, and the detection of the surface is more strongly resolved. However, increasing the scale, spreads out the surface response across more voxels, this leads to some uncertainty in the true surface location, and after NMS is applied, there is more localisation error in the result which offsets some of the advantages gained, and therefore differences in objective performance is marginal. However, on the MultiFlat scale analysis the $\chi^2$, $F$, $L$ and Steerable filter methods favoured a small scale, while the other statistical filters and the Canny method favoured a larger scale.

On the MRI data, the statistical methods consistently outperformed the baseline Canny and Steerable methods, with the parametric tests, notably the $DoB$ and $t$-tests offering improved noise suppression, while the non-parametric tests offered improved detection and connectivity, but at the expense of more noise in the surface map output. The best performing method was the $\chi^2$ method, which was able to achieve a strong signal to noise ratio in the output, signified by good noise suppression, but it was also able to combine this with good detection and surface connectivity.

A notable difference between results from evaluations 1-3 with that of evaluations 4-6, is the presence of $z$-spacing biasing in the results. As the MRI data is anisotropic in its resolution (Tab 6.1) which is typical for real patient cases, the baseline methods did not account for this in their design and as a consequence produced artefacts which negatively impacted the performance. These artefacts are illustrated in Fig 6.19, signified by a

shaded 2-D region and the effect is present throughout all patient cases in evaluations 4-6. The Vector Magnitude which incorporates scaling of the directional components of the statistical vector can largely minimise this effect, for superior results in the surface map outputs.

In the following and final chapter, the contributions found here and throughout are summarised and conclusions are drawn.

# Chapter 8

# Conclusions

## 8.1 Contributions

This chapter provides a conclusion of the different aspects of the work carried for this thesis, summarising each chapter and the contributions offered throughout.

### 8.1.1 Two Novel Approaches to Statistical Surface Detection

Chapters 1 and 2 review advancements in edge and surface detection. Existing work has shown that statistical methods for detecting edges to be superior over traditional methods when factors such as texture and noise are present in the image data. However, reviewing the literature revealed that development of 3-D statistical approaches to surface detection remained largely unexplored. This led to the development of two novel non-linear methods of surface detection which employ oriented, dual region statistical filters. Which were then comprehensively analysed to see if the advantages offered by statistical edge detection can be utilised in 3-D data. Also introduced is a novel 3-D orientation filtering method for 3-D non maximum suppression (Chapter 3).

The two distinct surface detection architectures were developed and were presented in Chapter 3, these include:

- Maximum Response method
- Vector Magnitude method

#### 8.1.1.1 Maximum Response Method

The maximum response method exists in 2-D, and has previously been comprehensively evaluated by Williams et al. (2014), in this thesis a novel adaptation of the 2-D statistical edge detection architecture is made in order to reconfigure the technique for 3-D surface detection and is a new contribution within this work. The Maximum Response method resolves surface interfaces by applying a rotating dual region voxel neighbour-

hood to the image and analysing the statistical properties of the regions covered by the mask. The number of orientations covered by this process is determined by the scale of the neighbourhood, but this implementation used a constrained case of 13-orientations where all surface directions are truncated to a 26-connectivity mask. The magnitude of the surface is determined to be the maximum response of the computation of an interchangeable statistical test applied across the 13 orientations. This technique determines the maximum statistical difference through different neighbourhood mask positions and tends to produce surfaces with a high magnitude. This is advantageous in terms of signal to noise ratio for the parametric tests which offer good noise suppression, but for the more sensitive non-parametric tests, this advantage is offset due to the noise also being of high magnitude. Computationally, the method is expensive if the number of orientations is not constrained, since as the neighbourhood mask scale is increased, the number of viable orientations which are available is also increased, due to the exponential increase computational cost with regard to scale, the unconstrained MR method is not a scaleable technique in practice.

#### 8.1.1.2 Vector Magnitude Method

The second method developed in this work is the Vector magnitude method. This architecture was developed primarily to improve the efficiency of statistical surface detection in 3-D data. The vector magnitude achieves this by fixing the number of region orientations used within the neighbourhood to the 3 orthogonal planes defined by 3-D data ($x$,$y$ and $z$ planes). By performing a statistical comparison in 3 orthogonal directions, 3 directional magnitude components are acquired. These components form a vector which describe both the overall magnitude and direction of a statistical differential. This work shows there are multiple benefits to the VM approach.

When the image resolution is anisotropic, applying surface detection using an isotropic mask distorts the surface map result, biasing the output toward surfaces which exist in the direction of the imaging plane with the smaller resolution. This is a common issue in real 3-D data and the effects of which should not be overlooked. In MRI data for example, the spatial resolution in the $z$-dimension of an axial viewed image volume is often lower than in the $x$ and $y$ directions, thus the voxel accounts for more physical space in the $z$-direction. This was also shown within the thesis in the MRI dataset (Table 6.1). Anisotropic resolution, where the $x$ and $y$ voxel resolution is superior to the $z$-direction, increases the likelihood of bigger shifts in image properties between voxels in adjacent layers, compared with adjacent voxels in the same image layer. A bigger change in image properties typically results in a larger differential computed by the surface detection method, which results in surfaces computed with a stronger $z$-component

than what would be expected from an isotropic image which has no inherent directional bias in the spatial properties. The VM architecture allows for the offsetting of this bias by applying coefficients which appropriately weight the directional components of the statistical vector based on the image resolution, prior to the L2-norm calculation. This in turn proportionately reduces the influence of the $z$-component in the calculation for a more balanced surface map. The benefit of this is that erroneous surface points in the surface map which manifest as filled regions can be avoided, which ws not the case for the baseline Canny and Steerable techniques as demonstrated in evaluations 4-6. This effect is best illustrated in Fig 6.19. The Vector Magnitude method was shown to offset this bias to achieve superior results in anisotropic MRI data compared with the baseline approaches.

In terms of efficiency the Vector Magnitude greatly improves over the Maximum response method, this can be seen in the results presented in Table 3.1, showing that the VM completion time is significantly shorter than that of the MR method. This is due to the fact the statistical test is always calculated 3 times in each location instead of 13. Additionally, as the scale of the mask is increased, there is no loss in accuracy in the directional component of the VM measurement, this is in contrast to the MR method, which either has to use a constrained set of orientations resulting in greater error magnitude as the scale increases, or by applying more orientations in the measurement, which would further increase the computational cost of the method. The impact of this are surfaces in results which are more seamless when obtained by the VM method compared with the MR method. These advantages make the VM the preferential choice of the two architectures, and was the method evaluated in Chapters 5 and 6.

### 8.1.2 Evaluation Methodology Framework

Chapters 1 and 2 present a strong case for undertaking an evaluation of statistical surface detection, exposing the present gap in literature and knowledge in the appraisal of these techniques. Currently there is no formal evaluation of 3-D statistical methods for surface detection, however, evaluation of surface detection methods is a non trivial process.

Chapter 4 explores the literature for the most appropriate methodologies for objectively evaluating surface detection methods. A number of issues have previously been raised such as those from Forbes and Draper (2000) and Bowyer et al. (2001) showing that objective measures and real world results do not always correlate due to the disparity in complexity between real imagery and synthetic data. Therefore in this work, a combined objective and qualitative approach was undertaken.

The criticisms associated with synthetic data analysis refer to the overall lack of complexity in the dataset interfaces. Notably the lack of topological considerations, specifically the lack of curvature, but also the relative strengths of interfaces. This observation led to a 3 stage approach within the objective methodology detailed in Chapter 5 and were as follows:

1. Evaluation 1: Measurement of the resolving power of statistical surface detection on interfaces between regions of different region profiles.
2. Evaluation 2: Measurement of effects of surface topology on the detection performance.
3. Evaluation 3: Measurement of detection performance with data that possesses multiple distinct region profiles with different strength interfaces.

Following the objective assessment, the filters were analysed qualitatively in a real case assessment in Chapter 6. Chapter 7 provides a discussion of the characteristics of the filters on both the synthetic and real image volumes. Comparisons are made and reveal that many of the filtering characteristics which occur in the synthetic domain also manifested in the results on real data. Indicating that objective analysis can inform performance to a degree for specific applications.

### 8.1.3 Efficient Paring Strategy

In the literature, F-measure analysis has been shown to be the most appropriate and reliable objective measure for accurately assessing surface detection performance (Pont-Tuset and Marques, 2016). Correct implementation of the F-measure metric requires proper collection of the TP, FP, TN and FN figures established through one to one correspondence matching of the filter results with the reference images as recommended by Liu and Haralick (2000).

Accurate correspondence matching can be solved for 2-D data alone using the Hungarian algorithm (Kuhn, 1955). However this is a brute force process which is very computationally expensive. The complexity of the problem is dependent on the size of the image and the number of edge or points present in the image. Due to these factors, in 3-D data the computational complexity is even greater, making the Hungarian algorithm impractical at currently available computational speeds for large datasets. This led to the development of the Efficient Paring Strategy (EPS) correspondence matching technique and its development is detailed in Chapter 4.

The EPS technique approximates very closely the Hungarian algorithm result in terms of accuracy with a Pearson Pairwise Correlation coefficient of 0.99, however it is solved with quadratic time complexity ($O(n^2)$) improving over the exponential time complexity

($O(c^n),\ c > 1$) achieved by Hungarian Algorithm variations (Fig 4.16). This allows for both the evaluation of large 3-D image volumes and the evaluation of large datasets with a greatly reduced completion time.

The EPS method is introduced and tested in Chapter 4, and then in Chapter 5 used as a novel contribution to evaluate objective response of surface detection methods.

### 8.1.4 Objective Evaluation of Surface Detection

This work has presented several techniques for statistical surface detection. Presented in Chapter 5, using the evaluation framework, a novel contribution in this thesis, the statistical tests were objectively evaluated with a direct comparison against the baseline 3-D Canny and Steerable filter methods. This was achieved utilising the novel EPS assisted F-measure, a further contribution, which newly allows for accurate objective analysis of 3-D surface detection with large 3-D data sets.

This evaluation is the first appraisal of statistical surface detection which has been undertaken in the literature and and is one of the primary contributions of the thesis. The evaluation provides recommendations for the following Vector Magnitude statistical tests:

- Difference of Boxes
- Fisher test
- Log Likelihood ratio test
- Student's $t$-test
- $\chi^2$ test
- Kolmogorov-Smirnov test
- Mann-Whitney $u$-test
- Robust rank order test

To support the quantitative findings, the surface detection methods were also visually assessed to identify whether visual characteristics found in the results correspond with the objective metric scores. Evaluation 1 was constructed to establish the resolving power of the different methods on interfaces with a strong intensity component (IntFl), a strong variance component to simulate stochastic texture (TexFl), and two interfaces which combine intensity and texture components at different amounts (Com1Fl and ComFl2). In 2-D data Williams et al. (2014) showed that the gradient based Canny method was effective at detecting the edges corresponding to intensity based interfaces, but was ineffective on texture based edges. For 3-D data the same trends continue,

with the Canny and Steerable methods performing strongly on interfaces with a strong intensity component, but poor on interfaces with a texture component. Evaluation 1 also revealed that the statistical tests excluding the $F$ test, were able to accurately detect both the intensity based interfaces and interfaces with a textured component. For the intensity based interface, the statistical $DoB$, $t$-test, $\chi^2$, $KS$, $u$-test and $RRO$ methods were all able to match the Canny method of achieving an optimal result. While the $KS$, $L$ and $F$ test methods were the only methods able to achieve optimal results for the texture interface, with the $\chi^2$-test achieving close to optimal detection. On the Combinational interfaces all methods with the exception of the $F$ test method outperformed the Steerable method, and all statistical methods with the exception of the $F$ method either offered improved detection or parity with the Canny method. Clearly indicating that statistical methods are more reliable at detecting surface interfaces over baseline methods, when texture is a factor in the properties of a region.

Evaluation 2 was constructed to identify the impact of interface topology on the detection. With 2-D statistical edge detection, Williams et al. (2014) identified that the non-parametric statistical tests were more effective with a larger scale parameter. However, a larger scale parameter introduces uncertainty in the location of a detected edge. The topology of an interface can contain details that may be lost due to the location uncertainty that arises at larger scales, which limits the effective range of the scale parameter of a detection method. Since a 3-D local neighbourhood possesses $N^3$ voxels compared to $N^2$ pixels for a 2-D neighbourhood, the number of elements processed in 3-D is greater than 2-D with an equal scale parameter. This allows for non parametric methods in 3-D to resolve surfaces with a relatively high signal to noise ratio when compared against 2-D. The same observation was made by Monga et al. (1991) in comparing 2-D and 3-D operators. Using a variety of 3-D interface topologies the results from Chapter 5 clearly show that for surface detection, the non-parametric tests are effective at smaller scales and are comparable to the parametric tests. This indicates that the earlier findings of Monga et al. (1991) hold true for statistical surface detection. There are some topological considerations when selecting a statistical test, for instance, the $t$-test was less effective on regular corners, compared to the other statistical methods. This is due to the test providing a high magnitude output when variance within a neighbourhood region is low. However, when positioned on a corner interface, the variance in one of the regions will be higher, and this results in a lower magnitude output. Consequentially corners are more weakly defined. However, when compared against baseline Canny and Steerable methods, the results replicate the general trends discovered in Evaluation 1, with region properties being the most important factor in resolving a surface interface.

Real images seldom contain 2 distinct regions, it is more typical for a real image to be complex, containing several distinct regions which may be different from one another by varying degrees of difference in the region properties. When surface detection is performed on an image of this kind, the relative magnitudes of the detected surface interfaces will be different, producing a surface map with surfaces of different strengths.

Different relative surface strengths are an important consideration when applying post processing techniques such as thresholding, or when the results are used by higher level processes such as region growing. This is because weaker surfaces can be removed by thresholding, or they may have a weaker response within a higher level process. Evaluation 3 introduced two synthetic image types which contained multiple regions, and this evaluation was used to identify the ability of a method to resolve multiple interface types within the same image. The most effective methods on these images were the non-parametric tests of the $\chi^2$, $KS$, $u$-test and $RRO$ test methods. This is due to the ranking nature of the tests which compare sorted data, instead of absolute values, which effectively creates normalisation at a scale equal to that of the local 3-D neighbourhood. As a result the non parametric tests are able to resolve boundaries considered "weaker", but at the expense of being more sensitive thus presenting with more noise. Evaluation 3 clearly shows that statistical methods outperform the baseline techniques on this image volume type.

### 8.1.5 Qualitative Analysis of Surface Detection with Multi-Modal MRI Data

Following the quantitative analysis in Chapter 5, the recommendations for the data can now be viewed in a qualitative analysis on multi-modal MRI data. Chapter 6 provided a real world application case for the evaluation of statistical surface detection which assessed whether the trends found with synthetic data in the quantitative analysis presented in Chapter 5 were also applicable to real imagery. It was found that many filter characteristics observed with the synthetic data also occurred with the different MRI modalities. Aspects such as sensitivity, noise suppression and artefacts such as duplicated and displaced surfaces were common to both.

The datasets consisted of real magnetic resonance image volumes from paediatric patients with Pilocytic Astrocytoma tumours. Each set of image data were comprised of 3 different MRI image modalities, T1-weighted, T2-weighted, and T1-weighted with an additional contrast agent. The statistical and baseline methods were assessed visually on their ability to resolve 3 different interface types within the images. The interface between the brain and tumour, the interface between the tumour and cysts, and the interface between the brain and the cysts.

The trends found in Chapter 5 evaluations 1,2 and 3 can also be found within the real imagery. T1-weighted images have a weak intensity component on interfaces which exist between the tumour and brain matter. Evaluation 1 showed that the baseline methods require a strong intensity component to delineate the surface, and when applied to T1-weighted images, the baseline methods were largely unable to detect the tumour interface, with statistical methods producing superior results. Most notably the Student's $t$-test which produced high magnitude connected surfaces on this type of image modality.

T2-weighted images possess a strong intensity component on the interface between tumour and brain, and the interface between cyst and brain, and a weak intensity component on the interface between tumour and cyst. The baseline methods which are better suited towards detecting a strong intensity differential were effective at resolving the tumour-brain and cyst-brain interfaces, but not the cyst-tumour boundaries. However, because the resolution of the real data was not isotropic, the $z$-spacing was greater than that of the $x$ and $y$ voxel spacing, this introduces a bias toward the detection of surface interfaces existing across the $z$-space. In the results this is apparent due to the large shaded areas, particularly visible on the cyst portion of the images. Whereas the Vector magnitude architecture was able to compensate for this bias, thus being able to detect the surface interfaces more clearly. The $\chi^2$ method was the most effective at resolving the 3 interface types of interest, while the parametric tests outperformed the baseline methods predominantly due to the bias compensation. The non parametric tests were effective at resolving all 3 interface types of interest, however this was at the cost of added noise, due to the sensitivity of the methods towards "weaker" surfaces, reflecting the results found in evaluation 3.

The T1-contrast enhanced modality provides a distinct intensity component for the each 3 interface types of interest. However, despite the baseline methods being optimised for intensity interfaces, they were largely outperformed by statistical methods on the T1-contrast enhanced modality. The parametric $DoB$ produced outputs with similar characteristics to the baseline methods. However, due to Vector Magnitude architecture allowing to offset the bias introduced by anisotropic resolution data, the $DoB$ method produces a preferential surface map response. While the $t$-test produced superior connectivity and magnitude in the interface response. The $\chi^2$-test offers potentially the best characteristics for T1-contrast data, due to producing strong connected boundaries with low noise outputs.

In Chapter 7, trends between the synthetic results and MRI results are discussed, highlighting the generalisability and transferability of the results. The key trends and characteristics were extracted from the objective and qualitative analyses and presented

in summary tables which contribute a useful set of guidelines for the expected trends in statistical and traditional surface detection. The summary tables strongly indicate the generalisability of the synthetic results to correspond with the MRI real case, with many of the characteristics found within the synthetic data results also being present in the MRI results. These tables therefore provide insight to the types of statistical test which are most effective for a particular type of interface, but also what the characteristics of the result would likely be if used if applied in a new domain, such as the amount of noise suppression, or the connectivity of the surfaces. These tables can therefore be used as a guideline to understand how to apply the statistical tests in different domains in the future.

## 8.2 Summary

Presented in this thesis were several novel statistics based techniques for surface detection which are well suited to resolving complex interfaces in 3-D data. These techniques were evaluated both objectively and qualitatively against two existing baseline methods (3-D Canny and Steerable filters) and is the first evaluation of statistical surface detection to have been undertaken. The objective evaluation utilised a 3 stage evaluation to analyse different aspects of the various filtering approaches. Each evaluation stage required the development of novel 3-D synthetic datasets which satisfied specific criteria aimed at addressing the issue of correlation between results obtained from that of both synthetic and real data. The factors which were analysed were the resolving power of the filters, the effect of interface topology on the filter response, and impact of multiple types of interface present within an image. Accompanying the objective evaluation was the need for an efficient and accurate performance metric for 3-D data, the most appropriate metric for surface detection is an F-measure score. However, for proper implementation of the metric, one to one correspondence matching between the filter response and reference image was required, this led to the development of an efficient algorithm for accurate 3-D correspondence matching called the EPS (Efficient Pairing Strategy) method. The statistical methods were also assessed qualitatively using imaging data from paediatric patients with Pilocytic Astrocytoma tumours. This case study observed the characteristics of the filter responses applied to individual T1-weighted, T2-weighted and contrast enhanced T1-weighted magnetic resonance imaging modalities, specifically examining the interfaces in the tumour and cyst regions of the brain image. The results of the evaluation strongly indicate that statistical approaches offer a number of benefits over the baseline methods, both in synthetic and real data. Notably, statistical methods outperform traditional methods in high intensity variance regions in synthetic data, and this translates to improved detection and less noisy results obtained

with both T1-weighted and contrast enhanced T1-weighted MRI modalities. Finally the statistical methods were also shown to be better suited to minimising distortion introduced by anisotropic resolution, a factor which is common in real 3-D images. The results suggest that statistical methods of surface detection should be adopted in place of traditional methods, if the type of surface is known to be defined by features in addition to or in place of intensity.

## 8.3 Future Work

While this work has presented a comprehensive evaluation of 3-D surface detection and compared the responses, trends and findings to work in 2-D edge detection, no full formal objective analysis on real data has been presented. This is largely due to the complexity of generating suitable and valid 2-D and 3-D ground truth data which is representative of the 3-D information required. Future, work should explore this and aim to provide a framework for acquisition of reliable real ground truth data for supporting objective comparisons of 2-D and 3-D operators.

At present, contouring tasks are supported in software with 2-D processes, and the potential benefits of 3-D detection methods, such as greater noise suppression, greater resolving power, and detection of surfaces between layers are yet to be realised. While an indicative application case has been presented in this thesis, the detection of brain tumour boundaries in MRI data, the understanding of the influence this 3-D surface detection approach can have has not yet been fully explored. While positive feedback was gained from the clinicians during discussions, a formal evaluation of how well this can improve contouring, by offering reliable detection of surfaces in T1, T2 and T1 contrast enhanced MRI data needs to be evaluated further.

One of the findings of this work was the impact of anisotropic image resolution on the quality of the results. The $z$-spacing is often inferior to the $x$ and $x$ resolution in real data and while this asymmetrical imbalance has no impact on the performance of 2-D operators, it can have a significant negative impact on the results for 3-D operators. This obstacle may be the reasoning behind why the wide scale incorporation of true 3-D operations has yet to be adopted into image analysis and visualisation software packages. Therefore, a future study which aims to quantify the effects of the bias, and the degree to which the Vector Magnitude method is able to mitigate it would be a useful contribution.

One of the most challenging aspects of fully utilising the potential of the statistical surface detection method are the considerations that need undertaking when selecting parameters for scale and the type of statistical test being applied. These parameters are

powerful with respect to their impact on the result. This thesis provides guidelines for optimising results, noting the most effective statistical tests on synthetically produced data, and across paediatric MRI modalities. However, selection of parameters are currently performed manually using human input and predicting the most effective scale and the most effective test to use requires an understanding of trade-off associated with scale selection and a detailed knowledge of the characteristics of each of the statistical filters. Future work should be focused toward streamlining this process into an automated or semi automated process whereby the parameters are optimised regardless of input data, the primary field of study for this would be a machine learning approach, with the end goal of a surface detection method suitable for all types of 3-D data. Additionally other statistical comparison tests could be evaluated.

Finally the benefits of statistical surface detection should be explored in alternate domains in order to determine how generalisable statistical surface detection would be to the follow-on high level processes, such as region growing that require surface information in 3-D. Understanding the impact that improved surfaces could have on the accuracy and reliability of the follow on processes would be a valuable contribution and would inform how to best utilise the method.

# Bibliography


Abdel-Gawad, A. H., L. A. Said and A. G. Radwan (2020). "Optimized Edge Detection Technique for Brain Tumor Detection in MR Images". In: *IEEE Access* 8, pp. 136243–136259.

Abdou, I. E. and W. K. Pratt (1979). "Quantitative design and evaluation of enhancement/thresholding edge detectors". In: *Proceedings of the IEEE* 67.5, pp. 753–763.

Aguet, F., M. Jacob and M. Unser (2005). "Three-dimensional feature detection using optimal steerable filters". In: *IEEE International Conference on Image Processing 2005*. Vol. 2, pp. II–1158.

Ali, M. and D. Clausi (2001). "Using the Canny edge detector for feature extraction and enhancement of remote sensing images". In: *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217)*. Vol. 5, 2298–2300 vol.5.

Bähnisch, C., P. Stelldinger and U. Köthe (2009). "Fast and Accurate 3D Edge Detection for Surface Reconstruction". In: *Pattern Recognition*. Springer Berlin Heidelberg, pp. 111–120.

Bauer, S., R. Wiest, L.-P. Nolte and M. Reyes (2013). "A survey of MRI-based medical image analysis for brain tumor studies". In: *Physics in Medicine and Biology* 58.13, R97–R129.

Beauchemin, M., K. Thomson and G. Edwards (1998). "On nonparametric edge detection in multilook SAR images". In: *IEEE Transactions on Geoscience and Remote Sensing* 36.5, pp. 1826–1829.

Bergholm, F. (1987). "Edge Focusing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9.6, pp. 726–741.

Bertels, J. et al. (2019). "Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice". In: *Medical image computing and computer assisted intervention*. Cham, Switzerland: Springer.

Bhattacharya, P. and D. Wild (1996). "A new edge detector for gray volumetric data". In: *Computers in Biology and Medicine* 26.4, pp. 315 –328.

Biederman, I. (1987). "Recognition-by-components: A theory of human image understanding." In: *Psychological Review* 94.2, 115–147.

Bomans, M., K. Hohne, U. Tiede and M. Riemer (1990). "3-D segmentation of MR images of the head for 3-D display". In: *IEEE Transactions on Medical Imaging* 9.2, pp. 177–183.

Borra, S. and S. Sarkar (1997). "A framework for performance characterization of intermediate-level grouping modules". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.11, pp. 1306–1312.

Bovik, A. (1987). "Streaking in median filtered images". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.4, pp. 493–503.

Bovik, A. and D. Munson (1985). "Boundary detection in speckle images". In: *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 10, pp. 893–896.

Bovik, A. C., T. S. Huang and D. C. Munson (1986). "Nonparametric tests for edge detection in noise". In: *Pattern Recognition* 19.3, pp. 209–219.

Bowring, N., E. Guest, P. Twigg, I. Williams and D. Gadsby (2004). "A new statistical method for edge detection on textured and cluttered images". In: *4th International Conference of Visualization, Imaging, and Image Processing,* pp. 435–440.

Bowyer, K., C. Kranenburg and S. Dougherty (2001). "Edge Detector Evaluation Using Empirical ROC Curves". In: *Computer Vision and Image Understanding* 84.1, pp. 77–103.

Braude, I. (2005). "Smooth 3D surface reconstruction from contours of biological data with MPU implicits". MA thesis. Philadelphia, PA: Drexel University, pp. 1–82.

Brejl, M. and M. Sonka (2000). "Directional 3D Edge Detection in Anisotropic Data: Detector Design and Performance Assessment". In: *Computer Vision and Image Understanding* 77.2, pp. 84–110.

Canny, J. (1986). "A Computational Approach to Edge Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6, pp. 679–698.

Cardoso, J. S. and L. Corte-Real (2005). "Toward a generic evaluation of image segmentation". In: *IEEE Transactions on Image Processing* 14.11, pp. 1773–1782.

Chang, S. H., Leiguang Gong, Maoqing Li, Xiaoying Hu and Jingwen Yan (2008). "Small retinal vessel extraction using modified Canny edge detection". In: *2008 International Conference on Audio, Language and Image Processing*, pp. 1255–1259.

Chen, Z. and H. Zhu (2019). "Visual Quality Evaluation for Semantic Segmentation: Subjective Assessment Database and Objective Assessment Measure". In: *IEEE Transactions on Image Processing* 28.12, pp. 5785–5796.

Cho, K., P. Meer and J. Cabrera (1995). "Quantitative evaluation of performance through bootstrapping: edge detection". In: *Proceedings of International Symposium on Computer Vision - ISCV*, pp. 491–496.

Clarke, L. and R. Velthuizen (1995). "MRI segmentation: methods and applications". In: *Magnetic Resonance Imaging* 13.3, pp. 343–368.

Coggins, J. M. (1983). “A Framework for Texture Analysis Based on Spatial Filtering”. PhD thesis. East Lansing, MI, USA.

Collins, D. L., A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes and A. C. Evans (1998). “Design and construction of a realistic digital brain phantom”. In: *IEEE Transactions on Medical Imaging* 17.3, pp. 463–468.

Cross, G. R. and A. K. Jain (1983). “Markov Random Field Texture Models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5.1, pp. 25–39.

Cyganski, D., B. King, R. F. Vaz and J. A. Orr (1995). “ROC analysis of ATR from SAR images using a model-based recognizer incorporating pose information”. In: *Algorithms for Synthetic Aperture Radar Imagery II*. Vol. 2487. SPIE.

De Souza, P. (1983). “Edge detection using sliding statistical tests”. In: *Computer Vision, Graphics, and Image Processing* 23.1, pp. 1–14.

DeAngelis, L. M. (2001). “Brain Tumors”. In: *New England Journal of Medicine* 344.2. PMID: 11150363, pp. 114–123. eprint: `https://doi.org/10.1056/NEJM200101113440207`.

Despotović, I., B. Goossens and W. Philips (2015). “MRI Segmentation of the Human Brain: Challenges, Methods, and Applications”. In: *Computational and Mathematical Methods in Medicine* 2015, pp. 1–23.

Dewey, M., T. Schink and C. F. Dewey (2007). “Claustrophobia during magnetic resonance imaging: Cohort study in over 55,000 patients”. In: *Journal of Magnetic Resonance Imaging* 26.5, pp. 1322–1327. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.21147`.

Ding, L. and A. Goshtasby (2001). “On the Canny edge detector”. In: *Pattern Recognition* 34.3, pp. 721 –725.

Du, J., Z. He, L. Wang, A. Gholipour, Z. Zhou, D. Chen and Y. Jia (2020). “Super-resolution reconstruction of single anisotropic 3D MR images using residual convolutional neural network”. In: *Neurocomputing* 392, pp. 209–220.

*FADGI* (2018). Digital image - Glossary - Federal Agencies Digitization Guidelines Initiative. URL: `http://www.digitizationguidelines.gov/term.php?term=digitalimage`.

Feng Ge, Song Wang and Tiecheng Liu (2006). "Image-Segmentation Evaluation From the Perspective of Salient Object Extraction". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 1, pp. 1146–1153.

Fernandez, M. A., R. M. Lopes and N. S. T. Hirata (2015). "Image Segmentation Assessment from the Perspective of a Higher Level Task". In: *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 111–118.

Fesharaki, M. N. and G. R. Hellestrand (1994). "A new edge detection algorithm based on a statistical approach". In: *Proceedings of ICSIPNN '94. International Conference on Speech, Image Processing and Neural Networks*. Vol. 1, pp. 21–24.

Fetit, A. E. (2015). "Radiomics in paediatric neuro-oncology : MRI textural features as diagnostic and prognostic biomarkers". PhD thesis. University of Warwick.

Fligner, M. A. and G. E. Policello (1981). "Robust Rank Procedures for the Behrens-Fisher Problem". In: *Journal of the American Statistical Association* 76.373, pp. 162–168.

Forbes, L. A. and B. A. Draper (2000). "Inconsistencies in edge detector evaluation". In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000*. Vol. 2, pp. 398–404.

Fram, J. and E. Deutsch (1975). "On the Quantitative Evaluation of Edge Detection Schemes and their Comparison with Human Performance". In: *IEEE Transactions on Computers* 24.06, pp. 616–628.

Freeman, W. T. and E. H. Adelson (1991). "The design and use of steerable filters". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.9, pp. 891–906.

Goldberg, A. V. and R. Kennedy (1995). "An efficient cost scaling algorithm for the assignment problem". In: *Mathematical Programming* 71.2, pp. 153–177.

Gruenwedel, S., P. Van Hese and W. Philips (2011). "An Edge-Based Approach for Robust Foreground Detection". In: *Advanced Concepts for Intelligent Vision Systems*. Springer Berlin Heidelberg, 554–565.

Guest, E. (1994). "Automatic Reconstruction from Serial Sections". PhD thesis. University of Edinburgh.

Haralick, R. M. (1979). "Statistical and structural approaches to texture". In: *Proceedings of the IEEE* 67.5, pp. 786–804.

Haralick, R. M. (1981). "Some Neighborhood Operators". In: *Real-Time Parallel Computing: Imaging Analysis*. Ed. by M. Onoe, K. Preston and A. Rosenfeld. Boston, MA: Springer US, pp. 11–35.

Haralick, R. (1994). "Propagating covariance in computer vision". In: *Proceedings of 12th International Conference on Pattern Recognition*. Vol. 1, pp. 493–498.

Heath, M., S. Sarkar, T. Sanocki and K. Bowyer (1996). "Comparison of edge detectors: a methodology and initial study". In: *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 143–148.

Hou, Z. and G. Wei (2002). "A new approach to edge detection". In: *Pattern Recognition* 35.7, pp. 1559–1570.

Huan, Z. and Y. Hou (2008). "An Segmentation Algorithm of Texture Image Based on DWT". In: *2008 Fourth International Conference on Natural Computation*. Vol. 5, pp. 433–436.

Jacob, M. and M. Unser (2004). "Design of steerable filters for feature detection using canny-like criteria." In: *IEEE transactions on pattern analysis and machine intelligence* 26.8, pp. 1007–19.

Jähne, B. (1997). *Digital Image Processing*. 6th. Heidelberg: Springer Berlin Heidelberg, p. 302.

Jähne, B., H. Scharr, S. Körkel, B. Jähne, H. Haußecker and P. Geißler (1999). "Principles of Filter Design". In: vol. 2. Academic Press, pp. 125–151.

Jain, R. and T. Binford (1991). "Ignorance, myopia, and naivete in computer vision systems". In: *CVGIP: Image Understanding* 53.1, pp. 112–117.

Jamil, N., H. Soh, T. Sembok and Z. Bakar (Jan. 2011). "A Modified Edge-Based Region Growing Segmentation of Geometric Objects." In: pp. 99–112.

Juneja, M. and P. S. Sandhu (2009). "Performance Evaluation of Edge Detection Techniques for Images in Spatial Domain". In: *International Journal of Computer Theory and Engineering*, pp. 614–621.

Kaur, P. and B. Kaur (2016). "2-D geometric shape recognition using canny edge detection technique". In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 3161–3164.

Kennedy, D. N., P. a. Filipek and V. R. Caviness (1989). "Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging." In: *IEEE transactions on medical imaging* 8.1, pp. 1–7.

Ker, J., L. Wang, J. Rao and T. Lim (2018). "Deep Learning Applications in Medical Image Analysis". In: *IEEE Access* 6, pp. 9375–9389.

Khotanzad, A. and J.-Y. Chen (1989). "Unsupervised segmentation of textured images by edge detection in multidimensional feature". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.

Kirsch, R. A. (1971). "Computer Determination Biological of the Constituent Images". In: *Computers and Biomedical Research 4* 328, pp. 315–328.

Kitchen, L. and A. Rosenfeld (1981). "Edge Evaluation using Local Edge Coherence". In: *IEEE Transactions on Systems, Man , And Cybernetics* SMC-11.9, pp. 597–605.

Kuhn, H. W. (1955). "The Hungarian Method for the assignment problem". In: *Naval Research Logistics Quarterly* 2, pp. 83–97.

Kundu, A. and S. K. Mitra (1987). "A New Algorithm for Image Edge Extraction Using a Statistical Classifier Approach". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9.4, pp. 569–577.

Lim, D. H. (2006). "Robust edge detection in noisy images". In: *Computational Statistics and Data Analysis* 50.3, pp. 803–812.

Liu, G. and R. M. Haralick (2000). "Assignment problem in edge detection performance evaluation". In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000.* Vol. 1, pp. 26–31.

Liu, G, W. Yang, S. Zhu and Q Huang (2013). "PET/CT image textures for the recognition of tumors and organs at risk for radiotherapy treatment planning". In: *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC).* IEEE, pp. 1–3.

Liu, G. and R. M. Haralick (2002). "Optimal matching problem in detection and recognition performance evaluation". In: *Pattern Recognition* 35.10, pp. 2125–2139.

Liu, H. and K. C. Jezek (2004). "Automated extraction of coastline from satellite imagery by integrating Canny edge detection and locally adaptive thresholding methods". In: *International Journal of Remote Sensing* 25.5, pp. 937–958. eprint: `https://doi.org/10.1080/0143116031000139890`.

Liu, H. K. (1977). "Two- and three-dimensional boundary detection". In: *Computer Graphics and Image Processing* 6.2, pp. 123–134.

Liu, X. and W. Chen (2006). "Elastic Registration Algorithm of Medical Images Based on Fuzzy Set". In: *Biomedical Image Registration.* Ed. by J. P. W. Pluim, B. Likar and F. A. Gerritsen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 214–221.

Liyuan Li and M. K. H. Leung (2002). "Integrating intensity and texture differences for robust change detection". In: *IEEE Transactions on Image Processing* 11.2, pp. 105–112.

Lopez-Molina, C., B. De Baets and H. Bustince (2013). "Quantitative error measures for edge detection". In: *Pattern Recognition* 46.4, pp. 1125–1139.

Lyra, M. (2012). "Single Photon Emission Tomography (SPECT) and 3D Images Evaluation in Nuclear Medicine". In: *Image Processing* December.

Maani, R., S. Kalra and Y. Yang (2014). "Robust Volumetric Texture Classification of Magnetic Resonance Images of the Brain Using Local Frequency Descriptor". In: *IEEE Transactions on Image Processing* 23.10, pp. 4625–4636.

Martin, A., H. Laanaya and A. Arnold-Bos (2006). "Evaluation for uncertain image classification and segmentation". In: *Pattern Recognition* 39.11, pp. 1987 –1995.

Martin, D. R. (2003). "An Empirical Approach to Grouping and Segmentation". PhD thesis. University of California.

Martin, D. R., C. C. Fowlkes and J. Malik (2004). "Learning to detect natural image boundaries using local brightness, color, and texture cues." In: *IEEE transactions on pattern analysis and machine intelligence* 26.5, pp. 530–49.

Meilă, M. (2005). "Comparing clusterings: an axiomatic view". In: *In ICML '05: Proceedings of the 22nd international conference on Machine learning*. ACM Press, pp. 577–584.

Meinhardt, E., E. Zacur, A. F. Frangi and V. Caselles (2008). "3D Edge Detection by Selection of Level Surface Patches". In: *Journal of Mathematical Imaging and Vision* 34.1, pp. 1–16.

Menze, H. et al. (2015). "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10, pp. 1993–2024.

Menze, H. et al. (2018). *Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*. eprint: `1811.02629`.

Milan, S., V. Hlavac and R. Boyle (1993). *Image processing, analysis, and machine vision*. Brooks/Cole.

Militello, C., S. Vitabile, L. Rundo, C. Gagliardo and S. Salerno (2015). “An edge-driven 3D region-growing approach for upper airway morphology and volume evaluation in patients with Pierre Robin sequence”. In: *International Journal of Adaptive and Innovative Systems* 2.3, pp. 232–253.

Mirmehdi, M., X. Xie and J. Suri (2008). *Handbook of Texture Analysis*. Imperial College Press.

Monga, O, R. Deriche and J. Rocchisani (1991). “3D edge detection using recursive filtering: application to scanner images”. In: *CVGIP: Image Understanding*, pp. 28–35.

Monga, O. and R. Deriche (1989). “3D Edge Detection Using Recursive Filtering”. In: *Computer Vision and Pattern Recognition*, pp. 28–35.

Morgenthaler, D. G. and A. Rosenfeld (1981). “Multidimensional Edge Detection by Hypersurface Fitting”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-3.4, pp. 482–486.

Morgenthaler, D. G. and A. Rosenfeld (1981). “Surfaces in three-dimensional digital images”. In: *Information and Control* 51.3, pp. 227 –247.

*MRI scanners in the UK 2000-2014 | Statistic* (n.d.). en. (Visited on 29/03/2019).

Munkres, J (1957). “Algorithms for the Assignment and Transportation Problems"”. In: *Journal of the Society for Industrial and Applied Mathematics* 5, pp. 32–38.

Murphy, K. J. and J. A. Brunberg (1997). “Adult claustrophobia, anxiety and sedation in MRI”. In: *Magnetic Resonance Imaging* 15.1, pp. 51 –54.

Nercessian, S., K. Panetta and S. Agaian (2009). “A non-reference measure for objective edge map evaluation”. In: October, pp. 4563–4568.

Nevatia, R. and K. R. Babu (1980). “Linear feature extraction and description”. In: *Computer Graphics and Image Processing* 13.3, pp. 257 –269.

Nikolic, M., E. Tuba and M. Tuba (2016). "Edge detection in medical ultrasound images using adjusted Canny edge detection algorithm". In: *2016 24th Telecommunications Forum (TELFOR)*, pp. 1–4.

Nixon, M. S. and A. S. Aguado (2002). *Feature Extraction and Image Processing*. Newnes, p. 360.

Nixon, M. S. and A. S. Aguado (2012). "Chapter 4 - Low-level feature extraction (including edge detection)". In: *Feature Extraction and Image Processing for Computer Vision (Third Edition)*. Ed. by M. S. Nixon and A. S. Aguado. Third Edition. Oxford: Academic Press, pp. 137 –216.

Otsu, N. (1979). "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1, pp. 62–66.

Papari, G. and N. Petkov (2011). "Edge and line oriented contour detection: State of the art". In: *Image and Vision Computing* 29.2-3, pp. 79–103.

Pavlidis, T. and Y.-T. Liow (1990). "Integrating region growing and edge detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.3, pp. 225–233.

Peli, T and D Malah (1982). "A study on Edge Detection Algorithms". In: *Computer Graphics and Image Processing* 1-21.

Petrou, M. and P. García Sevilla (2006a). *Image Processing: Dealing with Texture*, pp. 1–618.

Petrou, M. and P. García Sevilla (2006b). "Introduction: What is texture". In: *Image Processing: Dealing with Texture*, p. 1.

Petrou, M. and P. García Sevilla (2006c). "Non-stationary grey texture images". In: *Image Processing: Dealing with Texture*, p. 297.

Petrou, M., V. a. Kovalev and J. R. Reichenbach (2006). "Three-dimensional nonlinear invisible boundary detection." In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 15.10, pp. 3020–32.

Pont-Tuset, J. and F. Marques (2016). "Supervised Evaluation of Image Segmentation and Object Proposal Techniques". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.7, pp. 1465–1478.

Prakoonwit, S. and R. Benjamin (2007). "3D surface point and wireframe reconstruction from multiview photographic images". In: *Image and Vision Computing* 25.9, pp. 1509–1518.

Prakoonwit, S. and R. Benjamin (2012). "3D surface reconstruction from multiview photographic images using 2D edge contours". In: *3D Research* 3.4, pp. 1–12.

Prewitt, J. (1970). "Object Enhancement and Extraction". In: *Picture Processing and Psychopictorics*, pp. 75 –149.

Prieto, M. and A. Allen (2003). "A similarity metric for edge images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.10, pp. 1265–1273.

Punarselvam, E. and P. Suresh (2011). "Edge Detection of CT scan Spine disc image using Canny Edge Detection Algorithm based on Magnitude and Edge Length". In: *3rd International Conference on Trendz in Information Sciences Computing (TISC2011)*, pp. 136–140.

Ramesh, V. and R. M. Haralick (1994). "An integrated gradient edge detector. Theory and performance evaluation". In: *ARPA Image Understanding Workshop, Monterey, CA*, pp. 689–702.

Ramya, R. and P. S. Babu (2015). "Automatic tuberculosis screening using canny Edge detection method". In: *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, pp. 282–285.

Reyes-Aldasoro, C. and a. Bhalerao (2006). "The Bhattacharyya space for feature selection and its application to texture segmentation". In: *Pattern Recognition* 39.5, pp. 812–826.

Roberts, L. (1963). "Machine perception of three-dimensional solids". PhD thesis. Massachusetts Institute of Technology, Dept. of Electrical Engineering.

Robinson, G. S. (1977). "Edge detection by compass gradient masks". In: *Computer Graphics and Image Processing* 6.5, pp. 492 –501.

Rong, W., Z. Li, W. Zhang and L. Sun (2014). "An improved Canny edge detection algorithm". In: *2014 IEEE International Conference on Mechatronics and Automation*, pp. 577–582.

Rosenfeld, A. and A. C. Kak (1982). "Digital Picture Processing: Introduction". In: *Digital Picture Processing*, p. 435.

Rothwell, C., J. Mundy, W. Hoffman and V.-D. Nguyen (1995). "Driving Vision by Topology". In: *Proceedings of International Symposium on Computer Vision (ISCV)*, pp. 395–400.

Savitzky, A. and M. J. E. Golay (1964). "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." In: *Analytical Chemistry* 36.8, pp. 1627–1639.

Scharr, H. (2000). "Optimal operators in digital image processing". PhD thesis. Ruprecht-Karls-Universität Heidelberg.

Selvakumar, P. and S. H. Ganesh (2017). "Tamil Character Recognition Using Canny Edge Detection Algorithm". In: *2017 World Congress on Computing and Communication Technologies (WCCCT)*, pp. 250–254.

Setayesh, M., M. Zhang and M. Johnston (2011). "Detection of Continuous, Smooth and Thin Edges in Noisy Images Using Constrained Particle Swarm Optimisation". In: *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*. GECCO '11. Dublin, Ireland: ACM, pp. 45–52.

Sharma, N. and L. Aggarwal (Apr. 2010). "Automated medical image segmentation techniques". In: *Journal of medical physics / Association of Medical Physicists of India* 35, pp. 3–14.

Simmons, J., J. Waggoner, Y. Zhou, S. Wang and M. De Graef (2013). "3D Materials Image Segmentation by 2D Propagation: A Graph-Cut Approach Considering Homomorphism". In: *IEEE Transactions on Image Processing* 22.12, pp. 5282–5293.

Smith, S. M. and J. M. Brady (1995). "SUSAN - A New Approach to Low Level Image Processing". In: *International Journal of Computer Vision* 23, pp. 45–78.

Smith, S. and I. Williams (2015). "A Statistical Method for Improved 3D Surface Detection". In: *IEEE Signal Processing Letters* 22.8, pp. 1045–1049.

Smith., S. and I. Williams. (2020). "Efficient One-to-One Pair Matching for 2-D and 3-D Edge Detection Evaluation". In: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,* INSTICC. SciTePress, pp. 590–598.

Sobel, I. and G Feldman (1968). "A 3 × 3 Isotropic Gradient Operator for Image Processing." In: *The Stanford Artificial Intelligence Project*, pp. 271–272.

Suzuki, K., I. Horiba and N. Sugie (2003). "Neural Edge Enhancer for Supervised Edge Enhancement from Noisy Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.12, pp. 1582–1596.

Synopsys (19th Jan. 2021). *Simpleware™*.

Tahmid, T. and E. Hossain (2017). "Density based smart traffic control system using canny edge detection algorithm for congregating traffic information". In: *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–5.

Tang, H, E. Wu, Q. Ma, D Gallagher, G. Perera and T Zhuang (2000). "MRI brain image segmentation by multi-resolution edge detection and region selection". In: *Computerized Medical Imaging and Graphics* 24.6, pp. 349 –357.

Thangam, S. V., K. S. Deepak, G. N. H. Rai and P. P. Mirajkar (2009). "An Effective Edge Detection Methodology for Medical Images Based on Texture Discrimination". In: pp. 227–231.

Todman, A and E. Claridge (1997). "Cell segmentation in histological images of striated muscle tissue: a perceptual grouping approach". In: *Proceedings of Medical Image Understanding and Analysis*, pp. 101–104.

Udupa, J. K., V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch and J. Woodburn (2006). "A framework for evaluating image segmentation algorithms". In: *Computerized Medical Imaging and Graphics* 30.2, pp. 75 –87.

van Vliet, L. J., I. T. Young and G. L. Beckers (1989). "A nonlinear laplace operator as edge detector in noisy images". In: *Computer Vision, Graphics, and Image Processing* 45.2, pp. 167 –195.

Wang, K, A. Sims and A Murray (2009). "A comparison of 2D and 3D edge detectors in semi automated measurements of chamber volumes using 3D echocardiographic laboratory phantom images". In: *Computers in Cardiology, 2009*, pp. 1–4.

Wang, Z, E Wang and Y. Zhu (2020). "Image segmentation evaluation: a survey of methods". In: *Artificial Intelligence 53*, 5637–5674.

Wang Xiao and Xue Hui (2010). "An improved canny edge detection algorithm based on predisposal method for image corrupted by gaussian noise". In: *2010 World Automation Congress*, pp. 113–116.

Wen, P. Y., D. R. Macdonald, D. A. Reardon, T. F. Cloughesy, A. G. Sorensen, E. Galanis, J. DeGroot, W. Wick, M. R. Gilbert, A. B. Lassman, C. Tsien, T. Mikkelsen, E. T. Wong, M. C. Chamberlain, R. Stupp, K. R. Lamborn, M. A. Vogelbaum, M. J. van den Bent and S. M. Chang (2010). "Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group". In: *Journal of Clinical Oncology* 28.11. PMID: 20231676, pp. 1963–1972.

Willemink, M. J., W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin and M. P. Lungren (2020). "Preparing Medical Imaging Data for Machine Learning". In: *Radiology* 295.1. PMID: 32068507, pp. 4–15. eprint: `https://doi.org/10.1148/radiol.2020192224`.

Williams, I. (2008). "Edge Detection of Textured Images using Multiple Scales and Statistics". PhD. Manchester Metropolitan University.

Williams, I., N. Bowring and D. Svoboda (2014). “A performance evaluation of statistical tests for edge detection in textured images”. In: *Computer Vision and Image Understanding* 122, pp. 115 –130.

Witkin, A. (1984). “Scale-space filtering: A new approach to multi-scale description”. In: *ICASSP ’84. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 9, pp. 150–153.

Yi, S., R. M. Haralick and L. G. Shapiro (1994). “Error propagation in machine vision”. In: *Machine Vision and Applications* 7.2, pp. 93–114.

Zhang, H., J. E. Fritts and S. A. Goldman (2008). “Image segmentation evaluation: A survey of unsupervised methods”. In: *Computer Vision and Image Understanding* 110.2, pp. 260 –280.

Zhu, Q. (1996). “Efficient evaluations of edge connectivity and width uniformity”. In: *Image and Vision Computing* 14.1, pp. 21–34.

Ziou, D. and R. Mohr (1992). “An experience on automatic selection of the edge detectors”. In: *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis,* pp. 586–589.

Zucker, S. W. and R. a. Hummel (1981). “A Three-Dimensional Edge Operator”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-3.3, pp. 324–331.

# Appendix A

# T1 - Weighted Dataset

The following Appendices show the 5 paediatric MRI datasets, grouped by modality, T1-weighted, T2-weighted and contrast enhanced T1-weighted images.

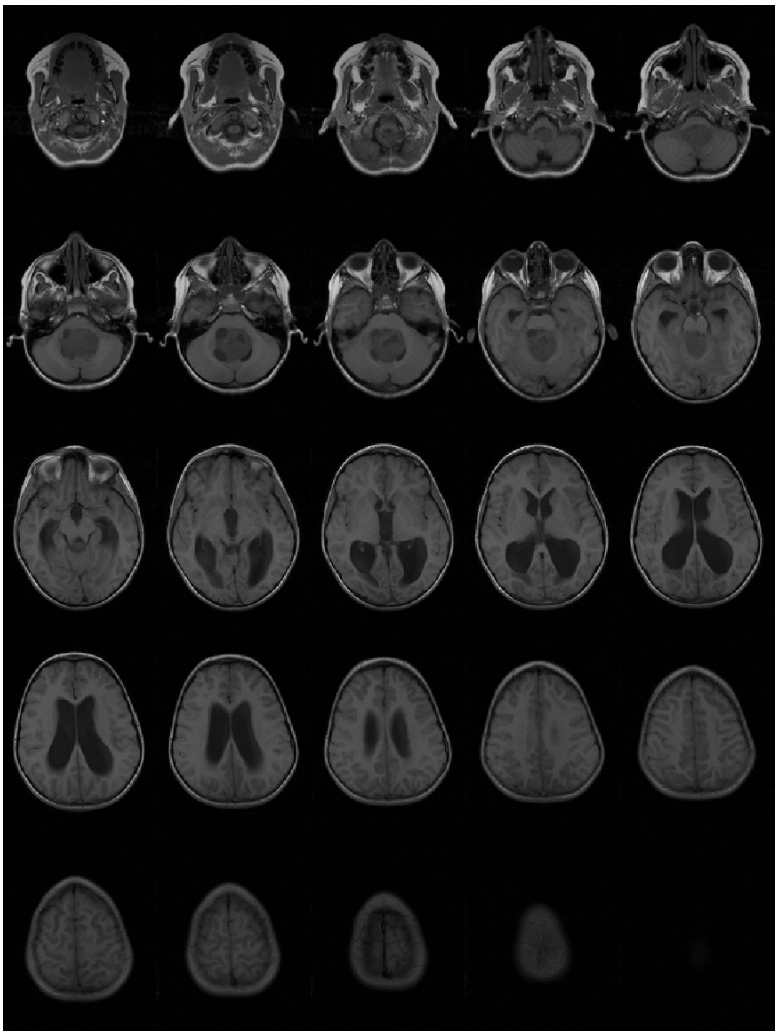


Figure A.1: Set 1, T1-weighted

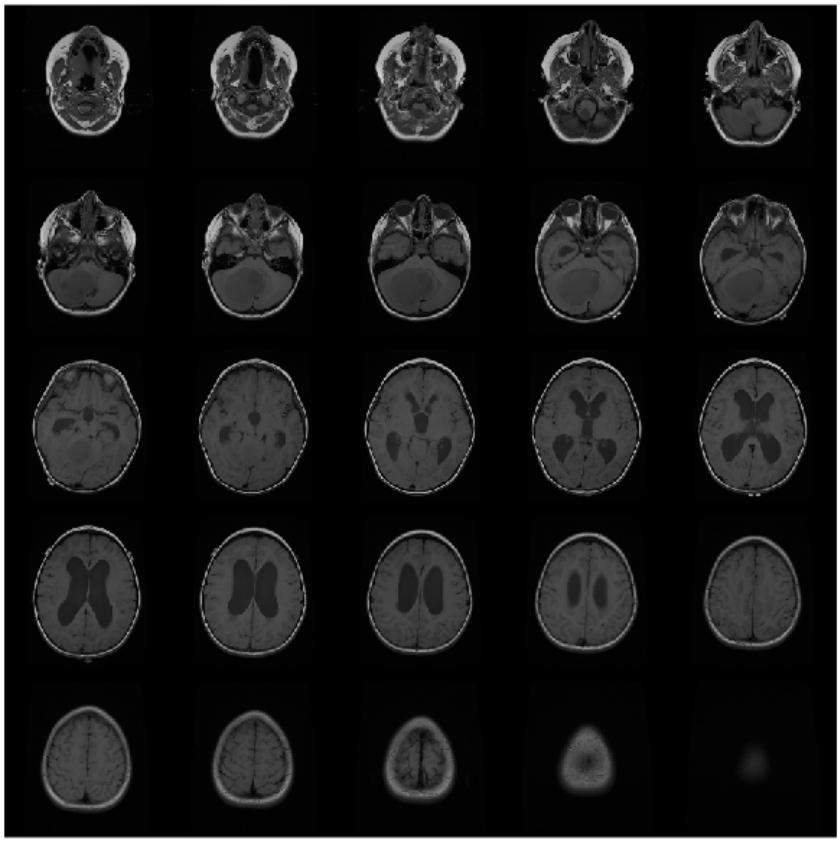


Figure A.2: Set 2, T1-weighted

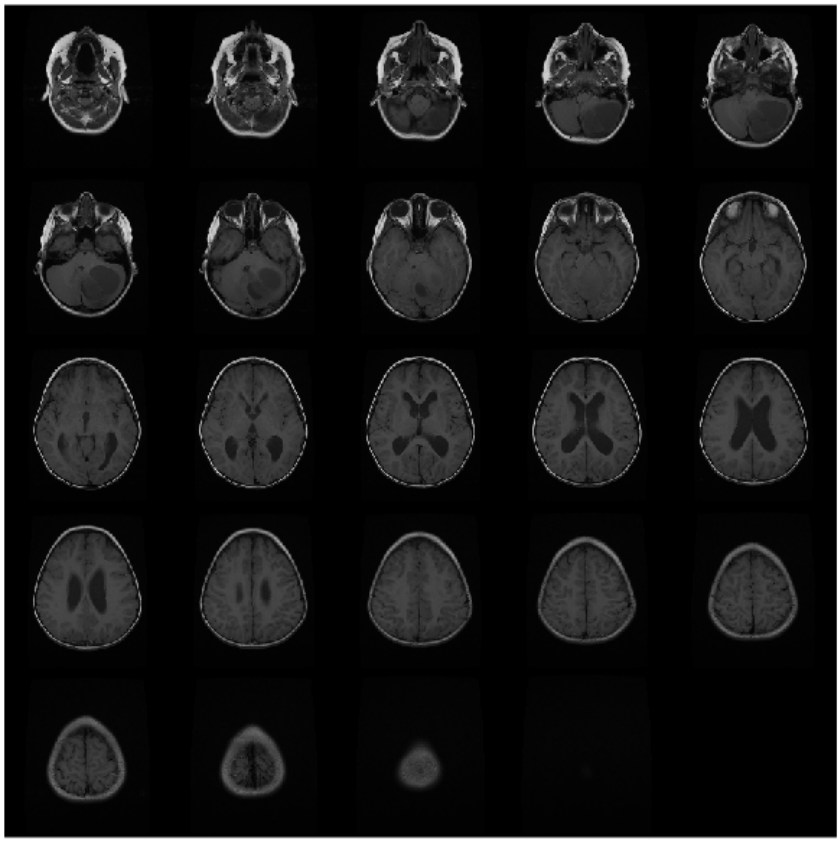


Figure A.3: Set 3, T1-weighted

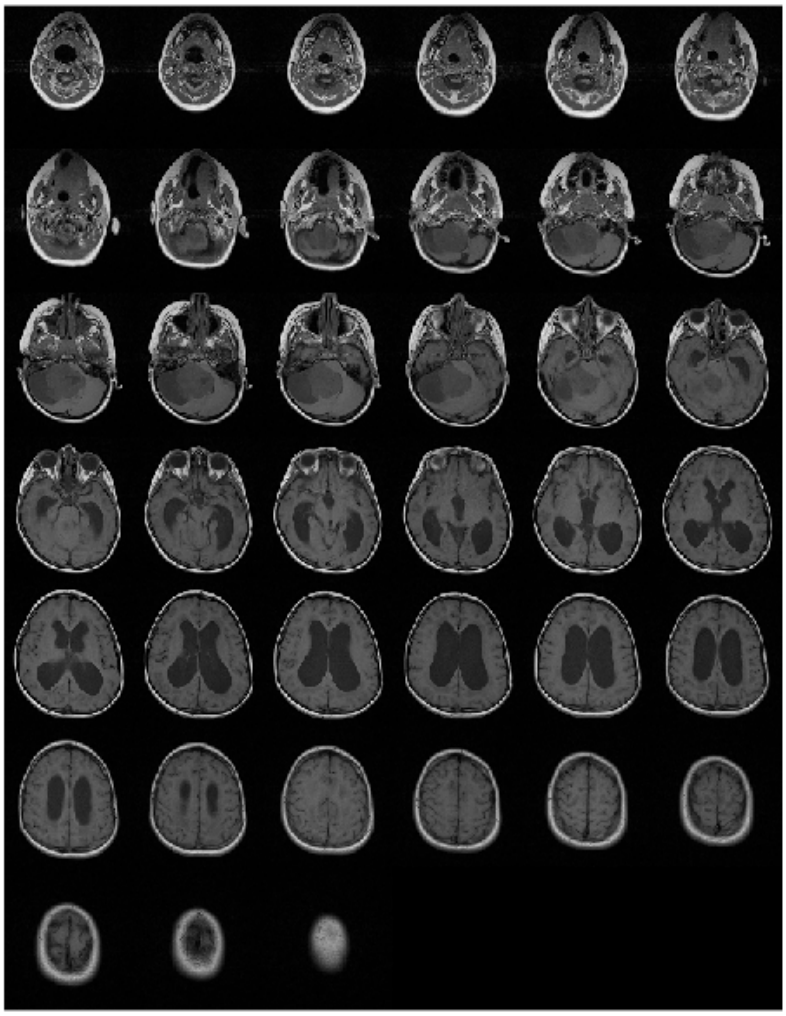


Figure A.4: Set 4, T1-weighted

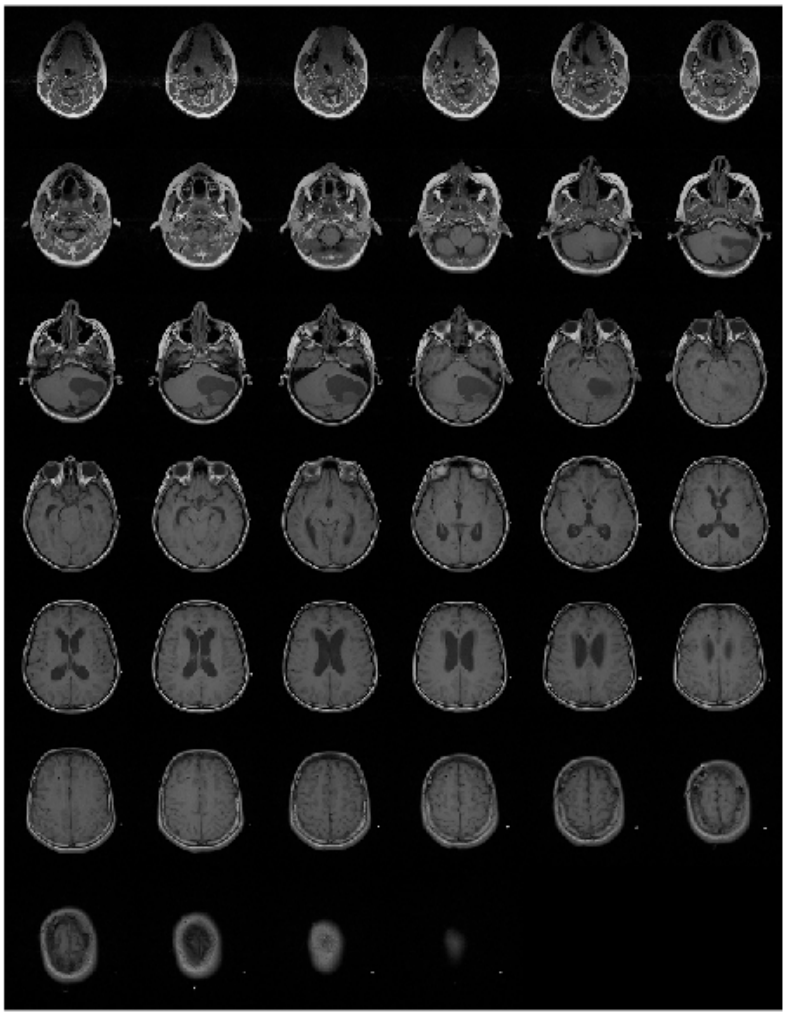
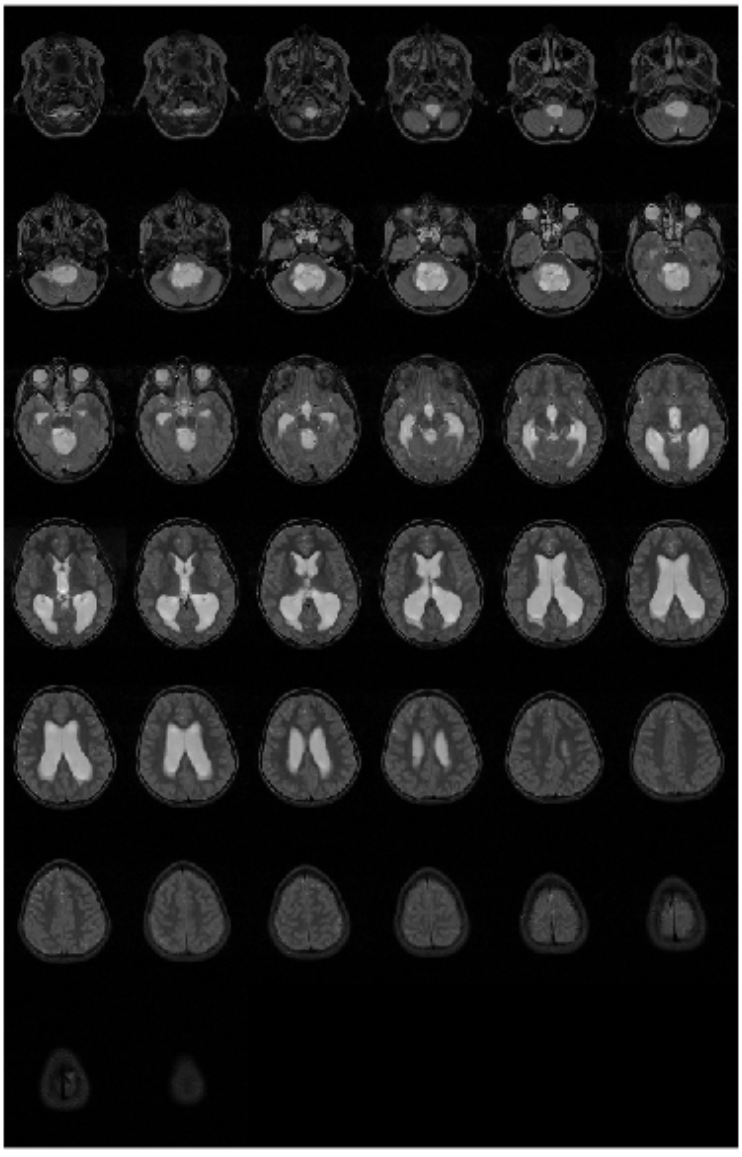


Figure A.5: Set 5, T1-weighted

# Appendix B

# T2 - Weighted Dataset



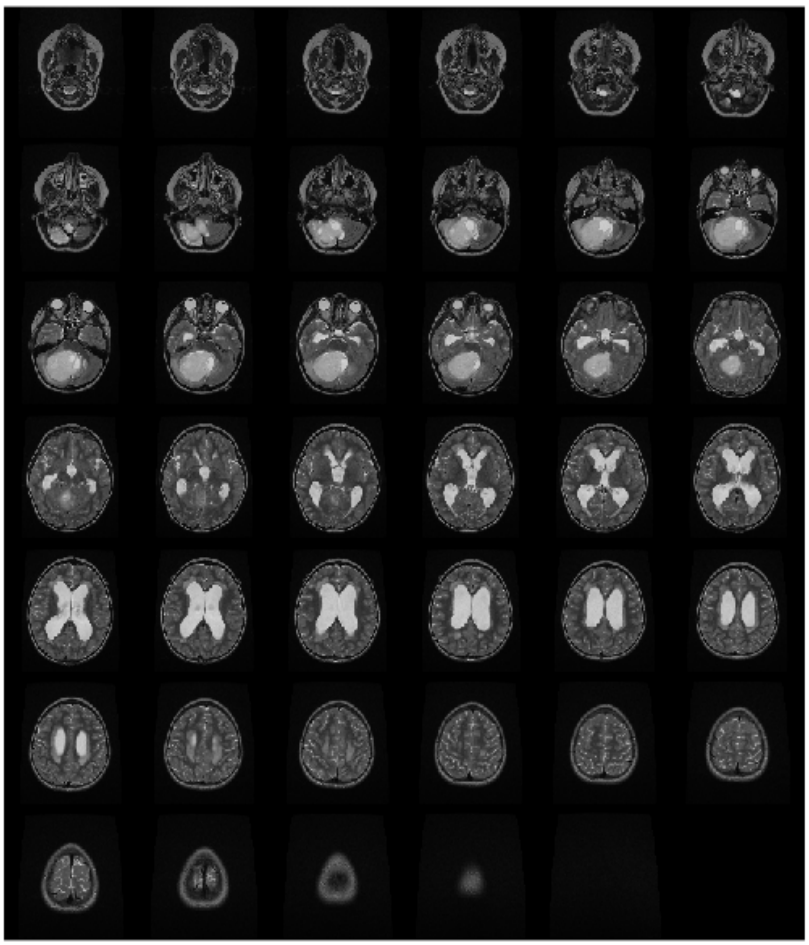


Figure B.2: Set 2, T2-weighted

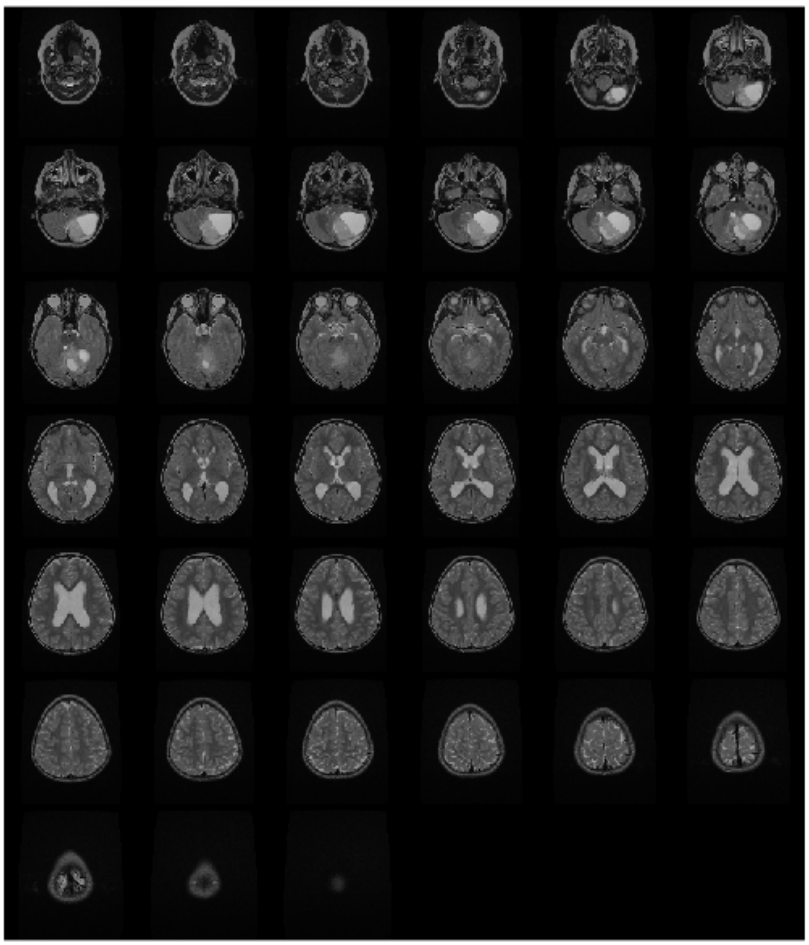


Figure B.3: Set 3, T2-weighted

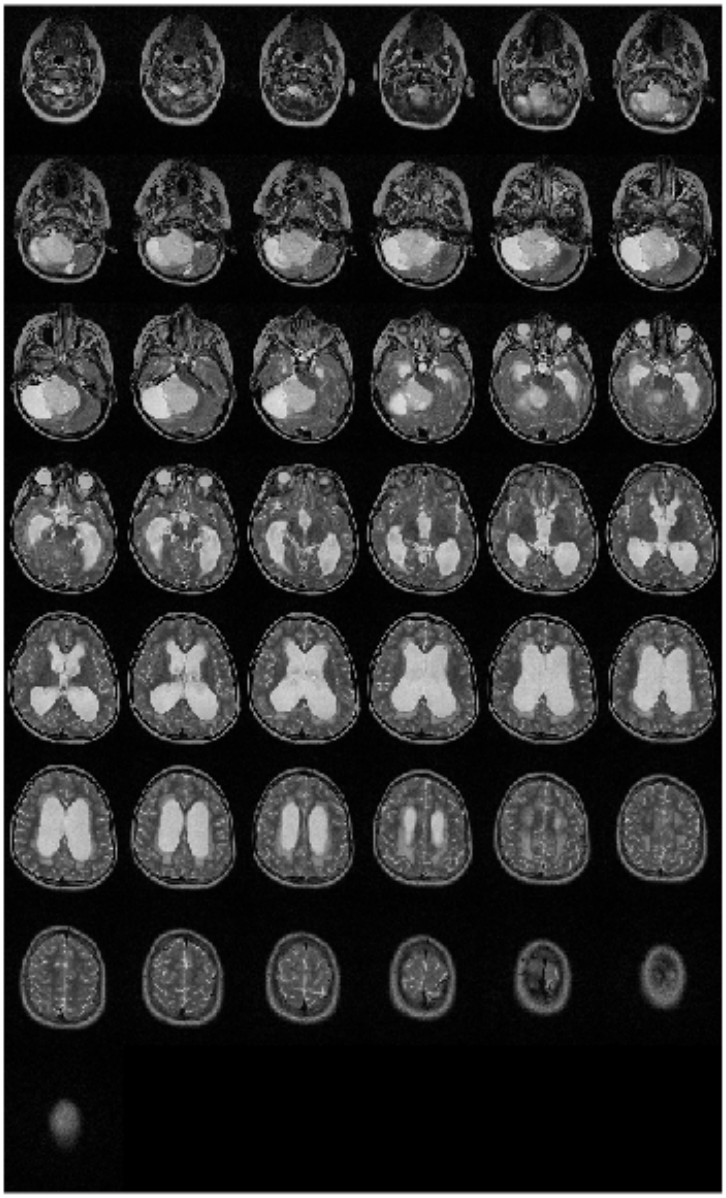


Figure B.4: Set 4, T2-weighted

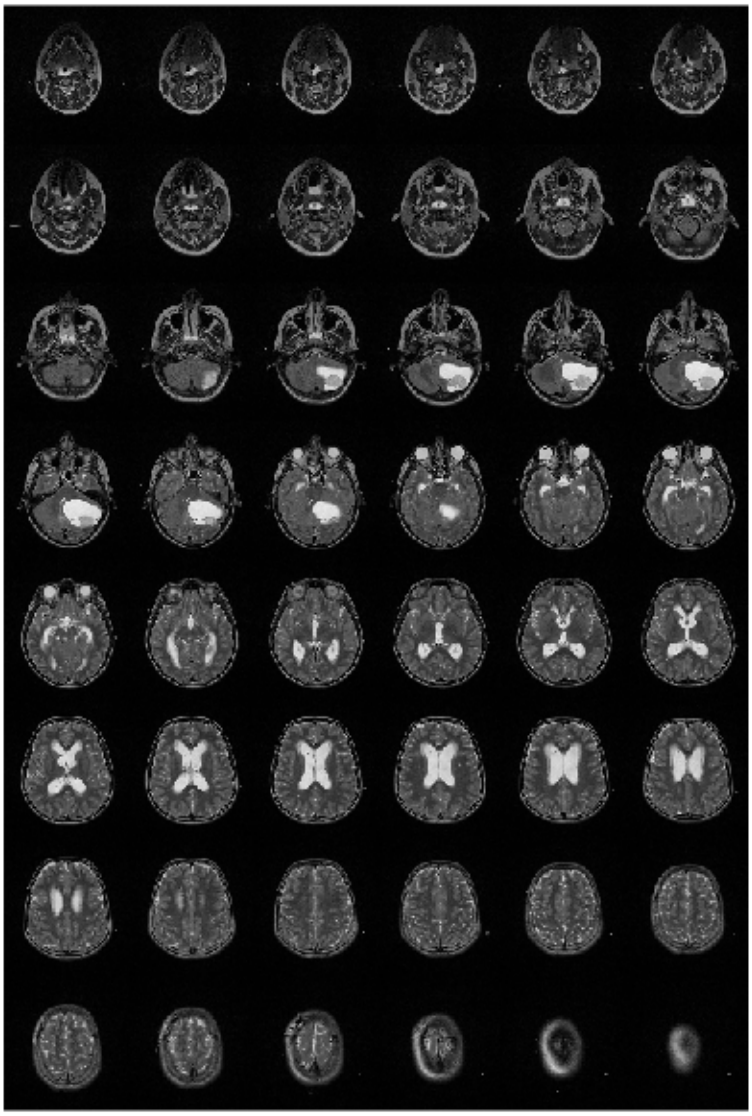
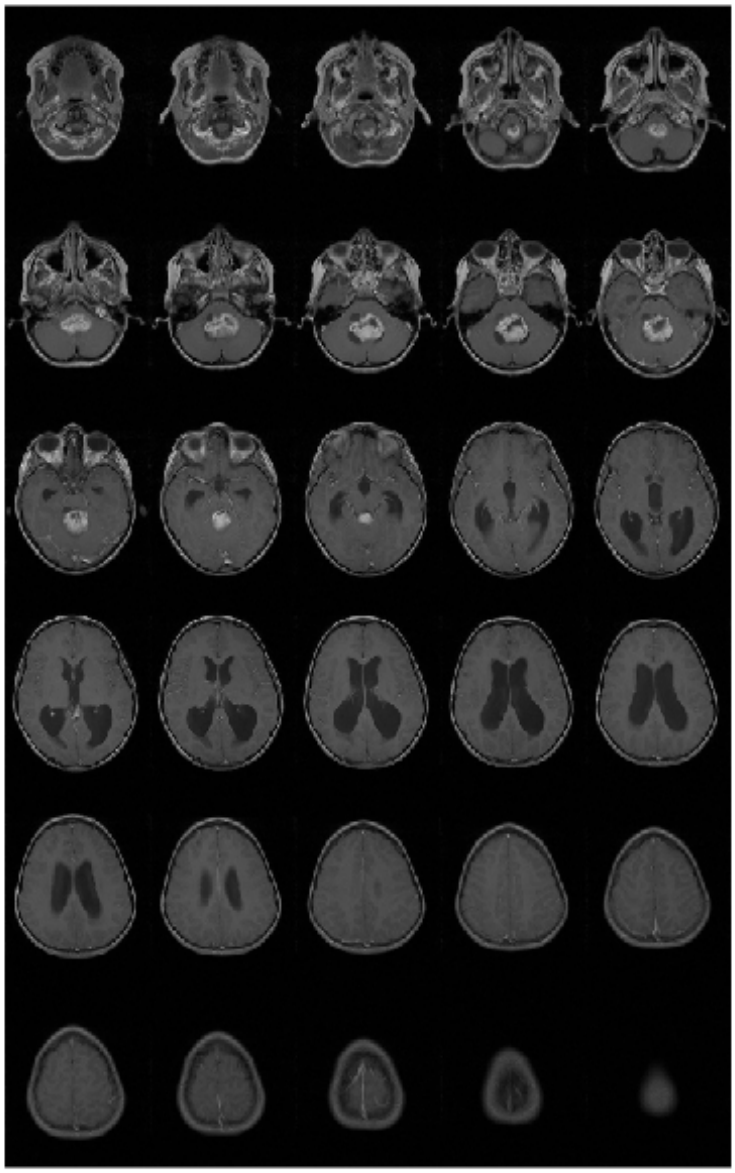


Figure B.5: Set 5, T2-weighted

# Appendix C

# Contrast Enhanced T1 - Weighted Dataset



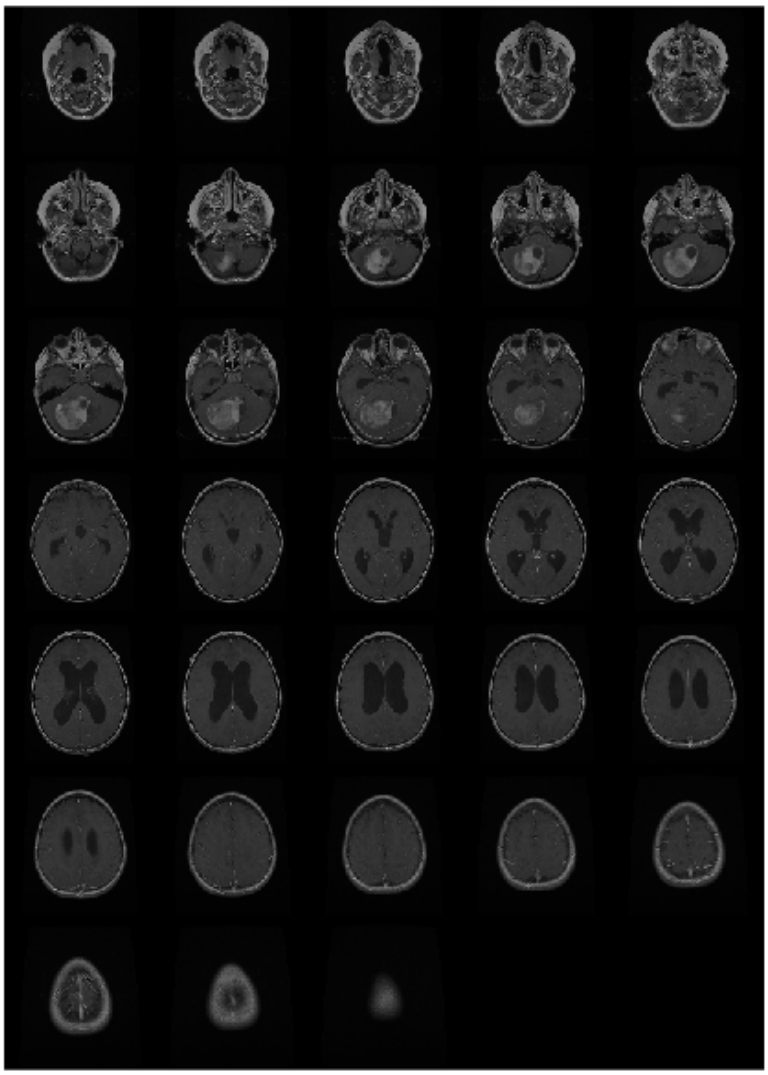


Figure C.2: Set 2, Contrast Enhanced T1-weighted

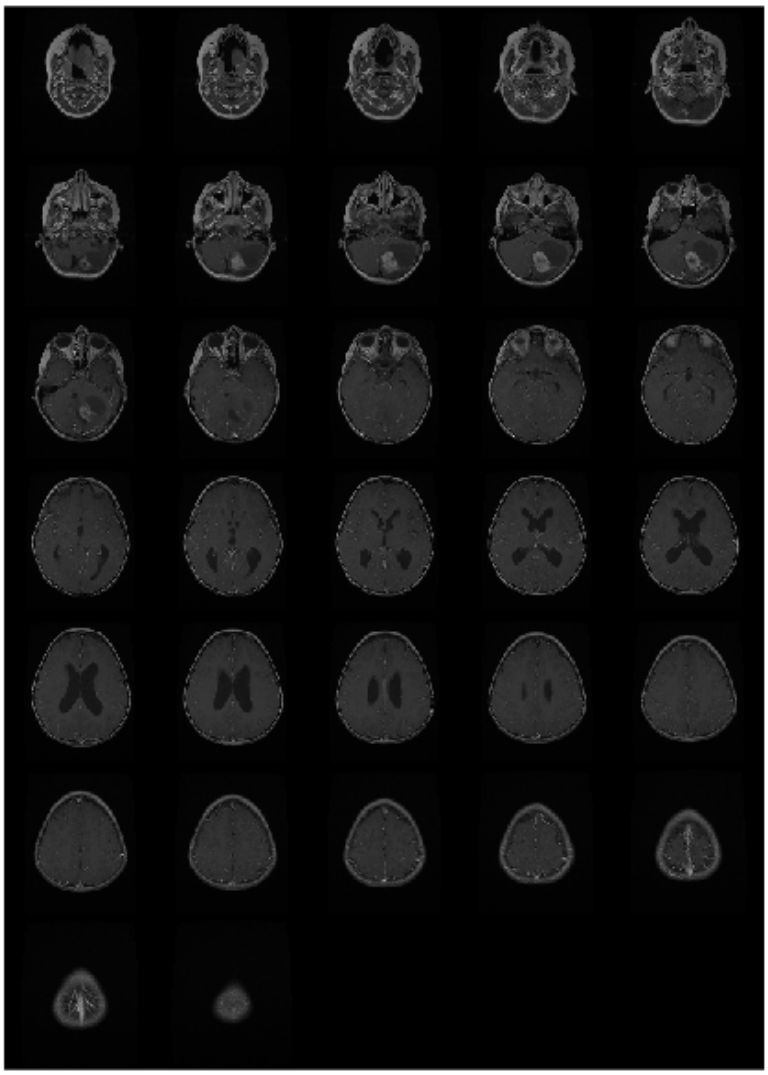


Figure C.3: Set 3, Contrast Enhanced T1-weighted

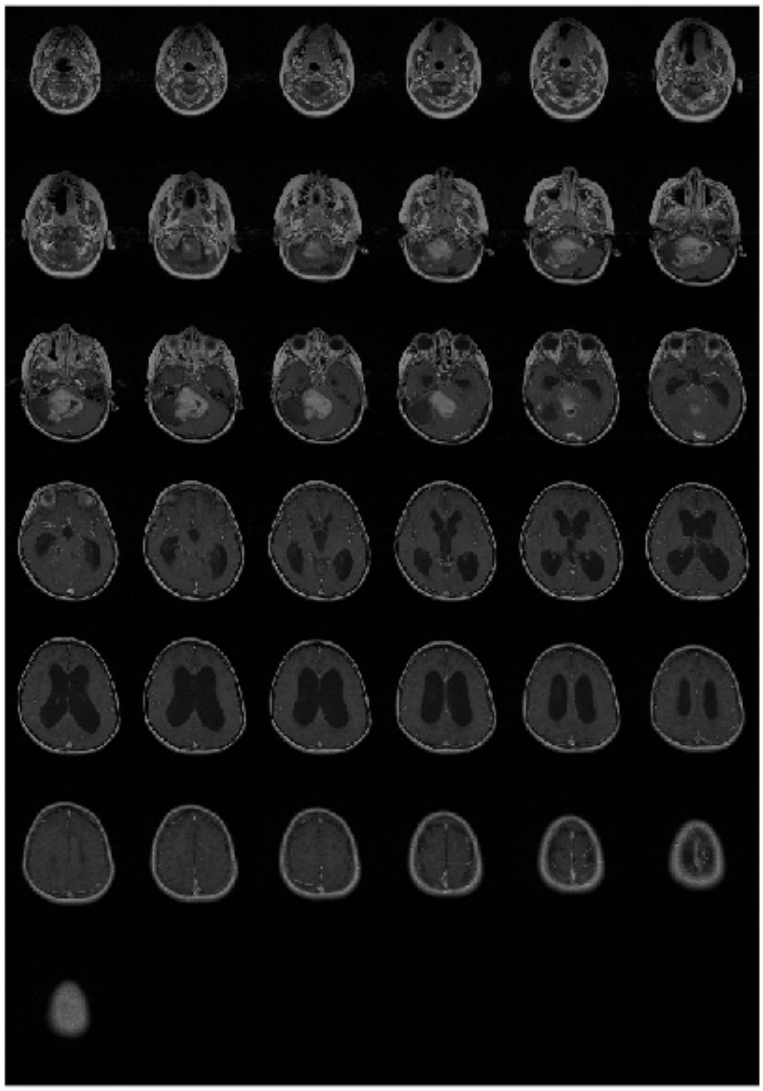


Figure C.4: Set 4, Contrast Enhanced T1-weighted

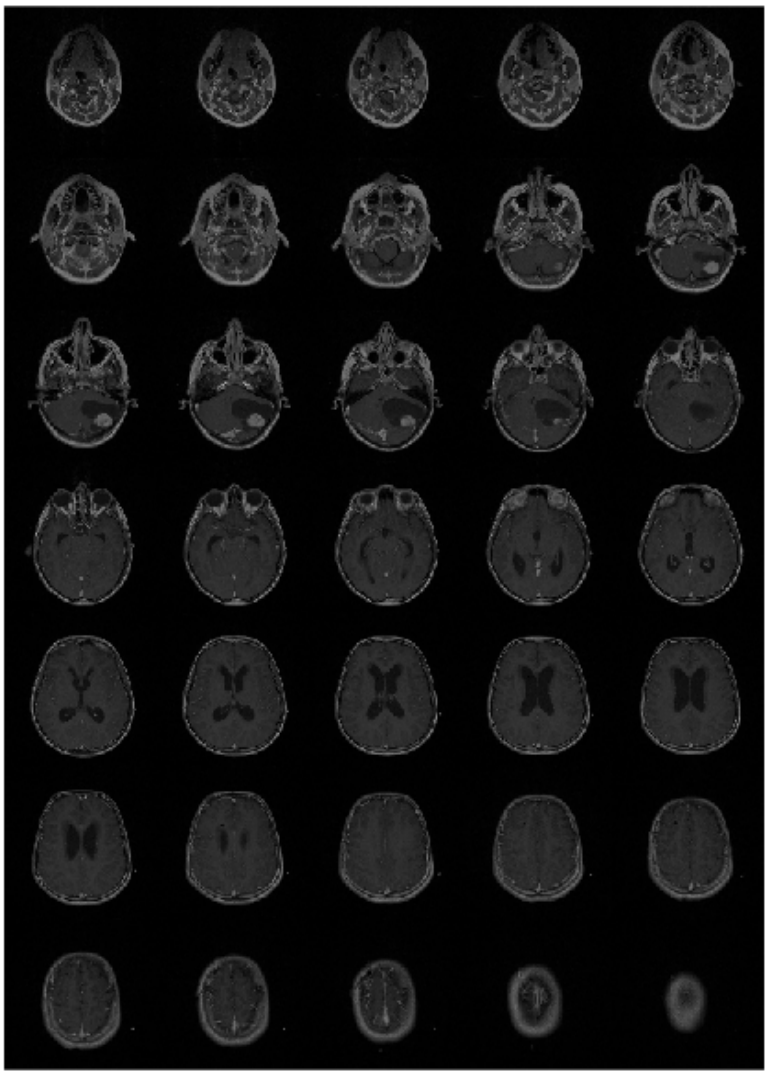


Figure C.5: Set 5, Contrast Enhanced T1-weighted