Fuzzy and probabilistic segmentation, and appropriate validation, applied to cardiac magnetic resonance images

Tasos Papastylianou

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2017

To my mother and father;
thank you for always believing in me.

Acknowledgements

This thesis and the many years of work it represents would not have been made possible had it not been for the encouragement and support — both at a professional and at a personal level — from some truly amazing people, to whom I am truly grateful for being part of this journey. I particularly wish to express my gratitude and appreciation to the following people:

Academics

My academic supervisor **Prof. Vicente Grau**, for his time, patience and never-ending guidance and support during my DPhil, and for allowing and encouraging me to pursue my own particular (and somewhat unorthodox) research direction.

My clinical co-supervisor **Dr Erica Dall' Armellina** for providing valuable clinical advice, materials, and support during my DPhil.

Funding body

My research was supported by the EPSRC via the RCUK Digital Economy Programme (grant number EP/G036861/1 – Oxford Centre for Doctoral Training in Healthcare Innovation).

University staff

The 'part programme-administrator / part super-hero' **Jo Armitage**, for being the most helpful and approachable person one could ever wish to have on their side from day one, and without whom, I suspect, the whole department would probably collapse. • The staff at Kellogg College, particularly **Sophie** for always welcoming people with a smile, and always having the piano keys ready for me; and **Lisa** and **Christopher**, who helped make my life a lot easier in the face of daunting paperwork.

Personal

My friends from the lab and Oxford in general: my labmates, especially **Chris**, **Jo**, **Ramon**, **Karl**, and **Carlos** who have been there since the beginning • **Ali**, **Marie**, **Paul**, **Simao**, **João**, and the rest of the gang from the CDT • **George** (a.k.a. "The Raccoon") for being my faithful Noodle-Nation buddy every Thursday come rain or shine, and for his genuine friendship • **Cooper** for being a true friend and never shying away from the truth • **Tingting** and **Aaron**; Tingting for being a friend I could count on, and for dragging me to all those interesting lectures and Aaron for being my go-to nerd; my Oxford experience would not have been the same without you guys • **Pavlos** (I miss our 3am coffee breaks) and **Artemis** • my flatmates and good friends over the years: **Alessandro** (a.k.a. my partner in crime) and **Tina** (seriously, you make the *best* cakes!); **Elona**; **Andrei** and **Monica** (thanks for being my biggest piano fan!); **Ben** and **Iulia**

(we *still* need to do that lasagna!) • the "Smart Handpumps" gang, particularly **Patrick** and **Joachim** • the Sentimoto gang, **Maxim**, **Lisa**, (and as always, Alessandro), for being crazy enough to co-found a company with me • everyone from the squash club, especially **Maxime**, **Kate**, and **Kinda** who kept it going against all odds • The girls at Mumus where I spent all my free time (it's so unfair they shut it down), especially **Alexandra**, **Micky**, and **Elena** for always making me feel welcome and sneaking extra ice-cream on my waffles • my Cypriot friends in Oxford, especially **Ioanna** (thanks for the motivational cat pics! Σταθερές αξίες!), **Nicolas** (I look forward to our next squash match!), **Christos**, **Christophoros**, **Xenia**, **Savvina**, **Andreas** and **Marios**, for making me feel like home.

My friends from a previous life of being a Bristol medic, who have stuck with me for the ride so far: **Adrien**, **Erasmus**, **Nik**, **Jo** • **Sophie**, **Richard**, **Dawn**, **Eleanor** • **Bhavin**, **Sameer**, **Jacek**, and **Zenonas**.

My old friends from Cyprus who are always there for me: **Christina** (you make me proud to be your friend), **Alex** (thanks for all the sci-fi books!), **Loukis** (thanks for being up for a walk and chat 24/7), **Kyriakos** (it is an honour to be your best man!), **Mike** (thanks for all the advice and pep talk!), **Refet** (thanks for always finding the time to geek out with me), and also **Stella**, **Giorgos**, **Louis**, and **Petros**.

My cousin **Stephanos** (for making the last year of the DPhil in quiet Leamington probably the most enjoyable year of my DPhil) and the rest of my extended family: **Stella**, **Ismini**, uncle **Andreas** and aunt **Chrystalla** • **Stelios**, **John**, **Catherine**, uncle **Christakis** and aunt **Thekley** • **Tasos**, **Eleni**, **Anastasia**, uncle **Nikos** and aunt **Aimee** — and all your loved ones, thank you all for standing by me as family.

Other friends, past and present, who stood by my side at one point or another, and helped me get through the DPhil one day at a time, for which I will always be grateful: **Irina**, **Laura**, **Kallia**, **Marina**, **Andreas** and **Valentina** (thanks for coming down to spend NewYear's with me!), **Dimitris**, **Sheena**, **Sonia-Stephanie**, **Roma**, **Maria**, **Allyssa**, **Hanelore**, and **Constantina**.

All the people from the lovely and welcoming Greek-orthodox communities in Oxford and Coventry, and the monastic community at the Monastery of St John the Baptist in Essex, especially **Father Ian** and **Father Nikolay** respectively, who were always happy to listen to my rants and imbue me with a sense of hope and calm. Also special thanks to **Olga** for the lovely company on several long trips to Essex!

My beloved grandmother **Katina** who sadly passed away during my DPhil.

Last but foremost, my mother **Katina**, my father **Ioannis**, and my brother **Stelios**, for always being there for me and supporting me at all times with their love, support, and advice, without my ever having to ask for it.

Other

I am indebted to **John McManigle** for the quality 'Oxford DPhil LaTeX template' he has made available online, which I used as a starting point and have heavily adapted for my own purposes • Also thanks to fellow PhD student **Aïcha Bentaieb** for our chat during MICCAI 2016, who pointed out further areas of impact for my work, which I had not yet considered at the time.

Abstract

Algorithms producing fuzzy and probabilistic (i.e. 'soft') segmentations are becoming increasingly popular. However, many of the unique strengths of such algorithms get overlooked, especially when used in the context of deterministic frameworks, which typically treat softness as label uncertainty, and tend to discard it as the final step. We maintain that such treatment results in loss of potentially useful information, which could be used to improve outcomes further. This is particularly the case with regard to validation algorithms, where such loss of information effectively renders validation unreliable.

When 'softness' is treated as a fuzzy measure, defined with respect to a suitable criterion, the uncertainty over such a measure can be further characterised and put to use. We explore the role of fuzziness and probability theory in characterising such notions of uncertainty over 'softness', and use it to a) fuse soft segmentations via their uncertainties, leading to improved outcomes compared to fusing by simple consensus, and b) to enable clinicians to optimise algorithms using clinically intuitive measures as opposed to fine-tuning unintuitive algorithmic parameters by hand. We also make theoretical predictions based on the semantics of fuzziness in the context of spatial overlap, and use them to propose the notion of a 'directional t-norm' and show that it leads to more reliable validation. Finally we propose ways of characterising modes of segmentation failure through the use of local-performance maps, fuzzy spatial / anatomical relationship masks, and validation sweeps.

In summary, this thesis provides a better understanding of the types of uncertainty that can be defined over an already 'soft' segmentation, and how they could be put to use to improve segmentation outcomes; and a better understanding of the semantics of fuzziness with respect to the validation of soft segmentations, leading to more reliable validation and evaluation in general.

While the applications presented in this dissertation have been demonstrated in the context of medical image segmentation, the methods and theory should also be more widely applicable to non-medical image analysis and computer vision in general.

Contents

List of Figures

List of Tables

Glossary and Abbreviations

AAM . . . . . *Active Appearance Models*

AMI . . . . . *Acute Myocardial Infarction*

ASM . . . . . *Active Shape Models*

CMR . . . . . *Cardiac Magnetic Resonance*

CPU . . . . . *Central Processing Unit* (computer component)

CT . . . . . *Computed Tomography*

DICOM . . . . *Digital Imaging and Communications in Medicine* (medical image file-format)

ECG . . . . . *Electrocardiogram*

EF . . . . . *Ejection Fraction*

EM . . . . . *Expectation-Maximization*

F_- . . . . . *False Negative*

F_+ . . . . . *False Positive*

FIMH . . . . . *Functional Imaging and Modeling of the Heart* (conference)

FISP . . . . . *Fast Imaging with Steady-state Precession*

G . . . . . *Gold-standard mask*

GPU . . . . . *Graphics Processing Unit* (computer component)

GTC . . . . . *Generalised Tanimoto Coefficient*

IHD . . . . . *Ischaemic Heart Disease*

LA . . . . . *Long Axis*

L . . . . . *Latent truth mask*

LGE . . . . . *Late Gadolinium Enhancement*

LGE . . . . . *MI* Myocardial Infarction

LV . . . . . *Left Ventricle*

MAP . . . . . *Maximum A Posteriori*

MICCAI . . . *Medical Image Computing and Computer Assisted Intervention* (conference / society)

MRI . . . . . *Magnetic Resonance Imaging*

PCA . . . . . *Percutaneous Coronary Angioplasty*

pT_c . . . . . *Pixelwise Tanimoto Coefficient*

PVE . . . . . *Partial Volume Effect*

RGB . . . . . *Red / Green / Blue* (pixel colour triplet)

ROC . . . . . *Receiver Operator Characteristic* (analysis)

ROI . . . . . *Region of Interest*

RV . . . . . *Right Ventricle*

SAD . . . . . *Sum of Absolute Differences*

SA . . . . . *Short Axis*

S . . . . . *Segmentation mask*

SNR . . . . . *Signal-to-Noise Ratio*

SPECT . . . . *Single Photon Emission Computed Tomography*

SSFP . . . . . *Steady-State Free-Precession*

STAPLE . . . *Simultaneous Truth and Performance Level Estimation* (algorithm)

STARE . . . . *STructured Analysis of the REtina* (project / dataset)

SV . . . . . *Stroke Volume*

T_c . . . . . *Tanimoto Coefficient*

T_- . . . . . *True Negative*

T_+ . . . . . *True Positive*

Ω . . . . . *Image domain*

1 Introduction

The two major themes in this thesis are the use of uncertainty in, and the appropriate validation of, soft segmentations in the context of medical images. This introduction outlines the motivation, thesis statement and claims, and a brief overview of the chapters ahead.

Contents

Prognosis, diagnosis, and management of medical conditions often relies on the interpretation and quantification of tissue properties, as evaluated through the use of medical images. This generally involves identification, characterisation and quantification of relevant tissue, or *objects* in the image, through the act of *segmentation* — the application of some labelling process on the image's pixels, such that each pixel is mapped onto a particular deterministic label or *class*. In the simplest case, the task is to divide the image into two constituent parts: a *foreground* class, made up of pixels denoting the object or tissue of interest, versus a *background* class.

More recently, the medical image analysis literature has been increasingly shifting

towards the notion of so-called *soft* segmentations [1–5]; i.e. rather than mapping each pixel exclusively to a single class, a pixel can be *partially* mapped to one or more classes instead, allowing for notions like "ambiguity", "tissue composition", "label probability", and "uncertainty" in general to be expressed over the classified pixels.

There are several mathematical frameworks that can be used to model this uncertainty in the more general sense; the main ones, and the ones we focus on in this thesis, are the notion of *fuzziness*, as defined through the various areas encompassed by *fuzzy theory*, particularly *fuzzy set theory*, *fuzzy measures*, and *fuzzy logic*; and the notion of *probability*, as defined through the various branches of *probability theory*, and in particular, the *frequentist*, and *Bayesian* interpretations.

This thesis aims to explore the semantics underlying such non-deterministic results within the scope of these frameworks and in the context of medical images, particularly with respect to expressing clinically relevant notions of uncertainty; and to put these to use, thus leading to improved, more clinically relevant, and more reliable / reliably validated segmentation algorithms.

1.1 Motivation

While there is an increasingly large number of segmentation methods producing soft / non-deterministic outputs, literature on the semantics of this 'softness', and how it can be put to use to gain further clinical insights still seems rather restricted. Many authors seem content to use such soft segmentations simply as an intermediate step enabling the researcher to obtain a "better" *deterministic* segmentation as the end-product, even though evaluation of clinical parameters directly from soft masks has been shown to be more accurate and precise (e.g. [6, 7]). And even when soft masks *are* used, estimates obtained from such segmentations still tend to lack appropriately defined confidence intervals [8], despite the fact that soft segmentations are a natural framework for obtaining such a confidence range over

the result — since each pixel value represents a quantifiable 'measure' that could be ascribed a confidence *range*, rather than simply a 'class' which can only be ascribed a *degree* of confidence[1]. Instead, the uncertainty over 'soft' pixels tends to be overlooked, or worse, treated as a nuisance factor; in fact, even worse, 'softness' in itself is often treated as a nuisance factor and 'corrected', typically by being thresholded out (e.g. [9–19]). One area where this happens consistently, is in the evaluation of such algorithms: whether the end result of an algorithm is intended to be soft or deterministic, validation often defaults to conversion from soft to deterministic anyway, so as to perform validation using traditional (but, as we claim here, unreliable and outdated) validation algorithms.

The motivation behind this thesis is to therefore explore the semantics behind 'softness', and in particular when, whether, and why they can be more useful / reliable in themselves compared to their corresponding deterministic outcomes; to explore soft segmentations as objects able to express useful, measurable forms of uncertainty, and, rather than treat this as a nuisance factor, to make use of such measures of uncertainty to improve segmentations and clinical outcomes arising from them; and finally to explore suitable methods for their validation, which appropriately take into account the semantics of 'softness' and particular nature of these algorithms on evaluation, therefore ensuring more specific, accurate and reliable validation.

1.2 Thesis statement and claims / contributions

Thesis statement:

Fuzzy and probabilistic, (i.e. 'soft') segmentation approaches confer significant technical and clinical advantages over traditional deterministic segmentation algorithms, which go beyond a simple assignment of label uncertainty, but can represent useful

[1] This is also often simply taken to be the soft mask used to generate the class, which in our opinion amounts to a rather limited use for softness.

information in their own right, with particular underlying semantic interpretations depending on context.

Of particular interest is the fact that one can characterise and quantify various measures of uncertainty over such fuzzy / probabilistic measures themselves, including 'clinical' uncertainty. Our thesis is that one can make use of this knowledge to improve their clinical relevance and outcomes.

Furthermore, such probabilistic and fuzzy approaches require novel evaluation approaches, as traditional validation measures used in the context of deterministic algorithms are shown to be inadequate and unreliable, and therefore inappropriate in the context of probabilistic / fuzzy segmentation algorithms

Claims and contributions

Chapter 3

i. A theoretical exploration of the semantics of 'softness' and 'uncertainty' in the context of medical image segmentation
ii. A framework for the fusion of soft segmentation algorithm outputs based on pixelwise measures of uncertainty
iii. A fuzzy generalization of the entropy-based "Intensity-based class-uncertainty" algorithm by Saha and Udupa [20], for use as a measure on uncertainty for the fusion of soft segmentations.
iv. A fuzzy spatial / anatomical uncertainty measure that allows adjusting for overall strength / spread of result (or of individual components).
v. A measure of 'clinical uncertainty' for segmentations defined as a fuzzy measure of the discrepancy between a set of clinical estimates versus predefined physiological constraints.
vi. A method for converting deterministic segmentation algorithms to more clinically relevant soft counterparts than simple averaging / consensus of the segmentation space, making use of the above uncertainty measure, but with

physiological constraints automatically obtained from the initial segmentation (published as [21]).

vii. A method making use of the above measure of 'clinical uncertainty', allowing a clinician to guide a segmentation algorithm (soft or otherwise) towards more clinically relevant / reliable results, representing near-optimal parameter-tuning of the algorithm's intrinsic parameters for the clinical scenario in question, using only clinical knowledge in the form of simple physiological constraints (published as [21]).

Chapters 4 & 5

i. A theoretical exploration of the spatial semantics of fuzziness in the context of overlap and validation of segmentations, including a proof for the existence of upper and lower bounds for the overlap of 'spatially' fuzzy pixels, corresponding to the most pessimistic and most optimistic overlap outcomes possible, and their significance with respect to the validity of 'fuzzy' validation operators and validation outcomes.

ii. An understanding of why the conventional approach to validation by "thresholding" is unreliable as a validation operator for soft segmentations, as an operator that violates the theoretical bounds outlined above, and a practical demonstration confirming the above prediction on synthetic and medical datasets (published as [22]).

iii. A representation of pixels at object boundaries as homogeneous fuzzy pixels with intrinsic orientation, which can be quantified or modelled appropriately, leading to a better understanding of the manner in which such boundary pixels overlap with respect to validation.

iv. A novel directional triangular norm based on the concept of boundary pixels, suitable for the validation of 'spatially' soft segmentations, shown to lead to more reliable validation compared to both "thresholding" and state of the art methods for 'soft' validation (published as [22])

Chapter 6

i. The use of local-performance maps for assessing the *distribution* of a segmentation's performance
ii. The use of fuzzy spatial / anatomical relationship masks, for characterising modes of segmentation failure related to the direction or distance from a gold standard or point of interest
iii. An evaluation of distance transforms in the context of evaluating distance from fuzzy objects
iv. A generalisation of the Hausdorff distance applicable to fuzzy objects
v. Validation sweeps and their role in the assessment of segmentation failures due to the unwanted presence of specific features

Publications arising from this thesis

- Tasos Papastylianou et al. "Fuzzy Segmentation of the Left Ventricle in Cardiac MRI Using Physiological Constraints". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2015, pp. 231–239
  — (relevant work presented in chapter 3).

- Tasos Papastylianou, Erica Dall' Armellina, and Vicente Grau. "Orientation-Sensitive Overlap Measures for the Validation of Medical Image Segmentations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 361–369
  — (relevant work presented in chapters 4 & 5).

1.3 Limitations and outline of future publications

Availability of data and algorithm sets

Cardiac MRI was largely used in this thesis to motivate the problem, in the context of a non-trivial segmentation setting (and in accordance with the lab's clinical topic

of interest); however, the work presented here is not specific to this particular clinical context, and is more broadly applicable to medical image segmentation in general.

A practical limitation which partly defined the nature of the work and experiments presented here was the availability and nature of available datasets with corresponding existing and available-for-use algorithms compatible with such datasets; the early stages of this work focused on bespoke algorithms with respect to bespoke data provided by our clinical partners, which was limited at the time. This had a few practical implications: the initial dearth of data with compatible algorithms meant that:

- some of the ideas presented here were necessarily presented as proof of concept, and while these ideas did lead to conference publications demonstrating the novel concepts presented here, more conclusive and substantial journal publications will require more thorough and comprehensive experimental demonstrations on larger cardiac datasets with compatible algorithms, and will therefore be the subject of future work stemming from this thesis.
- where it was judged necessary to quantify an effect, rather than simply demonstrate it via simple proof-of-concept experiments, we made use of alternative publicly existing datasets for which existing algorithms and outputs were available, but which were not necessarily cardiac in nature. In particular, in chapter 5, in order to quantify the effect of d-norms experimentally compared to more traditional and state-of-the-art approaches, we made use of the STARE (STructured Analysis of the REtina) Project [23], which provides a clinical dataset of 20 images of human retinae, and corresponding manual and algorithmic outputs, freely available online.

At the time of writing this introduction, the first (to this author's knowledge) *publicly* accessible standalone algorithmic implementation, designed explicitly to be compatible with the format of certain publicly available cardiac datasets (e.g. from existing segmentation challenges, like the MICCAI Left Ventricle [24] and Right

Ventricle [25] segmentation challenges) has been released ([26], a convolutional neural network based algorithm; first public implementation released April 2017). This greatly facilitates the above goal for generating a large number of suitable outputs on which to demonstrate our contributions in future journal publications. We present below the outline for one such intended publication stemming from this thesis, in accordance with the conference papers already published.

Outline of intended publication on d-norms and their role in the validation of soft segmentations

The contents of Chapters 4 & 5 are a significant contribution, both in terms of the state of the art, and in terms of flagging important issues with established current practice in the field of medical image segmentation. The journal publication stemming from the work in these chapters will build on the work published as Papastylianou *et al.*, 2016 ([22]).

Motivation: The aims of the publication will be to:

- present the theoretical contributions outlined in chapter 4, which explain why the established practice of thresholding soft segmentations for the purposes of validation, and some state of the art methods, are in fact practices leading to unreliable validation.
- motivate the concept of a boundary pixel as a homogeneous fuzzy pixel with intrinsic orientation, giving rise to the notion of a directional t-norm, and demonstrate how its use within overlap measures leads to more reliable validation, and an improvement over state of the art.
- demonstrate the above on a large cardiac dataset, with a known gold standard for each image set, and for a large number of segmentation 'algorithms' applied per image set; this serves two purposes: one, demonstrate the superiority of d-norm based validation experimentally, as per [22], but on a dataset of a significantly larger size and complexity; and secondly, explore instances where the unreliability of a validation operator leads to a false ranking between

competing 'algorithms' with respect to ground truth, and the extent to which d-norm based validation protects against this.

Materials and Methods: We will make use of the publicly available cardiac segmentation suite mentioned above ([26]) and the publicly available data with which it has been designed in mind (namely the MICCAI Left Ventricle [24] and Right Ventricle [25] segmentation challenges, both of which provide training and validation data).

The algorithm in [26] is a training-based algorithm; therefore one can separate the training set into a number of subsets (e.g. 10), and treat each subset independently as a single 'algorithm'[2].

Furthermore the algorithm seems stochastic in nature (although, if the degree of stochasticity is not adequate, bootstrapping can be employed), therefore ensuring that for each subset we can obtain a number of 'component' segmentations, all of which are related but not identical, such that an 'average' segmentation can be obtained for each subset.

One of the main challenges in quantifying the reliability of fuzzy validation *operators*, is that their outputs need to be compared against a *known* value, referred to as the *latent truth*. and representing the validation of the underlying 'true' (i.e. crisp, higher resolution) objects that their 'soft' equivalents happen to represent. In [22] we dealt with this problem by starting with high resolution segmentation / gold standard pairs first, and generating corresponding fuzzy segmentation / gold standard pairs from them, such that we could compare the outputs of various fuzzy validation operators acting on the fuzzy sets, against the 'true' high resolution set (which only has a single possible validation outcome since it is non-fuzzy).

Given the presence of multiple 'components' generated by each 'algorithm', we can also try a different approach. The gold standard 'contours' are generally given at subpixel resolution. This means a 'soft' mask can easily be generated where fuzzy

[2]If other independent algorithms compatible with these datasets are made available at the time of writing these can also be included in an appropriate manner.

values reflect pixel occupancy by volume. For the segmentations, we can similarly obtain two versions: one at subpixel resolution, by computing a 'shape-based average' from the components (e.g. as per [27]); and a fuzzy mask at the 'lower', native resolution, as a normal unweighted average of the component labels, which should closely correspond to occupancy by volume w.r.t. the high-resolution segmentation. This creates per image set, the required 'hi' and 'low' resolution gold-standard, and a 'hi' / 'low' resolution segmentation pair for each 'algorithm', enabling evaluation of the fuzzy operators on the 'low' resolution segmentation / gold standard pairs, and comparison with the 'latent truth' as obtained from the 'hi' resolution pairs.

Intended outcome: The publication should provide experimental evidence that:

- D-norm based fuzzy validation is accurate and precise with respect to the latent truth.
- Traditional threshold-based validation is unreliable, in that it is inaccurate and imprecise, as demonstrated in [22].
- Unreliable validation operators can result in mis-estimation of the relative quality and ranking between algorithms w.r.t. the latent truth; d-norm based validation operators should give more reliable comparison of performance between competing algorithms.

1.4 Overview of Chapters

The content of this thesis report is organised in the following manner:

Chapter 2 presents the clinical background relevant to the thesis, state of the art in heart segmentation from MRI to set the research into context, and the mathematical background which underpins the work, introducing basic notation and concepts to be used in later chapters.

Chapter 3 expands on the role and advantages of non-deterministic approaches in medical image segmentation, particularly in the context of the Partial Volume Effect, and reliable clinical parameter estimation. We discuss work done in the field, and focus on the interpretation of fuzziness and probability, and the distinction between the two, where the former is a broader concept and a superset of the latter. We particularly discuss how fuzziness relates to different types of uncertainty, particularly as clinical ambiguity or semantic uncertainty, and we discuss how to transform this uncertainty from being a nuisance factor into useful input, such that it can be put to use to improve segmentation outcomes. We propose two frameworks on that principle: a framework to combine soft segmentation algorithms, to produce a fused soft segmentation output for which individual pixel uncertainty is minimized; and a framework for the construction of fuzzy / probabilistic variants from traditional algorithms, that allows for more clinically relevant algorithms, guided by clinician input via clinically-relevant parameters, rather than unintuitive algorithmic ones, that are intrinsic to the algorithm's inner workings.

Chapter 4 functions as a treatise on validation as it applies to probabilistic and fuzzy segmentations. We discuss the limitations of current validation practice, and propose ways to improve reliability of validation for this category of segmentations. We put forward a theoretical framework underlied by a semantic interpretation of fuzziness specific to the context of fuzzy pixels at the object boundary, and use it to make claims and predictions about theoretical constraints in fuzzy validation, and propose on the basis of that, the novel concept of a *Directional T-Norm* (or *d-norm* for short).

Chapter 5 puts the above theoretical contributions to the test through experiments on synthetic and clinical data, aiming to demonstrate the existence and significance of the theoretical bounds to validation, and to quantify the improvement afforded by a d-norm. We show that d-norms significantly improve accuracy, precision, and therefore reliability of validation operators in

fuzzy segmentations, over established practice and state of the art validation algorithms. More importantly, we demonstrate during this process that the conventional 'thresholding' approach, which is currently the mainstay of validation in the segmentation literature, is *particularly* unreliable with respect to validation of soft segmentations, and we urge the community to stop using it and explore more suitable alternatives.

Chapter 6 goes beyond standard validation, which typically only reports 'to what extent' a segmentation succeeds or fails to match the gold standard, and examines how one might characterise segmentations further by identifying the *modes* in which a segmentation succeeds or fails, both in the qualitative and quantitative sense. This is done through the use of local-performance maps, fuzzy spatial / anatomical relationship masks dealing with failures in a particular direction or at a particular distance from a gold standard, and validation sweeps dealing with failures due to the unwanted presence of particular features. The notion of distance from fuzzy objects is also discussed, as well as a generalisation of the Hausdorff distance applicable to fuzzy objects.

Chapter 7 concludes the thesis by summarising the claims and contributions made herein, discussing open questions that remain or arise as a result of the presented findings, and providing a future outlook on how to address them.

If I had only one hour to save the world, I would spend fifty-five minutes defining the problem, and only five minutes solving it.

— (attributed to Albert Einstein)

2 Background

We divide this chapter into two parts: the first part discusses the need for segmentation in clinical practice, both in general and with particular reference to cardiac MRI as a demonstrative case, and discusses particular challenges, conventions, and traditional and state of the art approaches with respect to cardiac MRI segmentation.

The second part introduces the mathematical background that forms the basis for discussion with respect to probability and fuzziness in the remainder of this thesis.

Contents

2.1 Clinical background

The history of physio-anatomical imaging of the human body and its use in clinical practice started with the discovery of X-rays in the beginning of the 20th century [28]; by the latter half of the century, medical imaging had already established itself as a medical specialty, particularly through the use of tomographic (i.e. cross-sectional) imaging techniques, such as Computed Tomography, Ultrasound Scanning, and later on Magnetic Resonance Imaging [29]. While little 'processing' may have been necessary with early X-ray films, the need for image-processing was evident from the outset for cross-sectional 'scanning' techniques, as these tended to rely on the transmission, detection, and interpretation of "signals".

Initially, the purpose of such 'medical image processing' was of a purely *qualitative* nature, aimed at bringing out or isolating salient radiographic features and structures in the image, that would aid the clinician in their diagnosis, by allowing them to apply their expert knowledge of anatomy and pathophysiology to make sense of them.

For example, in computed tomography, the signal obtained represents the *radiodensity* (i.e. amount of X-radiation absorbed) at a particular 'unit' area in the scan (interpreted as a pixel in the reconstructed volume). This is measured in Hounsfield units, a normalized index of X-ray attenuation based on a scale of -1000 defined for air and 0 for water at standard pressure and temperature [30]. In order to visualise the relevant anatomy as an image on a computer screen, Hounsfield Units need to be reinterpreted as grey-level pixel intensities in the range 0–1; since different tissue types are characterised by different X-ray attenuation profiles, visualising a particular anatomical structure involves selecting an appropriate *window* around a particular value of interest (the *level*). This enables a clinician to visualise the relevant anatomy as clearly as possible, by isolating that particular tissue, and providing good contrast from surrounding tissues of a different type.

Increasingly, however, medical imaging is now also used to obtain *quantitative* information, relating to evaluating function, obtaining size measurements from the images, defining suitable biomarkers, etc. But, for such information to be usable in clinical practice, it needs to be relatively quick and easy to obtain and use; the need to perform image *analysis* to retrieve, quantify, and use such information in an efficient, consistent, and clinically viable manner, has led to (bio)medical image analysis becoming a research field in its own right [31]. Developments in computing over the past few decades, both in terms of hardware and suitable algorithms, have played a crucial role in the field; in particular, it has borrowed heavily from, and developed alongside the fields of pattern recognition, image processing and computer vision. The specific, and differentiating aim of biomedical image analysis is to provide streamlined tools and methods for automated and semi-automated analysis in medical images specifically, providing personalised tools for diagnosis and clinical management.

2.1.1 Segmentation in clinical practice

Before any measurements can be made on an image object, it needs to be identified on the image first. In other words, any pixels deemed to correspond to the object / tissue of interest need to be labelled as such. This process is called *segmentation*. In medical images, in most cases, the gold standard for segmentation is considered to be manual delineation of the object by an expert clinician, or even a weighted average of several manual delineations from multiple expert clinicians (e.g. [11, 32], etc). However, except for the most trivial cases, such manual segmentation is tedious and prohibitively time-consuming, therefore much research has focused on automated and semi-automated techniques for segmentation [17].

An equally important process central to medical image analysis is *registration*, that is, a one-to-one mapping of equivalent structures between *two* or more images. This usually takes the form of a *transformation* problem, i.e. the 'warping' of one image in some rigid or non-rigid fashion, so as to maximise overlap of corresponding structures in the target image. Registration techniques are, for the most part,

beyond the scope of this thesis, but we mention it here in passing, in recognition of its important role in image processing in general, and specifically in that it links with segmentation and validation in two important ways:

- Many segmentation algorithms often involve one or more registration steps. The reverse also applies, i.e. some registration algorithms rely on a segmentation step first, and rely on the segmentation output to transform one image onto another, such that the overlap between two "equivalent" segmentations is maximized.

- Registration relies on good quality *similarity metrics* for performance, since the overlap of the transformed image to the reference image needs to be evaluated as a measure of whether registration is successful or not. While the similarity metrics used in registration are not necessarily the same ones used in the validation of segmentation algorithms, many of the concepts described in this thesis with respect to validation apply equally well to registration (and indeed one method of driving / assessing registration quality is through validation of their derivative segmentations).

Segmentation, both medical and otherwise, is a large field in itself; while there are some common themes and classical approaches in medical image segmentation in general (e.g. region homogeneity / thresholding approaches (e.g. [20, 33]), traditional level set [34] and snake-based [35, 36] approaches, graph-cuts [37] etc), for most non-trivial medical datasets and applications, segmentation research tends to be both modality and specialty / investigation specific.

Therefore, for the rest of this thesis, we will discuss the relevant concepts in the context of segmentation in MRI images of the heart, and Cine MRI in particular as a demonstrative case, to illustrate the concepts of fuzziness and validation in non-trivial algorithmic contexts.

However, we reiterate that the work presented in this thesis is generally applicable, and not specific to any single modality or image type, and that Cardiac MRI is only used to motivate the problem more clearly in an area with an established segmentation literature. A high-level description of the clinical context is thus not necessary here, and is beyond the scope of this thesis; therefore we only provide a brief outline of the clinical context in the next few paragraphs, before moving on to the next section, where we will focus more on the state of the art methods for medical segmentation, as seen through the lens of the cardiac segmentation literature.

2.1.2 Cardiac Magnetic Resonance — clinical context and conventions in segmentation

Cardiac Magnetic Resonance (CMR) has become an important imaging modality in cardiac disease, as it provides a high-quality alternative imaging solution for a number of investigations traditionally undertaken using other modalities, such as 2D / 3D Echocardiography for chamber quantification; CT and nuclear scans for perfusion; and even invasive procedures like coronary angiography. CMR — with or without contrast — can be used to obtain the same clinical information with comparable accuracy or better [38], conferring several advantages in the process:

- it is a non-invasive imaging technique and involves no ionising radiation
- it provides clear, high-resolution images with good tissue contrast
- with 'Cine' methods, it allows the acquisition of 3D[1] images with a further 'time' dimension, meaning the heart can be visualised throughout the cardiac cycle

[1]Strictly speaking, many of the methods we refer to here as "3D" are in fact better described as a 'stack' of 2D slices along the 3rd dimension, rather than a true '3D snapshot' in the physical sense. This has important implications, since it can introduce further sources of error or limitations, such as highly anisotropic pixels due to the large distances between slices as compared to the distances between individual pixels in the 2D slices; potential misalignment between slices which then needs to be algorithmically corrected; potential variability between the various slices, caused by the fact that they are captured at different points in time, etc. However, for the sake of simplicity we will use the term '3D' throughout this thesis to refer to any such image modality which results in a three-dimensional dataset in the broader sense.

- it provides the ability to visualise and characterise tissue composition (e.g. infarcted from non-infarcted)

But perhaps more importantly, it allows this large range of investigations traditionally performed on separate modalities, to be performed in a *single* modality as part of a routine diagnostic work-up and provide multiple aspects of heart anatomy and function, including cardiac motion, blood flow, geometry, tissue characteristics, etc, in a single scanning session.

CineMRI '3D+t' images

The *Cine MRI sequence* in particular, allows the acquisition of short-axis functional slices, with a temporal resolution high enough to detect consecutive phases within the cardiac cycle [39] (i.e. a '3D + time' image sequence). Due to limitations with temporal resolution in the acquisition process, these are generally not collected in real-time, but reconstructed from data acquired over a number of cardiac cycles, where the appropriate timing for each phase within the cardiac cycle is achieved by synchronising the data acquisition pulse with the corresponding phase in the ECG [40]. All timeframes for each slice are acquired at a single breathhold to minimize through-slice movement of the heart resulting from respiration; however, for practical reasons each slice tends to be taken at a separate breathhold (since the acquisition time required for the kind of protocols routinely used for Cine MRI makes it difficult to hold one's breath for that long). The 3D+t images produced can be used to easily and accurately ascertain clinical parameters pertaining to heart function, such as Ejection Fraction, Stroke Volume, Ventricular wall mass, movement and thickness at the various stages of the cardiac cycle.

Conventions and segmentation protocols in CMR

One major problem encountered when developing automated and semi-automated algorithms for use in any particular discipline, is the need for *standardization*. In order to be able to trust and meaningfully compare the segmentation outputs of different algorithms (or even of experts), there needs to be some sort of guarantee

that they are in fact segmenting, and thus measuring, the exact same thing, to the exact same specification. This becomes even more important during validation, when the quality of a segmentation (and the process that produced it) is judged by comparing it to a gold standard; if the two were obtained according to differing specifications, such an evaluation would be misleading at best.

In the case of CMR, efforts towards such standardization are evident in the guidelines published by the American Heart Association [41], which is now an accepted clinical standard[2]; it was designed with the aim to "optimize and facilitate communication between cardiac imaging modalities for research and clinical applications". It defines a model of the heart consisting of 17 segments of physio-anatomical significance, and plane orientations for image slices that optimise identification and use of these, while also being clinically relevant, namely *Short Axis* and Horizontal / Vertical *Long Axes* orientations (see fig. 2.1).

Beyond the above guidelines, however, regional protocols apply; sometimes such protocols are clinically motivated and vague, variable, or unspecified at the time of analysis, whether this is conducted by automated algorithms or human personnel; in the case of cardiac segmentation, typical examples of such inconsistencies include the varying inclusion of trabeculae and papillary muscles of the left ventricle (e.g. many automated algorithms make an assumption of smooth, "circular" contours, often irrespective of the protocol used in the available gold standard), the varying inclosure of the most basal slice and outflow tract, anatomically vague boundaries for the right ventricle outflow tract, etc; adherence to arbitrary but clearly defined segmentation

[2] A (personal) criticism of this standard is that it is designed with *clinical* interpretation and needs in mind, rather than because of its suitability for optimal automated analysis; it is entirely possible that images obtained in the natural coordinates of the MRI machine would be more useful and fruitful from a computational analysis perspective (e.g. easier identification of position, size and orientation of the heart with respect to neighbouring anatomy and specific relevant landmarks). However, at the same time, this protocol ensures that in-plane resolution, which is considerably higher than between slices, is more likely to coincide with structures of clinical interest worthy of analysis; it also ensures some shape consistency, such as the 'circular' appearance of the left ventricle in Short Axis views, which are features exploited for segmentation.

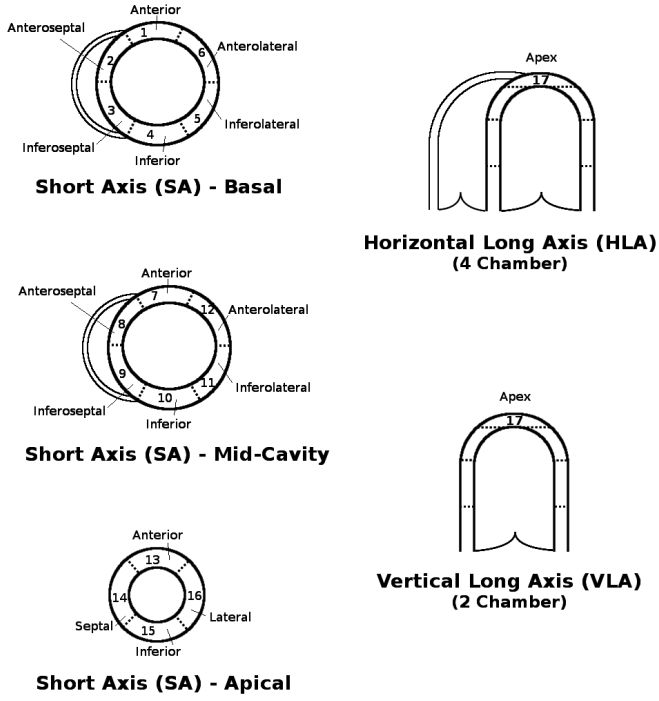


Figure 2.1: The AHA 17-segment standard, showing the recommended Short-Axis and Long-Axis orientations for cardiac imaging. Reproduced from Cerqueira *et al.* [41].

protocols by itself has been shown to improve the reliability of segmentations, both by automated and human agents [42].

While the large variance and arbitrariness of regional protocols is a problem, this is not prohibitive to relying on automated segmentation algorithms for clinical measurements, as long as such algorithms are consistent with the regional specifications, otherwise care must be exercised when interpreting such results. It is for this reason that segmentation challenges (such as the MICCAI Left Ventricle [24] and Right Ventricle [25] challenges) are typically required to state such protocols clearly, so that competing algorithms can be assessed on a focused, fair, and accurate basis.

2.2 State of the art in CMR segmentation

A recent review by Petitjean and Dacher [1] on state of the art approaches to cardiac segmentation, broadly categorises segmentations into algorithms that make

weak, or no prior assumptions, versus algorithms with strong prior assumptions; and algorithms that are fully automatic, versus algorithms which require either moderate or significant clinician input as an initialization step. We would add to this two more categories: algorithms that result in deterministic segmentations — that is, classifying each pixel as either belonging to a class entirely, or not, without allowing 'in between' situations — versus algorithms that result in probabilistic, or more generally soft, outcomes; and monolithic methods applying a single concept for segmentation, versus ensemble algorithms that rely on some sort of fusion of multiple classification subcomponents or approaches for their final result. Naturally, the above categories for the most part represent a spectrum of approaches rather than a dichotomy, and most algorithms fall somewhere in between. We summarize some of the more important approaches below:

2.2.1 Deformable models

Deformable models (sometimes called active surfaces [43]) are the conceptual analogue of active contours [35] (or 'snakes') to three-dimensional images. In the same way a 2D contour evolves according to an energy functional, 3D deformable models rely on the deformation of a pre-constructed model towards a 3D object's features, constrained in its evolution by prior knowledge of the model [44]. A limitation of the algorithm is that it relies on a good pre-existing initial 3D model for initialisation, which is already close enough to the object boundaries, therefore potentially making unsound assumptions, such as limited variability of heart shape during pathological conditions [45].

2.2.2 Active Shape Models and Active Appearance Models

Active Shape Models (ASM) [46] rely on a mathematical representation of object shape, defined as a vector of landmark points. Once this set of landmark points is trained on a large set of data, a distribution of such vectors, and by extension object shapes, is obtained, which can be used to constrain the evolution of a model towards object features in a way that retains the overall shape class of the model, as long as

initial placement of the curve is close enough to the object. 3D Active Shape Models have been used for cardiac segmentation [47], and variants incorporating pixel intensity information — called Active Appearance Models (AAM) accordingly [48] — with reasonable success. Like all training-based methods, a limitation of these models is that they rely heavily on the training set used to train the model, and the extent to which it is general enough to accommodate pathology, but without losing specificity in terms of object shape.

2.2.3 Atlas registration with segmentation propagation

Registration of an atlas to a target image refers to the process of shifting and deforming an atlas in such a way that the features of the object in question in the atlas match the corresponding features of the object in the target image [49]. An atlas in this sense then is an image where all the representative and important anatomy is shown, such that it can have a good chance of being matched to a reasonable array of targets; occasionally atlases are produced by averaging many source images together, in the hope that this will reflect to an extent the anatomy of all constituent images, and will therefore have a better chance of representing the more general and salient features of the object [12]. If the atlas is pre-segmented, then one can easily 'propagate' its segmentation to the target image by using the same deformation field resulting from the atlas registration (e.g. [45, 50–53]). Limitations of this method are the reliance on training data, which generally reduces its robustness to new pathology, or any anatomy not adequately represented in the atlas. Furthermore, as would be expected, even in the presence of accurate registration, labelling errors in the atlas itself will carry forward to the final segmentation result [53]. Ironically, registration becomes a much easier problem if performed on the respective segmentations of the source and target images; this is described as a sort of 'chicken-and-egg' problem, which some authors address by using iterative methods [54]. If not performed on segmentations, due to the variable intensity profiles and movement artefacts between different MRI studies, exact registration is less robust, and can be a source of errors which are then carried on to the segmentation result [53]. In

these circumstances, the registration can be improved by limiting the algorithm's operation to a Region of Interest (ROI) surrounding the heart, which can be attempted automatically; since heart localisation is often used as a pre-processing step for many segmentation and registration algorithms, it has received much research focus as a distinct problem in itself [1, 55].

2.2.4 Image / Pixel classification methods

Pixel, or *Voxel* (i.e. "volumetric pixel") based methods predominantly rely on statistical or morphological properties of the image. For instance, the distinct appearance of different tissues in the image theoretically means that different tissues could have particular intensity profiles. In practice, however, it is difficult to separate the different heart components from the whole image from intensity characteristics alone. Nevertheless, pixel methods can be very versatile, as a variety of features and metrics can be used; furthermore, simple thresholding operations are often used as a preprocessing step for many algorithms, particularly since in Cine MRI the blood pool appears brighter than the surrounding tissues. In Cocosco *et al.* [55], for instance (which we make use of later in this thesis), the authors take advantage of voxel variability in time (since Cine MRI has multiple timeframes for the same volume — information which goes unused in most non voxel-based algorithms), to compute a ROI. Then a simple thresholding operation based on Otsu's method [33], followed by simple morphological operations, is used to select the two ventricles as the two connected components with the largest variability in volume throughout the timeseries.

Apart from intensity, other traditional voxel-based metrics such as the gradient, orientation, local phase, etc can be used in this context, as well as more general machine-learning techniques. In Pednekar et al. [17] for instance, a Hough transform of the motion map is used to identify the Left Ventricle from its circular shape in the basal slice; a Gaussian-Mixtures Expectation-Maximization (EM) model is used to define a statistical model of intensities, then propagate the classification in the remaining slices using fuzzy-affinity; then a polar transform from the LV centroid is

used as the substrate to a dynamic-programming-based segmentation by minimizing a cost function responsive to region homogeneity, gradient and spatial continuity.

While a training bias generally isn't a problem in the same sense that some of the supervised methods mentioned above would be biased towards pathology encompassed by the training set, care needs to be taken when designing a voxel-based algorithm, since relying on voxel features characteristic of healthy myocardium for segmentation might lead to segmentation failures if these features do not apply to pathological tissue. Infarcted tissue is particularly problematic, since it tends to exhibit great inhomogeneity and its location and appearance varies greatly [45].

2.2.5 Ensemble methods

The term 'Ensemble methods' refers to approaches aiming to make use of multiple components or sources of information, such that these are optimally combined into a single output in some way, which is then expected to perform better than any of its individual constituent parts in isolation.

In general, and even outside the narrow context of segmentation, it is recognised that there are certain characteristics that individuals in a 'crowd' must possess, for the 'crowd' to have a collectively better decision ability (the 'not all crowds are wise' phenomenon) [56]; namely there needs to be:

- Diversity of opinion, where each member has a personal – and relatively unique – approach to interpreting the information at hand
- Independence, where each member's opinion is not influenced by other members' opinions
- Decentralisation, where members draw conclusions on local knowledge (and in the context of algorithms relating to the concept of avoiding a central — i.e. systematic — bias).
- Aggregation, that is an effective, organic mechanism by which the individual judgements rise to a collective decision

With respect to segmentation, ensemble methods generally rely on one of, or a hybrid of, the following two complementary principles:

- they *either* attempt to combine a collection of weaker *classifiers* in an optimal manner, with the aim of creating a stronger *classifier*
- *or* they take a number of segmentation *outputs* which are already considered to be of 'good quality', and try to optimally 'fuse' this information into a single optimal *output*

Combining weaker classifiers into a stronger classifier

The main principle behind this class of algorithms is that there will generally exist an optimal combination of weaker classifiers, that will result in a stronger classifier; the intuition behind this claim is that, as long as the components are relatively independent, there will generally be a way to combine two or more classifiers (e.g. via a weighted average), such that their combined bias is generally closer to the ground truth compared to the bias of the individual components (on average), making the combined classifier more *accurate*, and such that their variances cancel each other out to an extent, due to the random and independent nature of their individual noise, making the combined classifier more *precise* [57]. Popular ensemble methods in this category includes algorithms such as Adaboost [58] and Random Forests [14].

However, like all training-based models, they rely on the quality of the training set; they also rely on arbitrary parameters whose effect cannot be predicted accurately in the absence of validation, and can easily lead to under- or over-fitting of the data. All the above algorithms combine segmentations at the classifier level, rather than at the voxel level, therefore subcomponents which are strong in some regions but weak in others will either be up-voted or down-voted completely, negating some of the expected benefits of fusion.

Label fusion from 'presumed good-quality' outputs

The focus of this class of algorithms tends to be to produce a better estimate of the ground truth from a (potentially relatively small) collection of gold standards

— i.e. segmentations which are already assumed to be fairly close to the ground truth. In other words, the aim is to reach some form of 'consensus among expert opinions'. Each 'expert opinion' may be considered equally valued, or priors may be attributed (locally or otherwise).

Popular algorithms in this category includes the 'Simultaneous Truth and Performance Level Estimation' (STAPLE) algorithm [32], and its many variants (e.g. SoftSTAPLE [59], a probabilistic analogue; STEPS [11], a variant for registration-based segmentations; Continuous STAPLE [60] which adapts STAPLE to vector-based images; MAP STAPLE [61], which additionally incorporates prior information on the contributing experts, etc).

Implicit in the use of such algorithms is the assumption that the inputs are, for the most part, all of relatively 'high quality' to begin with. While this doesn't preclude them from being used as a segmentation refinement approach in itself, it does imply that using it with weaker, or biased classifiers would in fact be against the algorithm's assumptions and perhaps not necessarily as robust.

Other methods — multi-atlas segmentation

The two approaches described above represent the two complementary, 'canonical' routes for attempting fusion, i.e. at the *classifier*, and at the *output* level; but naturally, it is possible to have methods that form a conceptual hybrid between the two principles, or where the distinction between the two 'routes' is blurry. An interesting and prominent example of such a class of algorithms is *multi-atlas segmentation* (MAS) *with label fusion* — see [62] for a recent survey on this class of algorithms. This is essentially an ensemble approach to the simple atlas-based segmentation case mentioned above, and generally involves a collection of 'intensity-atlas / label-atlas' pairs, where, generally speaking:

- each atlas is registered to the target image
- a measure of registration quality is obtained for each atlas
- the labels are propagated to form a collection of segmentation outputs

- the outputs are fused appropriately, effectively reaching a 'consensus' while taking into account each component's registration quality as 'prior' information.

One could perceive this process as a case of combining weaker classifiers into a stronger one, if one thinks of each atlas pair as a conceptual 'algorithmic' component, whose quality can be evaluated (e.g. by assessing, in this case, the registration quality of the intensity-atlas, as evaluated by a suitable image-similarity metric), and thus combined with the other atlases in a sophisticated 'weighting' scheme, producing a stronger overall classifier. *Or*, one could also perceive this as fusing a collection of 'outputs' obtained from the individual atlases, where these are assumed to be of good quality and representative of the ground truth, and where the accompanying measure of quality for each atlas serves as additional 'prior' information regarding the reliability of the process that generated each output.

The overall accuracy in this particular class algorithms is therefore a function of *both* the quality of the registered components, *and* the quality of fusion. Bai *et al.* empirically point out in [63], that registration accuracy tends to have a large impact on segmentation performance, such that in the presence of accurate and consistent components, the difference between various label fusion strategies, including simple majority voting, is very subtle; whereas in the presence of less accurate components, for example when simple affine registration is used, more sophisticated label fusion strategies are required to improve segmentation performance. One advantage of the MAS approach is that one can evaluate measures of registration quality at the 'local' level, meaning that the fusion process could then even be performed even at the 'pixel' level. Furthermore, the presence of two independent 'routes' of optimisation naturally lends itself to 'refinement' approaches, where optimisation of one approach can inform the other in an iterative manner until some sort of convergence is reached.

2.3 State of the art in Validation methods

Research on validation tends to lag behind segmentation research [64, 65]. This is a big problem, since it often results in a relative inability for modern segmentation methods to be evaluated appropriately, unless certain simplifying assumptions are made to allow their evaluation under older validation frameworks.

A detailed discussion on the appropriate validation of modern segmentation algorithms, with a particular focus on probabilistic / fuzzy algorithms against corresponding gold-standards, and proposed novel methods to improve on this area, is one of the main contributions of this thesis, therefore state of the art in validation is discussed separately in Chapter 4, under the broader thematic unit of validation in medical image segmentation.

2.4 Mathematical background

Throughout this thesis we make use of theoretical notions from the fields of *probability*, *fuzzy set theory* and *fuzzy logic*. In this section we introduce the main mathematical definitions and concepts used throughout the thesis, and in particular we introduce the mathematical concept of Triangular Norms and Triangular Conorms, which are fuzzy generalisations of the binary intersection and union operators respectively.

We first discuss standard set theory, and how this translates into classical logic and binary mask operations in the context of images, so as to introduce the basic concepts and establish some notational formalisms that we will carry forward to the probabilistic and fuzzy cases.

2.4.1 Classical sets and binary masks

Standard set notation

Assume two ordinary sets A and B, such that A contains m elements (i.e. A has cardinality $|A| = m$), and B contains n elements (i.e. $|B| = n$). Furthermore, we define a parent set Ω, with cardinality $|\Omega| = k$, such that both A and B are subsets of Ω (i.e. $A \subseteq \Omega,\ B \subseteq \Omega,\ m \leqslant k,\ n \leqslant k$). Standard set theory defines the following valid set operations [66]:

- The **intersection** of two sets A and B, denoted $A \cap B$ — or equivalently $\cap(A, B)$ in functional form — refers to the set which consists of only the elements common to both A and B. For brevity and notational convenience, we may refer to both the operator and the resulting set as $\bigcap$ in the text, when the particular meaning and input sets are clear from the context.

- Similarly, the **union** of two sets A and B, denoted $A \cup B$ — or $\cup(A, B)$ in functional form — refers to the set which consists of only the elements contained in both A or in B, but without repetition. Similar to the $\bigcap$ operator, we may refer to either the operator or the resulting set as $\bigcup$ depending on the context.

- The **symmetric difference** (also known as *disjunctive union*) of two sets A and B, denoted $A \triangle B$, is the set of elements that are contained *either* in A *or* B, but *not* both, such that $|A \cap B| + |A \triangle B| = |A \cup B|$

- The **complement** of a set A (with respect to a parent set Ω), denoted A^c, is the set of all elements in Ω that are not found in A. In other words, A and A^c are *mutually exclusive and exhaustive*, such that $|A| + |A^c| = |\Omega|$.

The $\bigcup$ and $\bigcap$ operations in particular, are considered *dual* to each other (i.e., conceptually opposite / complementary operations), linked together mathematically

via *De Morgan's Laws*:

$$\begin{aligned} A \cup B &= (\ A^c \cap B^c\ )^c \\ A \cap B &= (\ A^c \cup B^c\ )^c \end{aligned} \tag{2.1}$$

All the above standard set operations can be combined to form more complex expressions in general. For example, the *Tanimoto* coefficient [67], a coefficient frequently used in image analysis to quantify the match between two segmentations, is a quantity defined as:

$$T_c = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \tag{2.2}$$

The *Dice* coefficient [68], another common coefficient used in the same manner, is a quantity defined as:

$$D_c = \frac{\left|A \cap B\right|}{\left(\frac{|A|+|B|}{2}\right)} \tag{2.3}$$

In chapter 4 we will be demonstrating how quantities that apply to standard sets, and which are defined via pure standard set operations like the above coefficients, can be generalised to fuzzy sets and fuzzy set operators for the purpose of more informative validation, adapted to probabilistic and fuzzy sets. We will be focusing on the Tanimoto coefficient, as this is more straightforward to use; however, the concepts we will discuss apply equally to the Dice coefficient, or any other set-based operator — indeed, these all tend to be largely interrelated anyway; for instance the Dice coefficient is related to the Tanimoto Coefficient via:

$$T_c = D_c/(2 - D_c) \qquad \left(\ \therefore \quad D_c = 2T_c/(1 + T_c)\ \right)$$

Classical sets represented as binary vectors

We can represent any set and its subsets in a '*Vector of Bits*' (or simply 'binary vector') notation; e.g. if a parent set Ω consists of elements $\Omega = \{x_1, x_2, \ldots, x_{10}\}$, we can represent this as a 10-element vector $\boldsymbol{\omega} = (1,1,1,1,1,1,1,1,1,1)$, where

each position in the vector corresponds to a particular element in Ω, and the boolean value at that position indicates whether that particular element is *present* ('1') or *absent* ('0') in the set. For the purposes of this thesis, we notate this transformational equivalence as a mapping from a set to a vector, i.e. $\Omega \mapsto \boldsymbol{\omega}$. This notation allows us to adequately represent subsets from a known parent set in suitable vector form, by simply 'flipping' the appropriate *bits* 'on' or 'off'. Note that a big difference between the set and vector constructs is that sets may or may not necessarily be inherently ordered — i.e., if there is no defined hierarchy between elements in a set, then two sets such as $\{x_1, x_2, x_3\}$ and $\{x_3, x_1, x_2\}$ would be considered equivalent; whereas the vector definition dictates a strict, pre-defined order in its elements. This is essential since the index of each vector element must always map to a specific element in the set, regardless of whether the underlying set is inherently considered to be ordered or not.

As an example, consider sets $A = \{x_3, x_4, x_5\}$ and $B = \{x_4, x_5, x_6, x_7\}$, where $A, B \subseteq \Omega$, for which we can obtain the following mappings:

$$\begin{array}{rll} A &= \{x_3, x_4, x_5\} & \mapsto \quad \mathbf{a} = (0,0,1,1,1,0,0,0,0,0) \\ B &= \{x_4, x_5, x_6, x_7\} & \mapsto \quad \mathbf{b} = (0,0,0,1,1,1,1,0,0,0) \\ & & \\ A \cap B &= \{x_4, x_5\} & \mapsto \quad (0,0,0,1,1,0,0,0,0,0) \\ A \cup B &= \{x_3, x_4, x_5, x_6, x_7\} & \mapsto \quad (0,0,1,1,1,1,1,0,0,0) \\ A \triangle B &= \{x_3, x_6, x_7\} & \mapsto \quad (0,0,1,0,0,1,1,0,0,0) \\ A^c &= \{x_1, x_2, x_6, x_7, x_8, x_9, x_{10}\} & \mapsto \quad (1,1,0,0,0,1,1,1,1,1) \\ B^c &= \{x_1, x_2, x_3, x_8, x_9, x_{10}\} & \mapsto \quad (1,1,1,0,0,0,0,1,1,1) \end{array} \tag{2.4}$$

In general, any set Γ resulting from a set operation on A and / or B, can be expressed as a vector via the following generator syntax:

$$\Gamma \mapsto \ \boldsymbol{\gamma} \equiv \ \forall i : \begin{cases} \gamma_i = 1, & \text{for } x_i \in \Gamma \\ \gamma_i = 0, & \text{for } x_i \notin \Gamma \end{cases} \tag{2.5}$$

The latter part of this equation is often called the *indicator function* (or *characteristic function*) of Γ, denoted $\mathbf{1}_\Gamma$; it is a mapping from any valid element x_i of the parent set Ω, to the binary set $\{0, 1\}$, such that:

$$\mathbf{1}_\Gamma(x_i) = \begin{cases} 1, & \text{for } x_i \in \Gamma \\ 0, & \text{for } x_i \notin \Gamma \end{cases} \tag{2.6}$$

Eq. 2.5 can therefore be equivalently rewritten as:

$$\Gamma \mapsto \boldsymbol{\gamma} \equiv \forall i : \gamma_i = \mathbf{1}_{\Gamma}(x_i) \tag{2.7}$$

We note in particular that $\Omega \mapsto \boldsymbol{\omega} \equiv \mathbf{1}_{\Omega}$

Correspondence of standard set and boolean logic operators

In terms of classical (i.e. boolean) logic, variables, e.g. $\alpha, \beta \in \{0, 1\}$, represent atomic *propositions*, i.e. simple, self-contained statements which may either evaluate to *True* or *False*. More complex statements (i.e. 'arguments') can be constructed from such simpler statements, by combining them together using logical connectives in a structured manner, such that the truth value of such a compound statement can be evaluated as a whole. Classical logic defines the following logical operations [66]:

- **conjunction**, implemented via the logical '*and*' connective (denoted $\alpha \wedge \beta$): represents the truth value of the statement "both α and β are *True*".

- **disjunction**, implemented via the logical '*or*' connective (denoted $\alpha \vee \beta$): represents the truth value of the statement "one of α or β (or both) are *True*".

- **exclusive disjunction**, implemented via the logical '*xor*' connective (denoted $\alpha \veebar \beta$); represents the truth value of the statement "either α or β (but not both) are *True*".

- **negation**, implemented via the logical '*not*' operator (denoted $\neg\alpha$); represents the truth value of the statement "α is *False*"

Their formal definitions are as follows:

$$\begin{aligned}
\alpha \wedge \beta &= \begin{cases} 1, & \text{for } \alpha = 1 \text{ and } \beta = 1 \\ 0, & \text{otherwise} \end{cases} \\
\alpha \vee \beta &= \begin{cases} 0, & \text{for } \alpha = 0 \text{ and } \beta = 0 \\ 1, & \text{otherwise} \end{cases} \\
\alpha \veebar \beta &= \begin{cases} 1, & \text{for } \alpha \neq \beta \\ 0, & \text{otherwise} \end{cases} \\
\neg\alpha &= \begin{cases} 1, & \text{for } \alpha = 0 \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{2.8}$$

It is straightforward to see, that the standard set operations of *intersection*, *union*, *symmetric difference* and *complementation*, can be straightforwardly implemented on their boolean vector equivalents, via the logical operations for *conjunction*, *disjunction*, *exclusive disjunction*, and *negation* respectively. Since individual positions in vectors represent distinct set elements / propositions, the above logical connectives can be applied to whole boolean vectors in an elementwise fashion directly, such that the vectors from eq. 2.4 can be straightforwardly expressed as follows:

$$\begin{array}{rcccl}
A \cap B & \mapsto & \mathbf{a} \wedge \mathbf{b} & \text{such that } \forall i: & (\mathbf{a} \wedge \mathbf{b})_i = a_i \wedge b_i \\
A \cup B & \mapsto & \mathbf{a} \vee \mathbf{b} & \text{such that } \forall i: & (\mathbf{a} \vee \mathbf{b})_i = a_i \vee b_i \\
A \triangle B & \mapsto & \mathbf{a} \veebar \mathbf{b} & \text{such that } \forall i: & (\mathbf{a} \veebar \mathbf{b})_i = a_i \veebar b_i \\
A^c & \mapsto & \neg\mathbf{a} & \text{such that } \forall i: & (\neg\mathbf{a})_i = \neg a_i \\
B^c & \mapsto & \neg\mathbf{b} & \text{such that } \forall i: & (\neg\mathbf{b})_i = \neg b_i
\end{array} \tag{2.9}$$

Furthermore, the cardinality of any set expressed in vector form, is simply the sum of the boolean values of all the elements in the vector (i.e. it's ℓ_1 norm), e.g.:

$$\left|A \cap B\right| = \|\mathbf{a} \wedge \mathbf{b}\|_1 = \sum_i a_i \wedge b_i \tag{2.10}$$

Masks are sets of pixels, represented as binary arrays

Any quantized 2D image I of resolution $M \times N$, is essentially a mapping from a set $\Omega = \{x_{ij}, \ldots, x_{MN}\}$, whose elements x_{ij} are the individual pixels that make up the *image domain*, to a real-valued domain, such that each pixel is associated with an 'intensity' value (or any other scalar or vector-valued metric in general).

Segmentation — i.e. the act of attributing to each pixel element in this image set one of two or more discrete labels — is therefore equivalent to creating two or more subsets of Ω which are *mutually exclusive and exhaustive*. For example, in the 2-class labelling problem (i.e. labelling each pixel as 'foreground' or 'background') this is equivalent to creating a *foreground* subset F consisting of all the pixels that relate to the object of interest, and a *background* subset B consisting of all the 'background' pixels, such that:

$$\begin{aligned} &F \subseteq \Omega \\ &B \subseteq \Omega, \quad \text{and} \\ &F \cap B = \emptyset, \\ &F \cup B = \Omega \end{aligned}$$

As per the vector representation defined above, we can represent F and B in 'array of bits' form, such that:

$$\begin{aligned} \Omega \mapsto \boldsymbol{\omega} &\equiv \forall i,j \;:\; \omega_{ij} = 1 \\ F \mapsto \mathbf{f} &\equiv \forall i,j \;:\; \begin{cases} f_{ij} = 1, & \text{for } x_{ij} \in F \\ f_{ij} = 0, & \text{for } x_{ij} \notin F \end{cases} \\ B \mapsto \mathbf{b} &\equiv \forall i,j \;:\; \begin{cases} b_{ij} = 1, & \text{for } x_{ij} \in B \\ b_{ij} = 0, & \text{for } x_{ij} \notin B \end{cases} \end{aligned}$$

We call any such binary array a *mask*. In the specific case of 2D images, masks are a particularly convenient way to visualise, or visually compare segmentation objects, by interpreting the mask as a black-and-white image, where black pixels correspond to positions with a value of '0' in the mask, and white pixels correspond to positions with a value of '1'. However, the above concepts can be extended naturally to "images" of higher dimensionality.

Fig 2.2 demonstrates how a segmentation S and a gold standard G defined over an image domain Ω can be inspected visually via their respective masks $\mathbf{s}$ and $\mathbf{g}$ in this manner, as well as the nature of their overlap and how this relates to the corresponding intersection ($\mathbf{s} \wedge \mathbf{g}$) and union ($\mathbf{s} \vee \mathbf{g}$) masks derived from the two binary inputs.

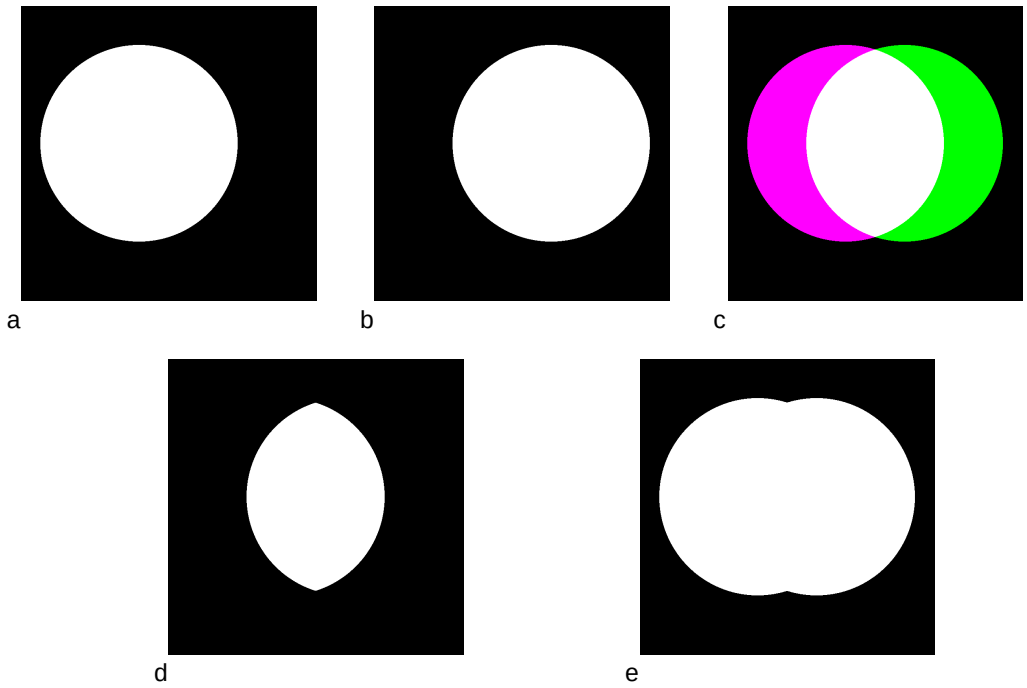


Figure 2.2: Binary segmentation masks: a) The segmentation candidate S ($\mapsto \mathbf{s}$). b) The gold standard G ($\mapsto \mathbf{g}$). c) Fused RGB colour-image of $\mathbf{s}$ (red and blue channels) and $\mathbf{g}$ (green channel), enabling visualisation of S and G on a single image. Each pixel is represented as an RGB triplet; the presence of red and blue channels *only* (i.e. the RGB triplet [R=1, G=0, B=1]) yield the colour purple on screen, whereas the simultaneous presence of red, green, and blue channels together (i.e. the RGB triplet [R=1, G=1, B=1]) yields the colour white. d) Mask $\mathbf{s} \wedge \mathbf{g}$ corresponding to the *intersection* of S and G (i.e. $S \cap G$). e) Mask $\mathbf{s} \vee \mathbf{g}$ corresponding to the *union* of S and G (i.e. $S \cup G$)

Notation: Segmentations, Gold-Standards, and Latent Truth

For consistency, we establish the following notation, which we will use throughout the thesis:

Ω — The *image domain*, i.e. the parent set of all pixels that comprise the image. Ω comes to represent such a discrete set of image pixels, as a direct result of the digitization process; the *number* of pixels contained in the image 'set' (i.e. $|\Omega|$), is a direct result of the resolution associated with the particular imaging method. All subsequent objects defined within the image domain are essentially strict subsets of Ω

I — The *image* itself, i.e. the actual data assigned to each pixel in Ω. Formally, it

is expressed as a mapping from the image domain to the real number domain, i.e. $I : \Omega \mapsto \mathbb{R}^D$, where D denotes the dimensionality of each datapoint. The most common scenario is when each pixel is associated with a particular *intensity*, which is a scalar quantity (i.e. $D = 1$) in the range [0,1], such that a '0' represents a dark (i.e. black) pixel, '1' represents a fully bright (i.e. white) pixel, and in-between values denote different points in the grayscale spectrum.

L — The *latent truth*. This can refer to either the *absolute ground truth* (i.e. the object or tissue in real-world terms), or the *relative ground truth*, subject to the limitations imposed by the digitization process (i.e. the most appropriate representation of the *absolute* ground truth, when expressed as a subset of Ω). We intentionally prefer the term 'latent' truth, over the term 'ground truth' in this thesis, for two reasons:

1. the term 'latent' truth makes explicit that fact that the ground truth, whether we refer to it in absolute or relative terms, is essentially an *unknown*, i.e. a hidden ("latent") variable, whose absolute nature cannot be observed or ascertained experimentally, but can only be implied or approximated through surrogate measures that *can* be acquired and observed instead, (such as the gold standard, or corroborating clinical variables).

2. The terms "ground truth" and "gold-standard" have occasionally been used interchangeably in the segmentation literature even though they technically refer to completely different concepts, as in some contexts, the latter is assumed to be a faithful representation of the former, and in such contexts any discrepancy between the two tends to be beyond the scope of that work. However, in this thesis, the distinction between the two is significant, particularly in the context of fuzzy validation, discussed in detail in Chapter 4. Therefore to avoid confusion, we use the term *latent truth* exclusively in this work instead.

When it is important to distinguish specifically between latent truth in absolute terms, and latent truth expressed in terms of an image domain Ω, we qualify these as L_∞ and L_Ω respectively (alluding to the fact that 'reality' can be thought of as the theoretical limit of an image domain of infinite resolution).

G — The *gold standard*; it is a segmentation object, which corresponds to the best method / approach in one's arsenal for approximating the latent truth L, subject to the limitations posed by Ω, as well as any other limitations, intrinsic for instance to the fact that it is a segmentation object (e.g. imposed deterministic definition, adherence to a particular segmentation protocol, etc). As G is typically expressed in terms of, and defined as a subset of Ω, it is specifically an approximation to L_Ω.

S — The *segmentation candidate*, i.e. a set of pixels thought to belong to the object or tissue of interest, as determined by a particular segmentation approach, typically an automated or semi-automated segmentation algorithm. For convenience, when this is clear from the context, we will refer to the segmentation candidate S simply as "the segmentation". The fact that both G and S are expressed in terms of Ω allows the two to be compared very straightforwardly for evaluation purposes, via standard set operations as described above.

Similarly, the array transformations for Ω, $L_{(\Omega)}$, G and S, (i.e. their corresponding 'masks') will be denoted by $\boldsymbol{\omega}$, $\mathbf{l}$, $\mathbf{g}$ and $\mathbf{s}$ respectively.

'Segmentation versus gold standard' binary classification components

One way to express the disparity between a segmentation object and a gold-standard object is to draw up a *classification matrix* (also known as a *confusion matrix*), like the one in table 2.1 below[3].

[3]Note that in this particular context we refer to the foreground and background labels of the gold standard as "*True*" and "*False*"; however, this is for notational consistency with standard classification literature only, and should not be confused with the notion of the *latent truth* (L) as differentiated in this thesis.

	Present (i.e. G)	Absent (i.e. G^c)
Positive (i.e. S)	$T_+ \ (= S \cap G)$	$F_+ \ (= S \cap G^c)$
Negative (i.e. S^c)	$F_- \ (= S^c \cap G)$	$T_- \ (= S^c \cap G^c)$

Table 2.1: Classification matrix treating the segmentation S as the *test*, and the gold standard G as the *baseline*, giving rise to the following classification components: True Positives (T_+), False Positives (F_+), False Negatives (F_-), and True Negatives (T_-).

Under this scheme, the segmentation set S acts as the "test", such that elements in S count as a "positive" result, and elements not in S as a "negative" result; similarly the gold-standard set G acts as a "baseline", such that elements in G signify the true "presence" of the quantity or quality being tested for, and elements not in G signify the "absence" of such a quantity or quality. With this framework in mind, the following *classification components* can be described:

True Positives (T_+): the set of pixels where the true object is deemed to be present (as per G), and the segmentation has "correctly" (as far as G is concerned) classified them as such (i.e. the test returned a "positive" outcome):

$$\begin{aligned} T_+ &\equiv S \cap G && \text{(in standard set notation)}, \\ \mathbf{t}_+ &= \mathbf{s} \wedge \mathbf{g} && \text{(in binary mask notation)}. \\ \textit{and} \quad |T_+| &= \left\| \mathbf{s} \wedge \mathbf{g} \right\|_1 && \text{(cardinality, as per eq. 2.10)} \end{aligned} \tag{2.11}$$

This component corresponds to the white area in fig. 2.2c.

False Positives (F_+): the set of pixels where the test *incorrectly* returned a *positive* outcome[4], i.e.:

$$\begin{aligned} F_+ &\equiv S \cap G^c, \\ \mathbf{f}_+ &= \mathbf{s} \wedge \neg\mathbf{g}, \\ |F_+| &= \left\| \mathbf{s} \wedge \neg\mathbf{g} \right\|_1 \end{aligned} \tag{2.12}$$

It corresponds to the violet area in fig. 2.2c.

False Negatives (F_-): the set of pixels where the test *incorrectly* returned a

[4] An alternative formulation for this is: "*the set of 'positive' results as a whole, minus the part that is known to be 'true'*", i.e. $F_+ = S - T_+$

negative outcome[5], i.e.:

$$\begin{aligned} F_- &\equiv \neg S \cap G, \\ \mathbf{f}_+ &= \neg\mathbf{s} \wedge \mathbf{g}, \\ |F_+| &= \left\|\neg\mathbf{s} \wedge \mathbf{g}\right\|_1 \end{aligned} \tag{2.13}$$

It corresponds to the green area in fig. 2.2c.

True Negatives (T_-): the set of pixels where the test *correctly* returned a *negative* outcome[6], i.e.:

$$\begin{aligned} T_- &\equiv S^c \cap G^c, \\ \mathbf{t}_- &= \neg\mathbf{s} \wedge \neg\mathbf{g}, \\ |T_-| &= \left\|\neg\mathbf{s} \wedge \neg\mathbf{g}\right\|_1 \end{aligned} \tag{2.14}$$

It corresponds to the black area in fig. 2.2c.

Many compound operators can be expressed using these classification components instead of the usual set operations. For example, the Tanimoto coefficient (eq. 2.2, p. 30) can be expressed using classification components as:

$$\frac{\left|T_+\right|}{\left|T_+\right| + \left|F_+\right| + \left|F_-\right|} \tag{2.15}$$

(for a more extensive list of similar examples see table 4.1 on p. 111)

Intersection and union operations on binary masks can have many equivalent implementations

In practice, when it comes to binary masks, from a mathematical point of view there are many ways to implement the logical operators described in eqs. 2.8 and 2.9 (p. 33). For example, the conjunction operator could be algorithmically implemented as follows:

$$\mathbf{a} \wedge \mathbf{b} \quad = \begin{cases} \min(\mathbf{a}, \mathbf{b}) & i.e.,\ \forall i : a_i \wedge b_i = \min(a_i, b_i) \\ \quad or & \\ \mathbf{a} \circ \mathbf{b} & i.e.,\ \forall i : a_i \wedge b_i = a_i b_i \\ \quad or & \\ \max(0, \mathbf{a} + \mathbf{b} - 1) & i.e.,\ \forall i : a_i \wedge b_i = \max(0, a_i + b_i - 1) \end{cases} \tag{2.16}$$

[5] An alternative formulation for this is: "*the set of 'true' results as a whole, minus the part that was correctly identified as 'positive'*", i.e. $F_- = G - T_+$

[6] An alternative formulation for this is: "*the part of the domain that was identified as neither true nor positive*", i.e. $T_- = \Omega - T_+ - F_+ - F_-$

Equally, the disjunction operator could be implemented as follows:

$$\mathbf{a} \vee \mathbf{b} \quad = \begin{cases} \max(\mathbf{a}, \mathbf{b}) & i.e.,\ \forall i : a_i \vee b_i = \max(a_i, b_i) \\ \quad or & \\ \mathbf{a} + \mathbf{b} - \mathbf{a} \circ \mathbf{b} & i.e.,\ \forall i : a_i \vee b_i = a_i + b_i - a_i b_i \\ \quad or & \\ \min(\mathbf{a} + \mathbf{b}, 1) & i.e.,\ \forall i : a_i \vee b_i = \min(a_i + b_i, 1) \end{cases} \tag{2.17}$$

Note that each respective conjunction / disjunction implementation pair between eqs. 2.16 and 2.17 forms a *dual* pair — i.e. consistent with De Morgan's Laws (eq. 2.1, p. 30).

While these three conjunction implementations — and similarly the three disjunction implementations — are not generally 'mathematically' equivalent operations for *any* real inputs $\mathbf{a}$ and $\mathbf{b}$, nevertheless, specifically for *binary mask* inputs they are effectively equivalent, and will yield the appropriate conjunction and disjunction results respectively, compatible with the definitions outlined in eq. 2.8 (p. 33).

2.4.2 Fuzzy sets and fuzzy masks

What is a "fuzzy set"?

Fuzzy set theory was formalised by Lotfi A. Zadeh in 1965, via his seminal work entitled "Fuzzy Sets" [69], where he introduced the concept of *fuzziness* as the notion of *partial* or *ambiguous* membership of an element with respect to a set. The need for such fuzzy sets arose from the fact that, in contrast to classical sets — or *crisp* sets as they are called in this context — many real-world objects and concepts cannot be uniquely and unambiguously classified into clear, distinct categories. In fuzzy theory, this traditionally reflects qualitative reasons, such as an ambiguous or vague definition for a class, or the potentially ambiguous or imprecise nature of some of the elements under consideration. For example, if the task is to classify images of animals versus plants, while it might be straightforward to classify a horse and a rose into their respective categories, it may not be as straightforward to classify an image of a starfish, or one of a bacterium, without first considering the question "to what extent is a starfish or a bacterium an animal; to what extent is it a plant?". Similarly, if one is classifying images of the brain into "healthy versus

Alzheimer's disease" based on the radiological features of the disease, while the extremes of health and advanced disease displaying all the hallmarks of Alzheimer's disease may be readily classified as such, intermediate stages of the disease may not be as easy to classify or definitively differentiated from "normality" — indeed, this is one of the reasons why Alzheimer's disease is predominantly a 'clinical' diagnosis rather than a purely radiological one [70].

In other cases, the source of fuzziness can be down to more quantitative reasons: in the context of segmentation for instance, where the classification essentially involves individual pixels, classification ambiguity arises as a result of the quantisation process. In other words, rather than ask "Which tissue does this pixel correspond to?" (i.e. with respect to a particular criterion), we are forced to ask "To which *tissues* does this pixel correspond to, and to what *extent*?" (we will be discussing the concept of *partial volume* in more detail in chapter 3).

Fuzzy theory attempts to transform such ambiguous and (functionally) imprecise terms, which are hard to process and evaluate, into more numerically useful expressions relating to more quantifiable notions of uncertainty; It does so by translating such expressions in terms of their degree of membership to a particular *fuzzy* set (i.e. a set with non-sharply defined boundaries), *or* by ascribing a degree of belief or evidence that an element is a member of a particular *crisp* set (i.e. a set with sharply defined boundaries). In the latter case, such measures of uncertainty are more specifically referred to as *fuzzy measures* [71]). Note that fuzziness, in this context, is a much broader concept than that of *uncertainty* as is typically used in the particular context of probability theory; we will be discussing this in some more detail in the next chapter.

Formal fuzzy set definition and notational conventions

Let Ω represent a parent set containing N elements, e.g. the domain $\{x_1, \ldots, x_N\}$. A fuzzy set Γ is any strict subset of Ω, such that any element x_i is neither necessarily fully present, or fully absent from Γ, but one can talk about the *extent* to which a

particular element is considered to be a member of Γ. Such a fuzzy set is defined in terms of Ω as a mapping from each element x_i in Ω to a *fuzzy value*, i.e. a value in the range [0,1], denoting the degree of *membership* of that element in the fuzzy set, where '0' means "definitely absent" and '1' means "definitely present". This mapping is known as the *membership function* of Γ, denoted μ_Γ; the membership function is a generalisation of the 'indicator function' (see eq. 2.6, p. 31), which is a property of classical sets. Formally:

$$\forall i, x_i \in \Omega;\ \mu_\Gamma : x_i \mapsto [0, 1] \tag{2.18}$$

such that, in terms of generator syntax, the following set of fuzzy mappings applies:

$$\Gamma \equiv\ \forall x_i \in \Omega :\ \left\{\ x_i \mapsto \gamma_i\ \middle|\ \gamma_i = \mu_\Gamma(x_i)\ \right\} \tag{2.19}$$

From fuzzy sets to fuzzy masks

We can see from eq. 2.19 that fuzzy sets extend naturally to the vector / mask representation discussed in section 2.4.1 (p. 33), by generalising eq. 2.7 (p. 32) to vectors / masks defined with fuzzy, rather than boolean values, i.e.:

$$\Gamma \mapsto\ \boldsymbol{\gamma} \equiv\ \forall i : \gamma_i = \mu_\Gamma(x_i) \tag{2.20}$$

Furthermore, in the same way binary masks were intuitively visualised as black-and-white images (fig. 2.2) fuzzy masks can be equally intuitively visualised as grayscale images, where the value range $[0, 1]$ becomes a gray value in the grayscale spectrum, such that '0' corresponds to black, '1' to white, and inbetween values to corresponding shades of gray[7]. Fig. 2.3 shows an example of such a fuzzy mask, corresponding to the same objects as fig. 2.2 (p. 35), but with somewhat more ambiguously defined borders.

[7]Note however that this still represents an object *mask*, rather than modality-dependent object *intensities*, which also tend to be represented using grayscale images, but have a very different interpretation.

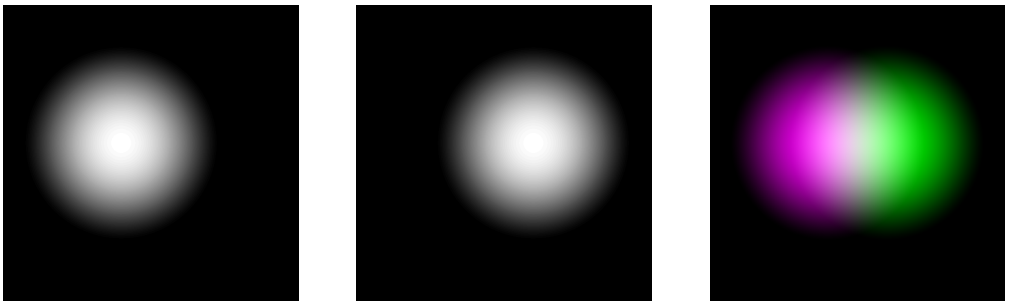


Figure 2.3: Fuzzy segmentation masks: a) The segmentation candidate S ($\mapsto \mathbf{s}$). b) The gold standard G ($\mapsto \mathbf{g}$). c) Fused RGB colour-image of $\mathbf{s}$ and $\mathbf{g}$ (as per fig. 2.2c)

Triangular Norms and Conorms: fuzzy generalisations of intersection and union operators

In section 2.4.1 (p. 39) we talked about how *intersection* and *union* operators can be implemented in a variety of different ways, yielding identical results in binary masks; eqs. 2.16 and 2.17 (p. 39) demonstrated three such intersection and union implementations respectively. Moving on from binary to fuzzy masks, however, we see that these implementations, apart from the extreme values of '0' and '1', are not otherwise equivalent for in-between values. Therefore if we implemented these as interpretations of the intersection and union operations respectively for *fuzzy* masks, we would obtain a different fuzzy mask output for each implementation; fig. 2.4 demonstrates graphically the difference between the three implementations (compare with figs 2.2 (p. 35) and 2.3).

In general, for fuzzy sets, there is no unique intersection or union operation, but rather there exists a family of intersection and union operators, known as *triangular norms* (abbreviated *t-norms*) and *triangular conorms* (abbreviated *t-conorms*, or also *s-norms*) respectively [72].

In other words, t-norms and t-conorms are generalisations of the intersection ($\bigcap$) and union ($\bigcup$) operators in Fuzzy Set Theory; they are defined as follows:

A **t-norm** is an operator / function, which:

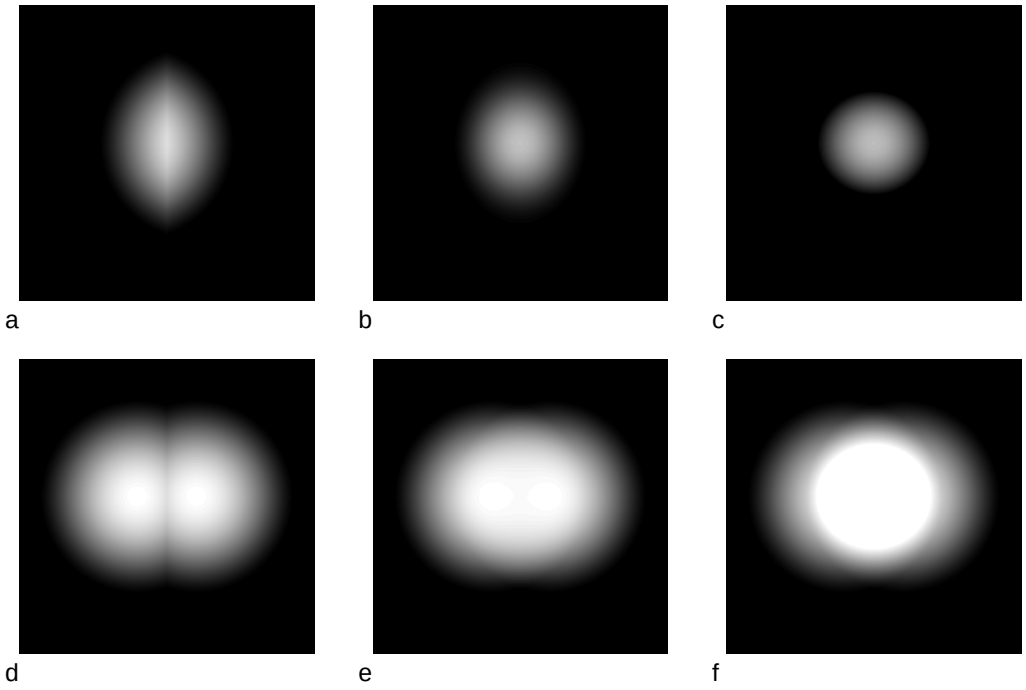


Figure 2.4: Intersection and union implementations for the fuzzy masks in fig. 2.3:
a) Intersection implemented as $\min(\mathbf{s}, \mathbf{g})$. b) Intersection implemented as $\mathbf{s} \circ \mathbf{g}$.
c) Intersection implemented as $\max(0, \mathbf{s} + \mathbf{g} - 1)$. d) Union implemented as $\max(\mathbf{s}, \mathbf{g})$.
e) Union implemented as $\mathbf{s}+\mathbf{g}-\mathbf{s}\circ\mathbf{g}$. f) Union implemented as $\min(\mathbf{s}+\mathbf{g}, 1)$.

Note: '$\circ$' denotes the 'Hadamard product' (i.e. entry-wise array multiplication).

- takes two fuzzy inputs and returns a fuzzy output,
- is commutative: (i.e. $A \cap B \equiv B \cap A$),
- is associative: (i.e. $A \cap (B \cap C) \equiv (A \cap B) \cap C$),
- is monotonically nondecreasing with respect to increasing inputs,
- treats 0 and 1 as null and unit elements respectively (i.e. $A \cap 0 \equiv 0,\ A \cap 1 \equiv A$).

A **t-conorm** is an operator / function, which

- takes two fuzzy inputs and returns a fuzzy output,
- is commutative: (i.e. $A \cup B \equiv B \cup A$),
- is associative: (i.e. $A \cup (B \cup C) \equiv (A \cup B) \cup C$),
- is monotonically nondecreasing with respect to increasing inputs,
- treats 1 and 0 as null and unit elements respectively (i.e. $A \cup 0 = A,\ A \cup 1 = 1$).

T-norms and t-conorms are *dual* operations, connected (like in the binary case) via *De Morgan's Laws* (see eq. 2.1, p. 30), where the *complement* Γ^c of a fuzzy set Γ is defined[8] as

$$\Gamma^c \quad \mapsto \quad \neg\boldsymbol{\gamma} \;= \mathbf{1}_\Omega - \boldsymbol{\gamma} \quad \equiv \quad \forall i: \; \neg\gamma_i \;= 1 - \gamma_i \tag{2.21}$$

Three special t-norms: Gödel, Product, and Łukasiewicz

The t-norms and t-conorms presented in eqs 2.16, 2.17 (p. 39) and fig. 2.4 (p. 44) are well-known in fuzzy theory and have special names:

The *Gödel* t-norm ($\bigcap_G$) and t-conorm ($\bigcup_G$):

$$\begin{array}{lcll} \mathbf{a} \cap_G \mathbf{b} & = & \min(\mathbf{a}, \mathbf{b}) & i.e.\ \forall i: a_i \cap_G b_i = \min(a_i, b_i), \\ \mathbf{a} \cup_G \mathbf{b} & = & \max(\mathbf{a}, \mathbf{b}) & i.e.\ \forall i: a_i \cup_G b_i = \max(a_i, b_i), \end{array} \tag{2.22}$$

The *Product* t-norm ($\bigcap_P$) and t-conorm ($\bigcup_P$):

$$\begin{array}{lcll} \mathbf{a} \cap_P \mathbf{b} & = & \mathbf{a} \circ \mathbf{b} & i.e.\ \forall i: a_i \cap_P b_i = a_i b_i, \\ \mathbf{a} \cup_P \mathbf{b} & = & \mathbf{a} + \mathbf{b} - \mathbf{a} \circ \mathbf{b} & i.e.\ \forall i: a_i \cup_P b_i = a_i + b_i - a_i b_i, \end{array} \tag{2.23}$$

The *Łukasiewicz* t-norm ($\bigcap_L$) and t-conorm ($\bigcup_L$):

$$\begin{array}{ll} \mathbf{a} \cap_L \mathbf{b} = \max(\mathbf{0}_\Omega,\ \mathbf{a} + \mathbf{b} - \mathbf{1}_\Omega) & i.e.\ \forall i: a_i \cap_L b_i = \max(0,\ a_i + b_i - 1), \\ \mathbf{a} \cup_L \mathbf{b} = \min(\mathbf{a} + \mathbf{b},\ \mathbf{1}_\Omega) & i.e.\ \forall i: a_i \cup_L b_i = \min(a_i + b_i,\ 1), \end{array}$$
$$(\text{where}: \mathbf{0}_\Omega = \neg\mathbf{1}_\Omega). \tag{2.24}$$

For brevity, when we refer to a named t-norm and t-conorm as a dual pair, we will refer to them simply as *norms*, i.e. 'the Gödel norms', 'Łukasiewicz norms' etc. We only make passing mention of these norms here, but will be discussing them in great detail in chapter 4 where we will explore their particular properties and semantics in the context of "fuzzy validation".

[8] while this is the most widely used definition, in theory there is a family of complementation operations, just like with t-norms and t-conorms, however these are of less relevance and beyond the scope of this thesis

2.4.3 Probability theory

Probability is a concept used in the study of *random variables*. A random variable X, is defined as a quantity whose value may vary each time it is measured, in a *random* (or, more specifically, a *stochastic*[9]) manner. In other words, there is uncertainty intrinsic to the variable itself, such that its value at any particular attempt to measure it can never be predicted with complete certainty, but at the same time conforming to a certain *distribution*, that is to say, some random outcomes are more likely than others. Such a distribution is referred to as the *probability distribution* of X.

Probability in this context is defined as a measure in the range [0,1], denoting a measure of frequency, proportionality or uncertainty over the outcome of a random variable, depending on the particular interpretation and context in which it is used (we discuss this distinction in somewhat more length in section 3.2.1). More formally, it is a function, mapping from a random event to the continuous range [0,1], where 0 denotes total uncertainty / lack of presence in a set, and 1 denotes full certainty / presence in a set.

When a particular measurement is performed, depending on the context, we say that the random variable X has "collapsed" onto, or has "generated" a particular value, or event. The usual notation for this is $p(X = x)$ — or simply $p(x)$ if the respective random variable X is implied from context — where small x denotes the particular value / event under consideration. However, more generally, provided the context is clear, most textbooks conventionally refer to probabilities over random variables and their distributions directly, rather than explicitly over specific events or values that the variable can take. Therefore the notations $p(X)$ and $p(X = x)$ tend to be used interchangeably when there is no space for confusion; when used in this way, a probability $p(X)$ is used to imply the validity of a particular formula with respect to *all* and *any* events that can occur in the context of the random

[9] from the Greek word meaning 'target' or 'bullseye', implying an event whose outcome cannot be pre-determined precisely, but can be described in terms of a particular accuracy / precision profile with respect to one or more targets.

variable X, and we often end up talking about random variables and the individual events / values they can take interchangeably.

The random variables studied may be of a continuous nature, in which case they are said to be described by *probability density functions* (*pdf*), or they can be discrete, in which case they are said to be described by a *probability mass function* (*pmf*). When dealing with continuous variables in particular, it is often more useful to also talk about a *pdf*'s corresponding *cumulative distribution function* (*cdf*), usually denoted as $P(X \leqslant x)$ or simply $P(X)$, which evaluates the probability that a random variable is *less than*, rather than exactly equal to a particular value, and can be evaluated via the integral of its corresponding *pdf* (or a summation in the case of a *pmf*).

Rules of probability

A valid probability measure must adhere to some fundamental rules and definitions, namely:

- **Joint probability:** The *joint* probability of two random variables A and B, denoted $p(A, B)$ is the probability that A and B occur jointly, i.e. the probability that their events co-occur. Formally:

$$p(A, B) = p(A \wedge B) \tag{2.25}$$

- **Probabilistic union:** The probability of *either* A or B occurring, can be defined in terms of the probabilities of the individual events and their joint probability:

$$\begin{array}{rcll} p(A \vee B) & = & p(A) + p(B) - p(A, B) & \\ p(A \vee B) & = & p(A) + p(B), & \text{(if } A \text{ and } B \text{ are } \textit{mutually exclusive} \text{ events)} \\ p(A \vee B) & = & 1 & \text{(if } A \text{ and } B \text{ are } \textit{mutually exhaustive} \text{ events)} \end{array} \tag{2.26}$$

- **Conditional Probability:** The probability of event A occurring given event B has occured, is given by:

$$p(A|B) = \frac{p(A, B)}{p(B)} \tag{2.27}$$

This is often rearranged as $p(A, B) = p(A|B)p(B)$; this particular formulation is also known as *the **product** rule of probability*.

- **Marginal Probability:** Given a joint probability $p(A, B)$, the probability of event A occurring irrespective of B, is given by:

$$p(A) = \sum_b \Big[p(A|B = b)p(B = b)\Big] \tag{2.28}$$

This process is often called *marginalisation*; the above rule is also known as *the **sum** rule of probability*.

The last two rules are also known as *Cox's axioms of probability*, named after Richard Threlkeld Cox who provided the first rigorous proof that probability theory could be regarded as an extension of Boolean logic to situations involving uncertainty. These axioms form the 'necessary and sufficient' conditions for probability measures to be applicable as meaningful measures of uncertainty [73].

Bayesian inference

A very important point which stems as a direct consequence of the product rule is the following: for any two random variables M and D, the following relation applies:

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)} \tag{2.29}$$

This rule is known as *Bayes' Rule*, or *Bayes' Theorem*. It is of crucial importance in the context of evaluating the probability of a model M given a dataset D (i.e. the suitability of M as a model in terms of how well it can explain / predict the data). It is often the case that the probability $p(M|D)$ is not easy to measure directly, but can be inferred from $p(D|M)$ which can be more straightforward to measure. The four terms comprising Bayes' theorem play a central role in probabilistic inference, and are therefore given special names:

- $p(D|M)$ is known as the *likelihood*; it reflects the probability of obtaining the particular dataset, assuming a particular model. It is usually straightforward

to estimate, as it typically just involves plugging in the observations directly onto the model under investigation, and observing its output.

- $p(M)$ is known as the *prior probability*; it reflects our prior knowledge about the model itself, i.e. it is a probability distribution over the model, in the sense that the model is treated as a random variable which can collapse into a distinct number of states (depending on the particular values / instantiations of its parameters), where each state is associated with a particular probability of occurring).

- $p(M|D)$ is known as the *posterior probability*; for the range of parameter states a particular model type can take, the posterior probability tells us how probable these states are given the particular dataset. In this sense, the posterior probability as a distribution, is a quantification of the uncertainty characterising the selected model as a whole, and therefore serves as a measure of its quality / reliability as a model.

- $p(D)$ is known as the *evidence*. In the first instance, this term serves in a mathematical sense as a normalisation constant, ensuring that the posterior probability as a function (i.e. the area under the curve over its range) sums up to 1. This is necessary, since for the posterior to be a valid probability describing a closed system of mutually exclusive and exhaustive possible events, the individual probabilities over all those events need to sum up to 1 (see eq. 2.26). However, as the name suggests, it is also important in itself, as the probability of obtaining the particular dataset under a particular model type *irrespective of the parameters chosen* (i.e. the parameters are said to be marginalised out). This probability will be different for a different model type, with more reliable models resulting in higher probability. Therefore this term acts as a quantification of the *evidence* for a particular model type, and is very useful in model selection, when choosing a particular model type over another.

♦

An in-depth treatise of mathematical probability theory and the specific probabilistic properties of specific models is beyond the scope of this section. From a mathematical point of view, fuzzy theory *subsumes* probability theory, i.e. the principles of probability theory can be expressed identically within the context of a fuzzy theoretical framework, where probabilities are a specialised case of ordinary fuzzy measures [71]. However, in practice, probability theory is a vast, specialised topic in itself, and as such, it tends to be studied in a largely independent manner from fuzzy theory.

Summary

- Segmentation is a crucial image analysis step for quantifying information from medical images. While general approaches to segmentation exist, non-trivial problems require approaches that are modality and specialty specific, therefore we choose to focus on a specific modality.

  Cardiac Magnetic Resonance imaging, through the use of imaging protocols such as Perfusion, Cine MRI, T_2 and LGE protocols, are established in the management of Ischaemic Heart Disease and clinical standards exist for standardized automated segmentation.

  State of the art approaches, briefly described here, include deformable models, active shape and appearance models, atlas registration with segmentation propagation, image / voxel classification methods, ensemble methods, and neural networks.

- Fuzzy sets are generalisations of classical sets, where each element can be partially, or ambiguously present in the set, as specified by a membership function μ; triangular norms and triangular co-norms are generalisations of the intersection and union operations to fuzzy sets, of which the Gödel, Product and Łukasiewicz t-norms (and respective t-conorms) are of special significance.

- Probabilities have found extensive use in the study of uncertainty, within the particular context of random variables. From a mathematical point of view, fuzzy theory subsumes probability theory; however, in practice, probability theory largely tends to be used in isolation.

"As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality."

— Albert Einstein; *"Geometry and Experience" lecture, delivered at the Prussian Academy of Science, Berlin, 27 Jan 1921*

3

Fuzziness, probability and uncertainty in segmentation

This chapter introduces the concept of the Partial Volume Effect (PVE) and the subtle differences between the concepts of fuzziness, probability, and uncertainty in general, as they relate to soft segmentations and their algorithmic implementations. We explore the idea that rather than treating segmentation uncertainty as a nuisance factor, it has the potential to be used as a source of information and be put to further use.

We propose two frameworks in this spirit: one approach to fuse segmentation algorithms based on their uncertainties at the pixel level, and one that treats the discrepancy between estimated and expected clinical parameters as clinical uncertainty, enabling clinicians to guide algorithms to more clinically relevant and reliable results[1].

Contents

[1]Published as: Tasos Papastylianou et al. "Fuzzy Segmentation of the Left Ventricle in Cardiac MRI Using Physiological Constraints". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2015, pp. 231–239

3.1 The Partial Volume Effect, and the need for soft segmentations

The canonical definition of 'segmentation' is "the partitioning of an image into constituent [i.e. mutually exclusive] regions or objects" [74], which typically involves labelling all pixels individually in an image as belonging to a single class from a set of discrete classes (the simplest case being a foreground and a background class). Any attempt to quantify tissue from a medical image will typically involve a segmentation / classification step, such that the appropriate pixels are labelled as belonging to the tissue of interest, which can then be used to obtain statistics and measurements for that tissue (e.g. a volume measurement could be a simple case of measuring all the constituent pixels).

Ideally there would be a one-to-one mapping from each tissue of interest to a single class per pixel; however, in practice, this is an oversimplification, and deterministic segmentations (i.e. a pixel can only belong to a single class to the exclusion of all others) cannot fully represent (and therefore reliably quantify) tissue distribution. One reason for this is that the segmentation technique itself might classify pixels deterministically, under circumstances where it is not necessarily clear how the pixel should be classified. For instance, in an ideal investigation,

two different tissue types might have mutually exclusive intensity profiles, making segmentation a simple case of thresholding by pixel intensity; in reality, due to the method of imaging, image noise, or other factors, two distinct tissue types may be represented by intensity profiles which display significant overlap, and any deterministic classification performed in the context of such ambiguity, is necessarily less accurate and less precise [75].

But perhaps more importantly, independent of the segmentation approach used, the limits in resolution imposed by the digitization / quantization process of constructing a medical image, particularly when the tissues of interest contain features at a smaller scale than what can be adequately represented at that resolution, lead to the problem of the *Partial Volume Effect* (PVE): pixels are likely to contain a mixture of tissues due to the finite resolution of the imaging process, where the particular mixture gives that particular pixel its appearance, in a manner that makes inferring the mixture proportions a non-straightforward task. This is particularly the case in Cardiac MRI, where the pixels tend to be anisotropic, i.e. the resolution *between* slices tends to be particularly low (typically an order of magnitude less), compared to the *in-slice* resolution. Some methods consider PVE to be a *corrupting* factor, and take measures to 'correct' for it. However, attempts to treat it as an image *feature*, attempting to model and *estimate* it, have been shown to lead to more accurate and precise estimates of clinical parameters (e.g. [6–8]).

'Soft' segmentations are an attempt to address this problem, by providing a means of representing this ambiguity / fragmentation in the segmentation output. A soft segmentation can essentially be thought of as a collection of fuzzy sets over the image domain (i.e. one fuzzy set per 'label' or tissue of interest), whose membership functions at each pixel in the domain describe the extent to which that pixel 'belongs' to each of the objects of interest (with respect to the particular classification criterion).

However, it is important to note that 'softness' merely describes a degree of membership, but the particular semantic interpretation of what 'softness' actually represents is completely contingent on the process that produced it, and cannot automatically be taken to imply a partial volume effect, let alone a straightforward measure of proportionality in a class mixture. Care therefore needs to be exercised when interpreting a soft segmentation to make inferences and measurements from the soft pixels. For example, let's consider a soft segmentation produced by classifying pixels into discrete classes based on intensity profiles. Let's assume two of the candidate classes display the same probability of manifesting in a particular intensity, such that a pixel with that intensity has 50% chance of belonging to class 1, and 50% to class 2 (assuming that the pixel can only hold a single class). In a deterministic segmentation algorithm, we would make a decision to label the pixel as one of the two candidate classes, and represent this as two separate binary masks, one containing a 'zero' value for that pixel, and one containing a 'one' value; in a soft segmentation approach, we could instead represent this situation as two fuzzy masks each containing a pixel with value 0.5; the important point here is that *this is not the same interpretation as saying that that pixel consists of a mixture of 50% class 1 and 50% of class 2*, since 'softness' here represents an entirely different process. In fact, a PVE interpretation here would be incompatible with the premise underlying the algorithm used, namely that pixels were assumed to only be able to hold a single tissue class. What we *could* say, however, is that it is possible that a mapping from one interpretation to the other exists, and that it would be possible therefore, to model such a mapping appropriately (for instance, involving each fuzzy value being mapped to an appropriate set of compatible tissue 'mixings').

3.2 Uncertainty in medical image segmentation

3.2.1 Probability as uncertainty

One of the commonest ways of representing uncertainty in engineering contexts is through the notion of *probability*, and in particular the *Bayesian interpretation of probability*. Under this interpretation, a probability $p(X)$ with respect to an *event* X, represents a measure of uncertainty over the nature of that event, expressed as a value in the interval $[0, 1]$, representing the range from total uncertainty to absolute certainty. However, probability is not always used to represent uncertainty in this sense. The exact meaning of probability is dependent on the particular *interpretation* we ascribe to it. In the context of the *classical* (or *frequentist*) interpretation of probability for instance, probabilities are seen as signifying the proportion of items within a larger (potentially infinite) set that conform to some criterion, or equivalently, the frequency of occurrence for a random, repeatable event (expressed as a ratio over the number of total attempted repetitions), or the expected / estimated frequency in the limit of infinite repetitions [76, 77].

In the context of PVE, between the Bayesian and the frequentist approach, perhaps the more useful interpretation of the two would be the latter one, as it lends itself more naturally to an expression of tissue proportion / distribution 'inside' a pixel. In this sense, probability could simply represent an "estimate" of the tissue percentage within the mixture. However, as is the case in most engineering applications, it is desirable for estimates to be expressed alongside their error margins or other expression of uncertainty in the estimate. If the process that produced the initial probabilistic (i.e. soft) segmentation mask is parameterised, and the uncertainty over those parameters is known or can be modelled appropriately, then a Bayesian framework could be used on top of a frequentist one, to produce a further measure of uncertainty over the probabilistic estimates themselves. This notion of obtaining an uncertainty mask over an already 'soft' mask is very useful; we show an application of such a mask in section 3.3.

3.2.2 Fuzziness versus probability

Probability deals with a very specific definition of uncertainty, namely the 'stochasticity' of a random variable. In probability theory terms, the uncertainty of a random variable is entirely defined by its probability distribution. In practical terms, this means that while each 'draw' of a random variable cannot be predicted precisely, the overall behaviour and expected outcomes of such a system can be analysed and described statistically.

However, as we have discussed in the previous chapter, not all forms of uncertainty can be adequately or appropriately expressed in terms of stochasticity. The expressive power of probability theory as a framework of uncertainty is therefore rather limited, as it encompasses a rather narrow definition of uncertainty.

As an example, let's consider the distinction between myocardium and papillary muscles, which are the muscular structures embedded in the walls of the ventricle and projecting out into the ventricular cavity. Many heart segmentation protocols dictate that the papillary muscles be excluded from the segmentation [78] (i.e. excluded from the myocardium 'object', and absorbed into the blood pool 'object', resulting in simpler shapes overall). However, while the papillary muscles could possibly be differentiated from core endocardium at the microscopic level, no such distinction is readily possible at the macroscopic level [79], let alone in terms of its appearance in a medical image of limited resolution. The definition of what constitutes 'myocardium', and what constitutes 'papillary muscle' is therefore vague and / or ambiguous by definition, particularly at the junction between the two, as there is no clear anatomically defined delineation specifying where one ends and the other starts. Therefore, even if we were able to observe the 'contents' of a pixel with perfect resolution, and determine the absolute ground truth, we would always be forced to classify the myocardial / papillary junction every time as "some sort of conceptual hybrid between a papillary muscle and myocardium". We could not describe this kind of uncertainty as "stochasticity", since the outcome is guaranteed

to be that particular, specific label (albeit vague) every time, with absolute certainty that that label is true (i.e. the outcome is still 'deterministic' in this sense).

However, clearly, in terms of characterising the extent to which a pixel is either "papillary muscle" or "myocardium", there is uncertainty which corresponds to the vagueness or ambiguity in the definition. This uncertainty can be "quantified" too, since, the more we move away from the junction towards the core of the myocardium, the more we can claim that the tissue is conceptually more myocardium than papillary muscle, and vice versa.

Fuzzy theory is uniquely equipped to deal with this kind of uncertainty, through the concept of fuzzy sets. More generally, together with fuzzy logic and fuzzy measures, it is able to describe a number of different types of uncertainty, as part of a complete and consistent theoretical framework; in fact, fuzzy theory subsumes probability theory, which is a specialised case of fuzzy measures. Furthermore, the use of probability makes certain intuitive assumptions over the *nature* of the variables involved, whereas fuzziness as a mathematical concept has a much broader scope. For this reason, we treat soft segmentations more generally as 'fuzzy', and only refer to softness as 'probability' when it is important that this is explicitly the case.

3.2.3 Types of uncertainty

Klir & Folger [71] distinguish between a number of different *types* of uncertainty, in terms of their semantic content, and how they relate to the various branches of fuzzy theory:

Uncertainty as vagueness

Vagueness is associated with the difficulty of making sharp or precise distinctions. For instance, if we're trying to decide if an element is a member of a class, vagueness would imply that actually this element can be defined as partly belonging to both classes (as per our myocardium / papillary muscle example above).

Fuzzy sets and fuzzy membership functions are a suitable framework to express and quantify this kind of uncertainty, and are thus suitable *measures of vagueness*.

Uncertainty as information

Uncertainty of this type, reflects the amount of information in a system that is unknown or missing, or equivalently the total amount of potential information that can be gained from that system (e.g. after observing all possible states or exploring all possible alternatives).

Uncertainty in this particular context links most naturally with classical sets and probability theory, and suitable measures of information exist in both frameworks, the two most representative from each category being *Hartley Information*, which is a measure of the uncertainty associated with a choice among a certain number of alternatives in a set, and *Shannon's Entropy*, which is a measure of uncertainty associated with the distribution of a random variable (although measures generalising the two, e.g. the Renyi Entropy [80], and formulations of entropy within the broader context of fuzzy theory [81] exist).

The basic unit of information in this context is the 'bit', which corresponds to the amount of information gained from an answer to a binary question, i.e. a question asking for a choice between two possible alternatives, such as 'true/false' or 'yes/no' questions. In particular, for non-trivial sets and probability distributions, a bit is more usefully defined as an 'optimal' binary question, i.e. one that partitions the set or probability space in 'half' [82]. This gives a quantitative meaning to uncertainty as a *measure of information*: it is the expected number of yes/no questions it would take (assuming an optimal question-asking strategy) to guess a specific outcome from a known set of alternatives (Hartley Information) or from a random variable given knowledge of the underlying distribution (Shannon Entropy) [83].

Uncertainty as ambiguity

Ambiguity is associated with the difficulty of choosing or distinguishing between two or more (distinct) alternatives based on available evidence. There are many

quantifiable concepts that could be described as measures of ambiguity with respect to evidence, and these are generally best described using appropriate fuzzy measures; Klir & Folger provide the following subcategories:

- **measures of non-specificity:** relate to the size of the space covered by the available evidence. For example, if the available evidence points to a set with a single element, the end outcome is much more certain than if the same amount of evidence (in quantifiable terms) points to an alternative set containing many elements, since in the former case all the available evidence is concentrated onto a single element (i.e. it is highly specific), whereas in the latter case it is distributed among all the members of the set.

- **measures of conflict:** relates to conflicting / distributed evidence with respect to choosing from a number of alternative candidates, under the assumption that only one such candidate holds true. In other words, if evidence exists for two candidates, evidence favouring one candidate conflicts with evidence favouring another candidate in this context.

- **measures of confusion:** relate to the distribution and strength of evidence, both *between* and *within* the elements of a set or the space of possible states of a random variable. It is an expression of uncertainty as confusion over the nature of the set (or random variable) as a whole, in terms of the available evidence. A common formulation generalises the Shannon Entropy, where the distribution of the evidence among the different states of a random variable is determined according to its probability distribution (i.e. denoting how likely each state is), whereas the strength of the evidence *within* each state is further determined according to its *possibility* distribution, which is another type of fuzzy measure relating to the degree of plausibility or (need for) believability of the evidence at hand.

The above classification of uncertainty as vagueness, information or ambiguity is largely theoretical; in principle, there is significant overlap between these categories,

depending on the semantics of what one is trying to represent. For instance, evidence to support a choice from a set of alternatives can be seen as both a measure of information relating to the available choices, *and* as a measure of ambiguity in terms of conflict of evidence. However, the distinction is useful in terms of prompting the keen researcher to explore uncertainties coming from a number of different contexts and sources, to consider the semantic interpretation of such quantified measures so as to better gauge how they impact the problem at hand, and to therefore choose the most appropriate frameworks and measures with respect to their analysis.

3.2.4 Sources of uncertainty in medical images

Uncertainty in medical images and their analysis, may result from, or be extracted from a number of different sources:

- **Error / confidence bounds:** A soft segmentation mask denoting PVE (or any other type of uncertainty) and corresponding estimates can rarely be given with complete precision, as estimating the exact nature of PVE is not straightforward, and needs to be modelled instead. Therefore error bounds on measurements are necessary, as in all engineering applications. Ballester *et al.* made the observation that most published medical image analysis methods develop segmentations and derive measurements of certain structures or lesions, but confidence bounds on such measurements are rarely provided [8]. A simple uncertainty range can be obtained by the two extremes of treating all PVE pixels as either all 0 or all 1. This is oversimplistic and would result in very wide confidence limits, but can serve as a starting point for modelling the volume distribution of the actual objects involved; for example, Ballester *et al.* modeled the underlying volume of objects with PVE by assuming a Gaussian probability distribution for the intensity of individual tissues, modelling the overall intensity distribution in PVE pixels as a linear combination of these Gaussian-distributed components for a particular mixing fraction, and then estimating a posterior distribution for the mixing fraction at each PVE pixel

via Bayes' Theorem. The intent was to obtain error measurements over the produced clinical estimates, but such an approach could easily be extended to produce uncertainty maps over the segmentations.

- **Hardware precision:** Uncertainty could derive from limits in the hardware itself [84]. For instance, sensors generally do not respond uniquely to exact signals, but rather, there may be a particular receptor profile with a particular probability distribution for a range of inputs, a probability of registering an input of a given intensity, etc. If the uncertainties underlying the hardware are known, these could potentially be propagated to the resulting intensity values in the image.

- **Consistency:** Uncertainty could reflect inconsistency in the algorithm. This is particularly the case where the algorithm depends on a stochastic step; if multiple runs of the algorithm result in areas where the resulting classification in those pixels is fairly consistent, then this could be interpreted as a measure of certainty, or robustness of the algorithm in those areas. Conversely, areas were the segmentation outcome is highly variable might indicate that the algorithm is less robust in its classification for that particular part of the segmentation. Therefore, quantification of this type of variability could be used as a measure of uncertainty denoting local inconsistency in the algorithm.

- **Soundness:** Many algorithms rely on regularisation, or other similar 'sanitization' steps. Such steps enforce 'soundness' in the final segmentation outcome, in that they typically prevent segmentations with certain characteristics that would not be realistic or meaningful, or simply not expected to be present in the ground truth. Such regularisation is often applied as a mask, i.e. different regularization is necessary at each pixel. For example, in registration-based algorithms, the regularization might penalize large deformations, and different areas of the registration may be penalized more heavily than others. Such regularisation masks can be extracted for use as a measure of uncertainty, since they indicate areas where the main body of the algorithm had resulted in a

potentially less 'sound', or meaningful result, one which had to be 'artificially' kept in check by some degree.

- **Entropy:** As discussed in section 3.2.3, entropy is a measure of the average information relayed by a variable of a particular probability distribution (or fuzzy distribution more generally). With respect to medical images, defining a suitable criterion that has a particular probability distribution at each pixel, enables us to express uncertainty for that criterion via the entropy at that pixel. Higher entropy denotes higher uncertainty, since if the information obtained with respect to that criterion once a value for that pixel has been observed, is very high on average, then this is equivalent to saying that the uncertainty surrounding that criterion *prior* to observation was also high, and following observation it has now been *reduced* by that amount. Conversely, if the entropy (i.e. average information obtained) is zero, it means we gain no new information by observing that value, and therefore there was no uncertainty about the criterion in question. Saha and Udupa used entropy to define a measure of class-uncertainty per pixel, given its intensity and a particular partition of the image into a particular mutually exclusive set of classes [20]. We will make use of a fuzzy generalisation of this method later in section 3.3.4, where we describe this in some more detail.

- **Clinical / anatomical ambiguity:** this kind of uncertainty would perhaps be the most useful for applying over segmentation masks intended to extract medical objects, but in practice it is hard to obtain. In section 3.4 we describe one such measure of uncertainty based on the suitability of clinical estimates with respect to a clinical reference standard, and show that it can be put to use to obtain improved segmentations. However, another simple, purely anatomical measure of uncertainty is the fuzzy labelling of pixels in terms of their distance, or direction with respect to one or more known landmarks. For instance, Moreno *et al.* successfully used fuzzy constraints and fuzzy spatial relationships such as the notion of "between-ness" to guide

segmentation [85, 86]. An example of how to define and combine such fuzzy spatial / anatomical masks, and particularly in a way that allows one to express the spread / strength of the overall result (or equally, the influence of its individual components, if needed) by adapting the 'strictness' of the Łukasiewicz operator, is shown implemented in figs. 3.1 and 3.2 (we will be discussing such masks in more detail in chapter 6). Beyond the fact that such fuzzy relations could be used to set simple constraints, they are also intriguing in that they could be used as part of segmentation strategies that mimic how *radiologists* identify objects in succession, based on their clinical characteristics and anatomical inter-relations to other structures, particularly structures serving as anatomical landmarks.

3.3 Combining soft segmentations using uncertainty

We have discussed in the previous chapter (section 2.2.5, p. 24) that ensemble methods, i.e. methods that combine individual algorithms with the aim of producing a better one, tend to rely on one of two principles: either the outputs (and by extension, the algorithms, or human experts that produced them) are considered to be of 'good' quality already, and the ensemble output simply represents an average or consensus approach over those segmentations, *or* a large selection of relatively 'naive' algorithms is evaluated against a validation set, to find an optimal weighted combination among those algorithms, such that the accuracy of the final ensemble performs optimally.

3.3.1 Motivation

Both approaches have pros and cons: the former requires no training, but relies on the assumption that participating algorithms are of acceptable quality already, such that the consensus process will isolate parts of individual segmentations that were

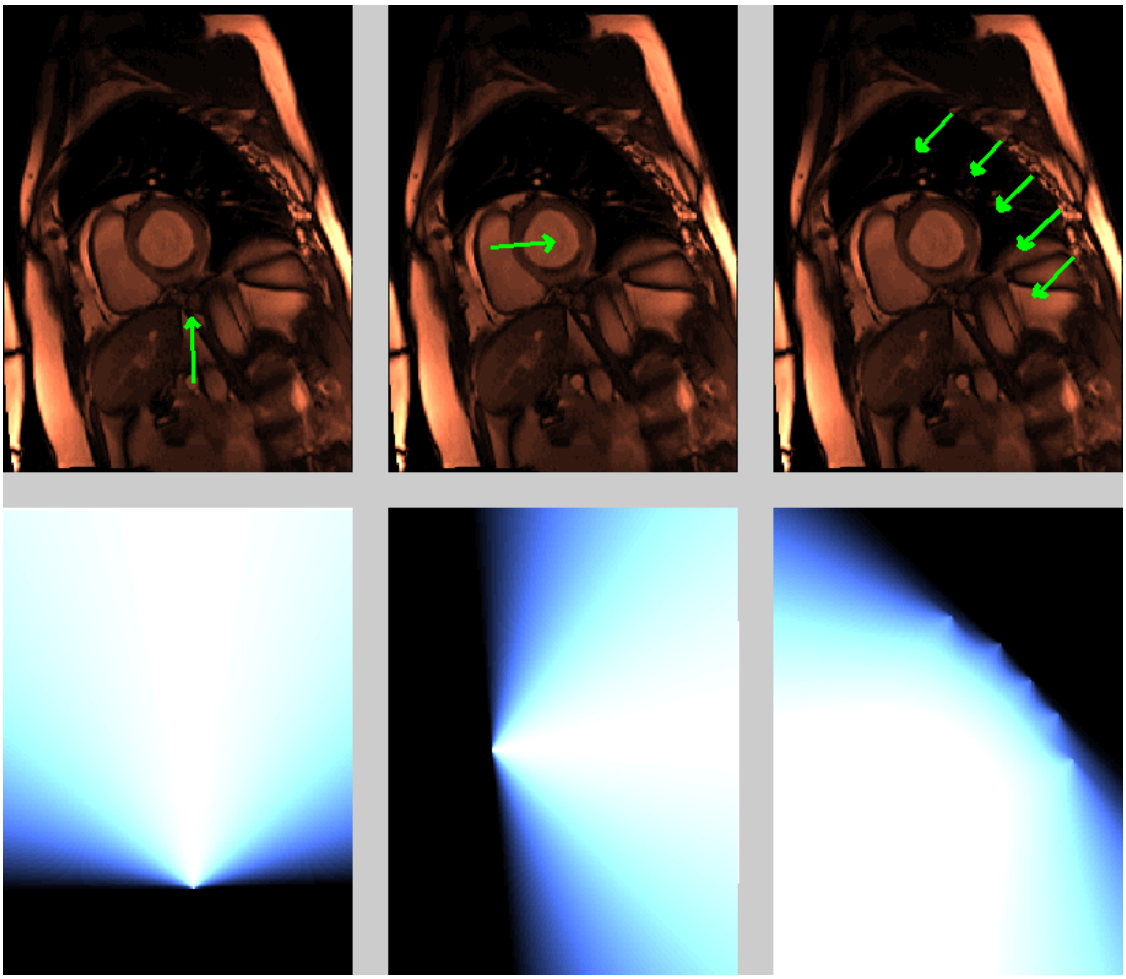


Figure 3.1: Spatial relations and corresponding fuzzy masks. From left to right: "In front of the aorta" "Left of the right ventricle" "Inside the ribcage" Fuzzy spatial masks were obtained using a simple 'cosine of the angle' rule (see chapter 6 for details); in the case of the ribcage, the overall mask is the union of individual fuzzy components.

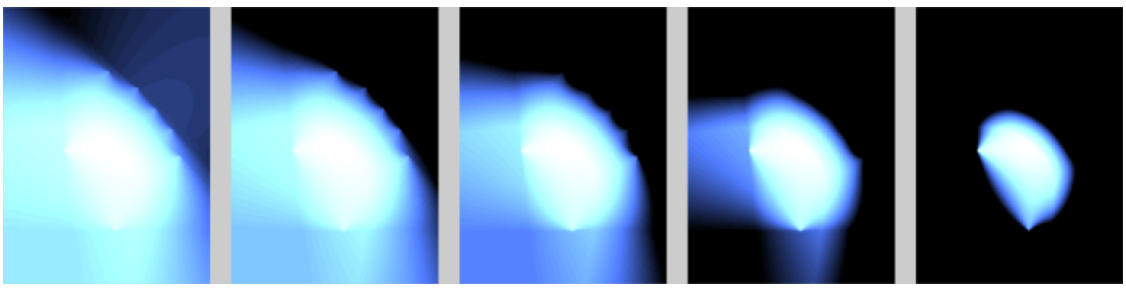


Figure 3.2: Different levels of spatial / anatomical uncertainty for all three masks of fig. 3.1 combined, using different levels of 'strictness' for the Łukasiewicz fuzzy intersection operator.

outliers of lesser quality, and only retain parts that were consistently present, and therefore considered to be more 'reliable'. The converse of this, of course, means that consensus among segmentations of bad quality cannot necessarily be expected to lead to a higher quality segmentation, particularly if such segmentations are consistent (e.g. some form of systematic bias is present); if anything such inputs guarantee a bad quality outcome. The latter approach mitigates against this by filtering out the less useful algorithms, and keeps the more useful classifiers, such that even weaker classifiers when combined can form a strong classifier. But, to achieve this filtering process, such approaches rely on training (which involves a validation step and a gold standard). Furthermore the output from this process is *not* an optimally fused, stronger *segmentation*, deriving from the fusion of individual, weaker *segmentation components*; rather it is an optimally fused, stronger *algorithm* deriving from the fusion of individual, weaker *algorithmic components*. This distinction means that, despite this optimisation process, the usual concerns about overfitting and underfitting from a single algorithm (albeit an optimally fused one) still apply (i.e. there is no guarantee that an algorithm that performs optimally during training, will perform equally optimally during an actual test).

We propose here an alternative approach, which avoids the limitations of both approaches above, and is suitable for combining outputs from soft segmentation algorithms specifically. This relies on the concepts described above, that rather than 'softness' being a measure of uncertainty itself, it is instead characterised more generally as a fuzzy measure, with the benefit that various measures of uncertainty can be applied onto the fuzzy measure itself. This means that we can combine segmentations using measures of uncertainty over the soft segmentation outcomes, rather than combining the actual soft values themselves directly, which would entail taking their precision and reliability at face value. In other words, we can combine segmentations, by selecting (or favouring) only segmentations for which we are relatively certain about their outcome, as described by a particular measure of uncertainty. More importantly, since such measures can be applied at the individual pixel level, this means that this filtering process can occur at the pixel level rather

than at the segmentation level. This is very useful since algorithms may be strong in some areas and weak in others (and this may be reflected as localised areas of higher uncertainty over the segmentation mask); by retaining only parts of segmentations that are strong for the overall output, we can get a much better ensemble output than if we were to reject a particular segmentation candidate altogether, effectively discarding useful information provided in parts where that algorithm is 'strong'.

The intuition behind this approach is similar to that for multi-atlas segmentation (MAS), where one attempts fusion over a set of label-maps produced by a collection of atlases (see section 2.2.5). Label fusion in the context of MAS makes use of prior information relating to the reliability of the registration of each atlas; this information is then used to better inform / bias the fusion process, e.g. by weighing the contribution of each label-set according to the reliability of its atlas. In the particular case of MAS, such prior information, is readily available as (potentially localised) measures of reliability pertaining to the registration itself, and is therefore information that is produced as an intrinsic part of the registration process for each atlas. Our work here is similar in its intent to provide localised measures of uncertainty that are intrinsic to an algorithm, but beyond the restrictive scope of deriving such reliability measures via atlas registration in a multi-atlas setting. Furthermore, In MAS, the reliability measures obtained are of an identical nature over all label-maps in the set. The work here is an attempt to find uncertainty measures which are more generally applicable to any class of algorithm, try to relate that to what kind of uncertainty it represents with respect to the discussion about the different contexts of uncertainty in the previous subsection, and test to what extent these types of uncertainty can be useful in the context of label fusion, even when attempting fusion using different classes of algorithm and different types of uncertainty.

3.3.2 Materials

We conducted a proof-of-concept study to demonstrate the approach. To this effect, we obtained two sets, which was deemed to be a sufficient number for this

purpose (see section 1.3); these were kindly provided by the University of Oxford Centre for Clinical Magnetic Resonance Research at the John Radcliffe Hospital, Oxford, produced on a 3.0T Siemens Tim Trio whole-body MRI scanner, and anonymised appropriately. '3D + time' images were obtained using the Cine MRI sequence protocol, as per section 2.1.2 (p. 18). Both studies were conducted as post-Percutaneous Coronary Angioplasty investigations, for patients with a diagnosis of an Inferior Myocardial Infarct; no other clinical or radiological details were available. Each set consisted of 25 timeframes; set 1 consisted of 10 Short-Axis slices per timeframe, and set 2 consisted of 8; each slice had a resolution of 256×192, with a voxel size of $1.5625 \times 1.5625 \times 8mm$. Images were extracted from the DICOM files using Laszlo Balkay's DICOM reader [87] for Matlab [88] / Octave [89]. Where registrations were performed, this was done using the DROP3D registration suite [90, 91]. All operations were performed on a Linux workstation running on an Intel Xeon CPU and 16Gb RAM; no GPU processing was used.

3.3.3 Segmentation algorithms used

We demonstrate this approach in the segmentation of the left and right ventricles of the heart from Cine MRI using two segmentation algorithms from the literature: one relying on an image-based approach, and one relying on the 'Atlas registration with segmentation propagation' approach.

The Cocosco *et al.* algorithm

The algorithm developed by Cocosco *et al.* [55] is a fast, simple technique relying on standard morphological operations (i.e. erosion and constrained dilation), and simple voxel-based statistics (such as intensity variance and volume differences of connected components). It makes use of the fact that Cine MRI has multiple timeframes spanning the entire cardiac cycle — a resource which, as noted by the paper authors, most algorithms of this area seem to ignore. The presence of timeframes is used in two ways. Firstly it is used to calculate a suitable *Region Of Interest* (ROI) based on pixel variability: this relies on the assumption that

high variability in the time dimension indicates pixels associated with a moving, beating heart; therefore, the standard deviation in the time-dimension is calculated at each pixel, resulting in a 3D object whose *Maximum Intensity Projection* (MIP) along the z-axis is subsequently thresholded and used as a ROI mask over all slices. Secondly, for the main segmentation task, Cocosco *et al.* used timeframes to identify connected components which change their volume significantly over time; the assumption being that the most 'pulsatile' components are highly likely to correspond to the two pulsating ventricles.

The algorithm for the segmentation of the left and right ventricles can be briefly summarised as follows:

- A ROI is applied as above
- An intensity threshold is applied on the 3D+t Cine MRI image using Otsu's method [33]
- A series of erosion / dilation steps are applied to each resulting connected component for shape-refinement.
- For each connected 3D component, the timeframes with the largest and smallest volumes for that component are found, and the components are subtracted to calculate a Contractility (F) value.
- The two components with the highest F values are selected; the rightmost one is labelled as Left Ventricle, and the other as Right Ventricle.

The initial, unrefined set of connected components is obtained via a simple thresholding operation, on the assumption that the blood pool is expected to be brighter than the surrounding tissues on Cine MRI; the authors used Otsu's method [33], on the grounds that it is "robust, simple, and does not have any application-dependent parameters". However, the large variability in datasets means there is no single thresholding technique which can be guaranteed to produce an optimal result for all datasets. Furthermore, even the optimal threshold which will lead to the most 'accurate' result, cannot necessarily guarantee a 'near perfect' segmentation, as the intensity characteristics of the region of interest might be such that a single,

simple thresholding operation cannot produce the desired isolated components, or capture the more subtle intensity differences between competing regions (for instance, the authors point out that the misclassification rate of the algorithm increases in the presence of bright 'patches' of fat). A straightforward 'fuzzy' modification of the Cocosco algorithm, is therefore to merge outputs produced for a range of 'optimal' thresholds.

The algorithm was reproduced from the paper, and tested for a range of thresholds as above. Fig. 3.3 shows the algorithm output for set 1 (set 2 produced a similar picture), in terms of accuracy with respect to a gold standard and certain physiological parameters relating to the left and right ventricles, as derived from the segmentation at each threshold step. We note that there is a short range of thresholds (here 0.6 to $\approx$0.8) for which accuracy is significantly higher.

Of interest is the observation that in this small range of thresholds resulting in higher accuracy, the physiological parameters obtained from the segmentations are more likely to be within more physiologically feasible / meaningful values. For instance, as seen in Fig. 3.3, this optimal range corresponds to contractility values of (on average) 2400 pixels for the LV, and 1700 for the RV; for the pixel sizes used, i.e. $1.5625mm \times 1.5625mm \times 8mm$ (corresponding to a pixel volume of $19.5 \times 10^{-3}mL$), this corresponds to a stroke volume of 47mL vs 33mL, i.e. a total stroke volume of 80mL, and stroke volume ratio of 0.70. These are of similar order as physiological values; normal range is around 100mL per ventricle, and a ratio of 1 [92], however, the lower values here seem genuine (as confirmed from the manual segmentations) and likely represent a clinically reduced stroke volume for these patients, due to the myocardial infarction. It can be seen from Fig. 3.3 that the other measures (such as volume ratios etc) corresponding to this region of higher accuracy, are also characterised by values that are closer to physiologically meaningful / feasible values, compared to the rest of the 'intensity threshold' space. In particular, we also note that the default threshold used in the paper (here just under 0.5, based on

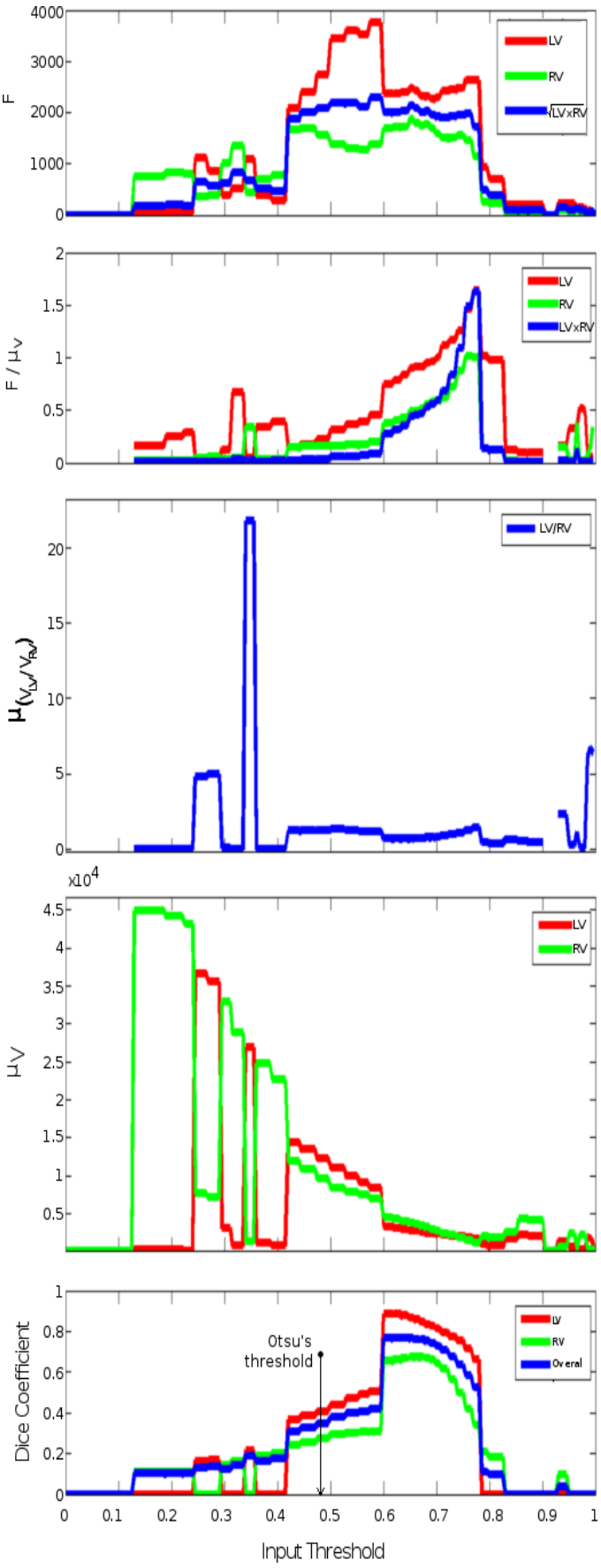


Figure 3.3: Effect of varying the intensity threshold in the Cocosco Segmentation algorithm. For this image, the threshold chosen by Otsu's method (i.e. the method used by the original algorithm) is 0.48, corresponding to an overall Dice Coefficient ≈ 0.4.

Subfigures from top to bottom:
a) Contractility: F (versus input intensity threshold)
b) Contractility normalised by the average volume for each component over all timeframes: F/μ_V.
c) Left-to-Right ventricular volume ratio, averaged over all timeframes: $\mu_{(V_{LV}/V_{RV})}$
d) Mean volume of each component, averaged over all timeframes: μ_V.
e) Dice Coefficient for the segmentation, against a manually delineated validation set.

Otsu's method) results both in suboptimal accuracy, *and* values for the physiological parameters that are not realistic in a physiological sense.

For the purposes of creating a high-quality soft segmentation based on the Cocosco *et al.* algorithm to demonstrate the concept of combining probabilistic algorithms through their uncertainty, in this instance we simply opted to obtain a fuzzy segmentation output simply by averaging the segmentation outputs produced for the particular 'optimal' range observed at validation, so we don't make particular use of the "physiological link" noted above at this stage. However, we make use of this concept later on in section 3.4 where we propose a method for putting this information to use, in order to obtain more accurate and clinically relevant segmentations.

An atlas registration and segmentation propagation method

The second algorithm used in this experiment was an *atlas registration and segmentation propagation* based algorithm. The general principles underlying this approach have been covered in section 2.2.3 (p. 22), but we outline the basic steps here again for clarity:

- Obtain a representative image, (the *intensity atlas*) with a known segmentation (the *label atlas*).
- Register the intensity atlas to the target image and retain the deformation field
- Apply the deformation field on the label atlas, to 'propagate' the labels onto the target image and obtain a segmentation

Our particular implementation relied on the DROP3D registration suite for image registration, which implements a registration algorithm based on a discrete optimisation approach using a b-splines model and multiple scales [90, 91]. Both sets were pre-processed to apply a ROI (as per the Cocosco *et al.* method above), and to perform histogram equalization, such that the intensity profiles would be similar to the intensity atlas used to improve registration; the similarity measure chosen to

drive the registration was the Sum of Absolute Differences (SAD), as this gave the most meaningful result on inspection, and did not result in 'holes' in the registration.

A limitation of this approach in general is that if there is no perfect structural correspondence between the two structures to be segmented (and an appropriate mask to narrow the registration focus onto those structures), then the segmentation will by definition not be able to identify the different structures in the target image. The accuracy is further compromised if the quality of the two images is different enough to cause the registration to converge onto irrelevant features (such as between images of significantly differing intensity profiles, with intensity being the image feature driving the registration algorithm). Approaches to overcoming these problems include: probabilistic atlases averaged from a number of atlas sources [51], in the hope that more structural features will be present for registration; registration from multiple atlases with fusion of the individual propagated segmentations [62]; and techniques relying on features other than intensity (such as local phase [93]). However, in our case, due to lack of a large number of appropriately pre-segmented sets, or appropriate number of segmentations per set, 3D registration was simply performed for each set, using the other set as a registration source (i.e. serving as the atlas). Since the deformation may result in labels occurring at non-integer coordinates, interpolation was performed to obtain the corresponding values at the standard integer coordinates. Intensity interpolation resulted in a 'split' myocardium appearance, therefore shape-based interpolation was applied instead using the technique introduced by Herman *et al.*[94], which resulted in visually appropriate segmentations.

In order to obtain a soft segmentation at this stage, we relied on registration of each atlas for a range of λ values, where λ controls the influence of the penalty function used for regularization of the deformation (chosen here to be a simple Pott's model [95], penalizing deviations in displacement). A regulariser function generally serves to constrain the deformation, to ensure the atlas coordinate system is not deformed in a 'meaningless' way; since this also represents 'semantic' information

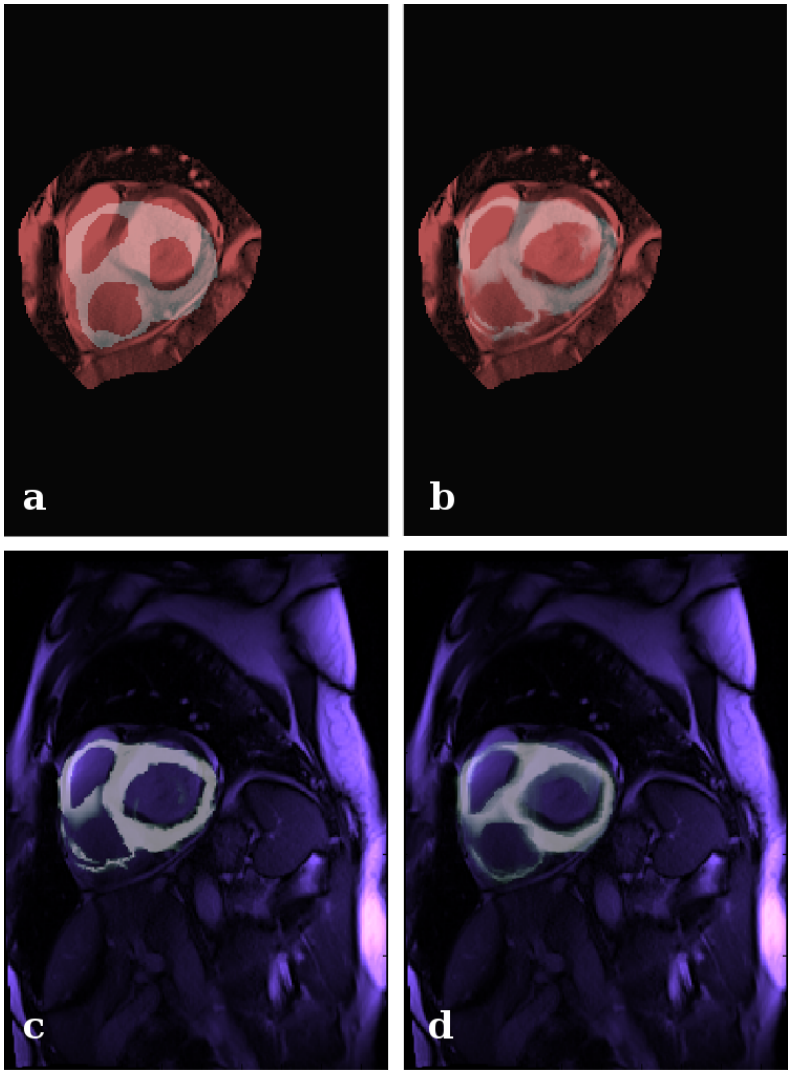


Figure 3.4: Results of the Atlas-based segmentation (shown on a representative slice).
a) Label atlas (white mask) before deformation (overlaid over target set, with ROI applied)
b) Label atlas *after* registration and deformation (for a particular regulariser value)
c) Same as *(b)*, but shown over the whole Short-Axis slice (i.e. with no ROI applied)
d) The final soft segmentation output after averaging.

to an extent, the resulting distribution is likely to demonstrate higher 'certainty' in areas where the registration is more robust regardless of regularisation, and higher 'uncertainty' in areas where specific regulariser ranges affected the consistency of the algorithm. For our purposes we used a range of regulariser values, from 0 to 0.01, and the resulting segmentations were fused through simple averaging, which resulted in visually appropriate soft segmentations.

We note this approach is similar in spirit to the work of Rohlfing and Maurer Jr (2005) [96], where a range of parameterisations was used in the context of an atlas registration algorithm, in order to obtain a collection of label-*confidence* maps (where these take the form of a vector of weights per pixel, with each weight

corresponding to 'confidence for each class', taking values in the range $[0, 1]$). In [96] this process was loosely referred to as 'bootstrapping' (even though the process is not related to classical bootstrapping), in that it resulted in a collection of related outputs, thought to represent the same underlying algorithmic process[2].

An important difference between our approach and [96], is the nature of the intended outcome: in [96], the intended outcome is a binary label from the fusion of a set of 'fuzzy' label-confidence maps; the fusion involves computing the unweighted sum over all confidence maps, and selecting the class with the highest 'confidence' at each pixel. In our work, our intention is the reverse: to produce a '*fuzzy* label output' and 'corresponding uncertainty map' pair, for subsequent fusion with 'label map / uncertainty map' pairs originating from other algorithms (and, more to the point, for potentially different 'types' of uncertainty map). Therefore, for this particular algorithm, the intent was to produce a collection of *actual* label-map components (soft or otherwise), from which we obtain mean, and variance maps, denoting the fuzzy 'aggregate' label-map, and an uncertainty / 'reliability' map corresponding to how consistent the underlying components were at each position.

Segmentation propagation was generally not optimal, as can be seen from the registration in Fig. 3.4 (note that the myocardial class is shown in these results, as it is easier to visualise areas of weakness in the segmentation; however, the focus of comparison later is the ventricle segmentations). The soft segmentation result, shown in Fig. 3.4, was a slightly better shape, but still displayed systematic bias. One possible reason for this bias could be that DROP3D is software for general medical image registration, but not specific to heart images; the availability of a more bespoke Cine MRI Heart registration software might have produced better results. Furthermore, obtaining a fuzzy segmentation in this manner is not very time-efficient (an order of magnitude more than the Cocosco algorithm as reproduced specifically for the needs of this experiment); however, at the moment

[2] Similarly, the aggregation of these outputs was correspondingly referred to as 'bagging' (i.e. from **b**ootstrap **agg**regat**ing**), in keeping with terminology used in more classical machine learning.

this may be partly due to the high cost of the external registration software and a bespoke solution could potentially be optimised accordingly.

Having said that, we reiterate that our goal here was not to construct a 'perfect' atlas-based segmentation, but to use this result to show whether 'weaker' outputs can still be combined via their uncertainties to provide better segmentation results. The choice and optimality of segmentation algorithms is therefore not that important in this context, but is more similar in vein as using a combination of weak classifiers to obtain a stronger one (except here we are using potentially weak 'outputs' instead).

3.3.4 Measures of uncertainty used

In order to combine the two segmentation algorithm outputs on a pixel-by-pixel basis based on their individual uncertainties, suitable pixelwise measures of uncertainty must be used.

Variance as pixelwise classification inconsistency / imprecision

Given that we construct our soft segmentations in both cases using an averaging approach (i.e. averaging the classification outputs at each pixel), the simplest measure to use for the *uncertainty* at each pixel is the *variance* of the classification values. The particular interpretation of this measure as a measure of uncertainty is that it is a measure of inconsistency / imprecision for that algorithm at a particular pixel location. In other words, pixels where the algorithm is consistent in its classification output irrespective of the particular value chosen for the parameter(s) being varied, can be said to be more 'certain' in its classification decision; conversely, if the classification decision at a particular pixel varies significantly, then we could say that the algorithm is less robust at that particular pixel, since the classification outcome for that pixel is highly dependent on the choice of parameter.

Entropy as class uncertainty given pixel intensity

A somewhat more sophisticated measure was proposed by Saha and Udupa for calculating an intensity-based class uncertainty based on *entropy* [20]. This is

defined on a per-voxel basis, as the entropy resulting from the classes represented by that voxel given its intensity, and some basic assumptions about the distribution of intensities per class (Gaussian, in the original paper). Intuitively, entropy here is a measure of how much information (e.g. in bits, i.e. the number of optimal 'true/false' questions; see section 3.2.3), one would require at such a voxel given the available knowledge, in order to reach the right conclusion about the class represented by that voxel with full certainty. Saha and Udupa used this to develop an optimal thresholding strategy, minimizing total image uncertainty (while also taking into account region homogeneity), and subsequently used the same concept as a functional for an active contour segmentation algorithm [97].

Since the Saha and Udupa paper deals with deterministic segmentations, instead of calculating a mean and standard deviation of intensities for each class 'partition', the algorithm was generalised for soft segmentations and entropies of fuzzy partitions, by applying the main ideas outlined in [98]. Namely, for an image I with N voxels such that $I_i, i \in 1, \ldots, n$ represents the intensities at each voxel, and M number of distinct classes C_m, such that $C_{m,i}$ represents the class probability for class m at voxel i, then mean μ_{C_m} of intensities for class m and standard deviation σ_{C_m} are defined as:

$$\mu_{C_m} = \frac{\sum_{i=1}^{n}\left[I_i C_{m,i}\right]}{\sum_{i=1}^{n}\left[C_{m,i}\right]}, \qquad \sigma_{C_m} = \sqrt{\frac{\sum_{i=1}^{n}\left[\left(I_i - \mu_{C_m}\right)^2 C_{m,i}\right]}{\sum_{i=1}^{n}\left[C_{m,i}\right]}} \tag{3.1}$$

These values were then used as parameters for the Gaussian distributions; where $\sigma_{C_m} = 0$, to avoid *division-by-zero* errors, voxels with $I_i = \mu_{C_m}$ were manually assigned a probability of 1, and all other voxels assigned a probability of 0. Furthermore, where probabilities of 0 occurred in the calculation of entropy, this was set to 0, since $\lim_{p\to 0}\left[p \log p\right] = 0$.

3.3.5 Merging strategy

Two fuzzy segmentation masks were obtained from the two soft segmentation algorithms ('Cocosco-based' and 'Atlas-based') as above, and for each one, Variance-

based and Entropy-based uncertainty masks were produced. The uncertainty masks were then converted to 'certainty' maps, normalised to be in the range [0,1] as follows:

- for the Variance-Based uncertainty mask, a 'certainty' mask was produced by taking the complement of the uncertainty mask (i.e. $1-\sigma^2$); this is guaranteed to be in the range [0,1] since the variance of a Bernoulli or fuzzy random variable is always less than 1.
- for the Entropy-Based uncertainty mask, since entropy is not bounded by an upper limit, the certainty value for a class at a particular pixel was defined as the sum of all other classes' uncertainties, normalised by the sum of the uncertainties of all classes combined (i.e. ranging from 0 to 1).

We reiterate that this is effectively an information fusion problem; specifically, label fusion, with or without 'prior' / 'reliability' information, as discussed in the previous chapter (see section 2.2.5 on Ensembles). Whilst more elaborate methods for fusion exist[3], particularly for taking into account 'prior' information over the label maps in the context of a probabilistic, Bayesian framework, (such as the STAPLE 'Maximum A-Posteriori' variant [60] mentioned earlier), in our case the 'softness' in both label masks was specifically obtained as an unweighted average over a distribution of binary components, rather than as an 'a posteriori' estimation of label probability[4]. A consequence of this is that the resulting 'softness' in these masks need not be interpreted in terms of denoting 'stochastic uncertainty', as is the typical view of probability in the Bayesian context, but only as 'probabilities' in the 'frequentist' sense at most, i.e. denoting a summary statistic of the label

[3] Both in terms of more general literature on classifier fusion, (e.g. [99], and more recent probabilistic approaches, e.g. [100], [101]), *and* in terms of the specific context of medical image segmentation; although, we note that a large part of the literature in the latter category is specifically contingent on the context of multi-atlas segmentation (see [62]), in that they require an underlying atlas to be present per fusion component, which is not applicable here.

[4] Note that we would not refer to this step as a 'fusion' step yet as such; the unweighted average and variance maps obtained in this step are simply intended to produce a soft segmentation 'summary' describing the distribution of the binary subcomponents produced by each algorithm, in terms of their accuracy and precision, to be used for subsequent fusion between different algorithms, rather than to produce already "optimally" fused outputs at this point.

frequency in the context of independent sampling events, and therefore more akin to class representation / presence within the pixel instead.

Given the above, and since we are more interested in studying whether the two different 'weighting' approaches can have a beneficial effect on fusion (i.e. to compare fusion weighted only using internal information, against fusion weighted via the uncertainty maps above), rather than exploring a Bayesian probabilistic framework at this point, we opt to compare using more traditional, 'additive' approaches, inspired from earlier fusion literature, like 'majority vote' (e.g. [102]), and 'sum fusion' (i.e. sum or average of class 'confidence' weights, selecting the highest summed weight if a binary label is needed, e.g. [96]), which in this case, should be more intuitive and straightforward to perform on 'soft' labels of this kind, while still demonstrating the underlying effects clearly, and avoiding any potential problems arising from the disproportionate influence of pixels with near-zero soft values / uncertainties when using multiplicative methods (a phenomenon more commonly known in the information fusion literature as the '*veto effect*' [103]), which are more relevant to Bayesian methods than simple 'additive' ones.

The two soft segmentations were combined using the following approaches for comparison:

i. For each voxel, keep the segmentation with the highest soft value (i.e. 'majority vote' w.r.t. components, using soft values)
ii. For each voxel, keep the segmentation with the highest Variance-based certainty value (i.e. 'majority vote' using external information for voting)
iii. For each voxel, average the two soft values (i.e. unweighted 'sum fusion' w.r.t. subcomponent labels).
iv. For each voxel, average the two soft values weighted by their relative 'confidence' (i.e. as above, but weighted by final soft value, assuming it represents algorithm 'confidence')

v. For each voxel, average the two soft values weighted by their relative Variance-based 'certainty' (i.e. weighted 'sum fusion')

vi. For each voxel, keep the segmentation with the highest Entropy-based 'certainty' (i.e. 'majority vote' using external information)

vii. For each voxel, average the two soft values weighted by their relative Entropy-based 'certainty' (i.e. weighted 'sum fusion')

3.3.6 Results

In general, in our experiments, set 2 had a worse segmentation than set 1, and the Cocosco-based algorithm performed better than the Atlas-based one. However, in terms of the effect of fusion, the increase in accuracy for uncertainty-driven fusion methods over pure probability-driven methods is evident; Fig. 3.5 shows the segmentations produced along with their respective Dice scores.

3.3.7 Discussion

We note that, while the uncertainty-based approaches seem to perform better in general than the more naive averaging / consensus approaches to fusion, which do not use any other external information, the fusion results in this particular experiment, generally come close to, but are still less accurate than the unfused Cocosco-based soft-segmentation method. This is somewhat disappointing, and rather surprising at first glance, as we might have hoped that the end-fused result would be stronger than either of its components, given that the fusion weighting is done at the pixel level. However, unfortunately it not entirely unexpected in this particular set of experiments at closer analysis; we can identify some weaknesses explaining this effect, both in the performance of the component algorithms, and in the uncertainty measures used.

Firstly, this effect may partly be explained by the systemic bias in the Atlas-based scheme (for the reasons explained earlier), leading to reduced overall accuracy and relevance of the soft segmentation as a whole, and therefore its role in the fusion process.

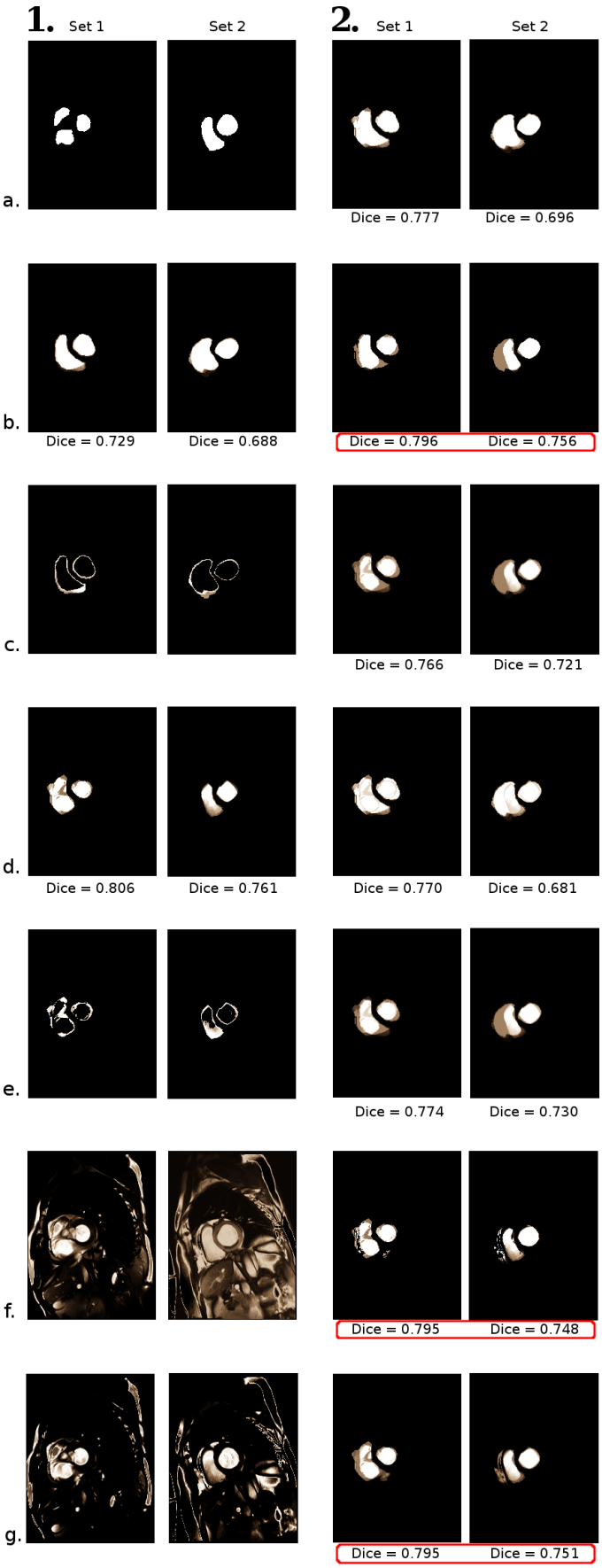


Figure 3.5: Results from Fusion Experiments:

1. Components used:

a. Gold Standard (manual delineation).
b. Atlas-based result
c. Atlas-based certainty (Variance).
d. Cocosco-based result.
e. Cocosco-based certainty (Variance).
f. Atlas-based certainty (Entropy).
g. Cocosco-based certainty (Entropy).

2. Fusion outcomes:

a. Method i: (keep highest fuzzy value).
b. Method ii: (keep fuzzy value associated with highest 'Variance' certainty).
c. Method iii: (unweighted average of fuzzy values).
d. Method iv: (fuzzy-value weighted average of fuzzy values).
e. Method v: ('Variance'-certainty weighted average of fuzzy values).
f. Method vi: (keep fuzzy value associated with highest 'Entropy' certainty).
g. Method vii: ('Entropy'-certainty weighted average of fuzzy values).

Best-scoring outcomes are outlined in red — these all correspond to uncertainty-based fusion approaches.

Equally, with regard to the Cocosco *et al.*algorithm, a systematic weakness identified by the original authors — and also confirmed in this study — is that the presence of bright fatty regions near the Right Ventricle, causes them to be mistakenly, but consistently identified as part of the ventricle.

Secondly, with respect to the uncertainty measures used, while we have shown that merging using uncertainty is of value in general, in that it represents the utilization of extra information which can be used to improve the fusion outcome, clearly, more appropriate measures of uncertainty are more likely to lead to even better results. For instance, we mentioned that the Variance-based uncertainty measures algorithmic inconsistency. While this is a useful measure of uncertainty, it is important to clarify this does not equate to areas where the segmentations are 'better' or 'more confident' from a *clinical*, or *anatomical* point of view; it is only a measure of algorithmic consistency, and therefore only a measure of certainty in the sense that it reveals areas in which the algorithm is more or less robust. However, in the presence of systematic bias in the algorithm, it is possible for mis-segmented areas to appear 'certain' in this context, in that, while the segmentation is wrong, it is wrong in a consistent manner over the parameter range in question.

Similarly, the Entropy-based uncertainty measure, measures uncertainty in relation to pixel intensity, but this is also not a clinical or anatomical definition of uncertainty, and potentially subject to the intensity limitations discussed earlier in the context of the PVE and the unpredictable intensity appearance of mixed tissue.

Therefore, a more useful type of uncertainty would be 'semantic uncertainty', rather than algorithmic, i.e. one that expresses some form of localised clinical or anatomical uncertainty.

In other words, how certain are we for each segmentation that it is a clinically meaningful result, rather than simply inconsistent to other results in the set or expectations deriving from pixel intensities. Such an uncertainty would be much more useful as a weighting factor for combining with other segmentation strategies,

especially if it can be applied on a voxel-by-voxel basis rather per segmentation result as a whole. Naturally, hybrid uncertainty measures that take into account both 'clinical' and 'algorithmic' types of uncertainty (with a particular weighting), may yield even more relevant results depending on the particular circumstances.

In the next section, we examine therefore, how we might obtain such a "clinically relevant" measure of uncertainty, and how we might put it further to use to improve segmentation quality and enable clinicians to fine-tune control of segmentation algorithms using clinical knowledge, as opposed to requiring detailed knowledge of the internal workings of algorithms.

3.4 Guiding segmentation outcomes using clinical parameters

In this section we describe a general framework[5] for adapting existing segmentation algorithms, such that the need for optimisation of intrinsic, potentially unintuitive algorithmic parameters is minimized, focusing instead on applying basic clinical knowledge. This allows clinicians to easily influence existing tools of their choice towards outcomes with physiological properties that are more relevant to their particular clinical contexts, without having to deal with the optimisation specifics of a particular algorithm's intrinsic parameters. Similar ideas have been explored elsewhere, e.g. the EFIS system [104, 105], which requires the presence of a gold standard, and relies on progressive clinician feedback in the presence of 'bad' segmentations, to improve its classification approach and accuracy, by adapting fuzzy rules in charge of selecting optimal parameters. In our case, we propose a method that does not rely on a gold standard, and clinician input is only required as a simple initialization step, which involves the provision of clinically relevant physiological constraints over the clinical parameters that can be estimated from a

[5]Published as: Tasos Papastylianou et al. "Fuzzy Segmentation of the Left Ventricle in Cardiac MRI Using Physiological Constraints". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2015, pp. 231–239

segmentation. The discrepancy between the constraints provided and the estimates obtained, serves as a form of 'clinical uncertainty', which can be used to guide the segmentation to a more clinically relevant, and therefore optimal result. This is achieved by a structured exploration of the parameter space resulting in a subspace of relevant segmentations, and by subsequent fusion biased towards segmentations that best adhere to the imposed constraints.

Specifically, our framework aims to enhance existing segmentation algorithms, with the following goals in mind:

- in the case of deterministic algorithms, propose a generalisable method to produce a fuzzy / probabilistic equivalent, by exploring the parameter space in a structured manner so as to produce a segmentation space, and then fusing the results appropriately using an ensemble approach;
- reduce the need for predefined, optimal, problem-specific parameter sets, by weighing each segmentation in the segmentation space according to its compatibility with respect to intuitively defined physiological constraints, before fusing them together into a fuzzy segmentation result;
- use the above to improve the output and usability of *existing* algorithms from a *clinical* viewpoint, by allowing clinicians to guide segmentation algorithms towards results that are more relevant to their particular clinical context by simply defining such a context in intuitive clinical terms, rather than attempt algorithmic optimization via trial and error over an unintuitive set of parameters.

We demonstrate this framework on an existing segmentation algorithm proposed by Heiberg et al. [106]. This algorithm was chosen both because of its simplicity, making demonstration of the concept straightforward, and also because an implementation is made freely available online by the authors (Segment — `http://segment.heiberg.se`), which has been shown to be robust and produce

accurate results (an extensive list of publications citing the algorithm is made available on the project website)[6].

3.4.1 The Heiberg algorithm's intrinsic parameters

The 2005 algorithm proposed by Heiberg et al. [106] (henceforth called the Heiberg algorithm) is essentially a *deformable model*-based segmentation approach. The model, consisting of a set of 2D active contours (one per Short-Axis slice), seeks to achieve an equilibrium between two competing sets of forces acting on its surface, while taking into account within-slice and temporal information; at each iteration, external forces guide the evolution of the model towards image-dependent features, whereas internal forces constrain the evolution, such that model smoothness and shape are relatively preserved.

The model has two external forces, an inflating *Balloon force*, and an *Edge force*. The Balloon force is dependent on local intensity, favouring expansion of the contour in areas closer to the estimated object's average intensity (as initialised by the user by selecting a single voxel lying within the left ventricle from the image). The Edge force is defined in terms of edge images derived from the image. Four edge images are produced, corresponding to estimating image edges in 4 different directions. At the point of calculation of the Edge force, the most appropriate edges to evolve towards are chosen given the direction of evolution of the model. Temporal information is introduced to the model by smoothing the edge force at each node-point of the model over several timeframes.

There are four internal forces with the purpose of ensuring spatial and temporal smoothness: a *Curvature force* which promotes smoothness in the overall contour shape, a *Damping force* and an *Acceleration force*, which ensure spatial continuity of

[6]This also removes inherent problems with interpretation that stem from replicating an algorithm described only theoretically but for which a public implementation is not available, since one can never claim with full certainty that any discrepancy in results from published literature is not down to the particular implementation as opposed to a genuine weakness of the algorithm / dataset under examination. For this reason, replication / unofficial implementations are generally of limited value when evaluating algorithms from the literature.

the model's nodes within timeframes, and a *Slice force* which relatively discourages node movement between the slices (i.e. in the z-plane).

The above six forces are then combined in a "modality dependent" manner to control model evolution; here, "modality dependent" means choosing a set of modifiers for each force (i.e. the algorithm's parameters), which are most effective at leading the model towards a successful segmentation, given a particular investigation or image type.

Details of the mathematical implementation of these forces are beyond the scope of the present section — particularly in the context of proposing a generalised framework aiming to minimize the role played by individual parameters, and by extension their particular role in the underlying mechanics of the algorithm in question; we refer the interested reader to the original paper for implementation details.

3.4.2 Motivation: from intrinsic algorithmic parameters to intuitive clinical parameters

As with most segmentation algorithms, the Heiberg algorithm relies on a careful selection of algorithmic parameters (in this case, the individual weightings for the different contributing forces). To a large extent, the choice of parameters represents partial knowledge about the nature of the problem, or about the environment in which segmentation is to take place. For example, for algorithms that are generalisable such that they can be used in more than one modality, object, or clinical problem, a common approach is to find a generally optimal set of parameters for each scenario, suitably defined on a test database through trial and error or by machine learning. In the case of the Heiberg algorithm and their implementation, provided freely online, a selection of pre-defined parameter sets is provided, each optimised for a particular *general* scenario: different types of MRI, segmentation of Left Ventricle versus Right Ventricle, segmentation from CT, etc.

There are drawbacks to such 'scenario-based' approaches: Firstly, while a parameter set optimised on a training set adhering to a particular scenario serves as a good

starting point, as we will demonstrate further on, it does not guarantee optimal results on particular images (even within the limits of the particular algorithm), or for particular setups and clinical contexts. Secondly, selecting an optimal set of parameters is normally a process which is largely intrinsic to the inner workings of an algorithm, offering little to no intuition on how they should be adjusted to accommodate changes in clinical context to ensure a more relevant outcome. Therefore, if a particular clinical environment has a slightly different setup to the one used for the algorithm training phase, and therefore has slightly different parameter requirements for optimal results (within the limits of the algorithm) than the ones provided by the manufacturer, tweaking that default parameter set to adjust it for their own setup is usually beyond the abilities of the clinician, because it does not translate to relatable clinical information. Therefore the clinician is more likely to simply accept the suboptimal parameter set (and by extension, a suboptimal segmentation result) *as is*, and simply try to take this into account *clinically* when weighing up the information. The main motivation behind our approach, therefore, is to enable the clinician to steer a segmentation algorithm towards results which are more relevant to their particular clinical context, by allowing them to introduce intuitive and clinically relatable information to the process; this could be performed once to adjust the default parameter set to one more suitable to a particular clinical setup, or it could be performed on a per-case basis as required.

The intuition for our approach lies in the following key observation: Segmentation results which are 'better' — better, here, defined as results that are closer, in a mathematical sense, to the gold standard — will also produce estimates of *physiological* parameters — such as Ejection Fraction (EF) and Stroke Volume (SV) — which are 'better'. Our first premise, therefore, is derived by following this logic in reverse:

Premise 1: *A segmentation result producing a large number of physiological parameters, which both individually and as a group are all 'better', is more likely to correspond to a 'better' segmentation.*

If we had the theoretical ability to explore all the possible values and combinations for each of the algorithm's parameters, we would obtain a set of segmentation results, covering all possible segmentation outcomes that are possible for a particular algorithm on a given image. We refer to this *finite* set, as the algorithm's *Segmentation Space*. Equivalently, the complete *Parameter Space* is the set consisting of all possible parameter sets, each mapping to a segmentation in the segmentation space; note that contrary to the Segmentation Space which is finite (since it is defined over a finite image domain), the Parameter Space may well be infinite, depending on the nature of the individual parameters involved.

While exploration of the full parameter space is therefore generally infeasible, we can select samples from a focused region, which is most likely to correspond to more accurate segmentations. If we treat a set of N parameters as an N-dimensional vector, then a simple way of doing this is by selecting random samples from an N-dimensional Gaussian probability distribution centred at a point of interest (we show later how the spread of such a probability distribution might be appropriately determined in practice). A reasonable choice for this point of interest would be the default parameter set/vector suggested by the algorithm itself. This hopefully should restrict the segmentation space to a subset of generally more accurate segmentations, which we could then fuse to obtain a fuzzy segmentation.

It follows from Premise 1, that if we introduce a bias in the fusion process, to favour segmentations that are 'better', the fused result should logically be biased towards a 'better' fused result.

Premise 2: *In the presence of a segmentation subspace, biasing segmentation fusion towards results associated with better physiological parameters should result in a better fused result overall, compared to an unbiased fusion*

The practical implication of applying the above insights to any algorithm, is that we shift the focus from having to optimise highly unintuitive parameters intrinsic to the segmentation algorithm, to something that is more intuitive within the context

of the task at hand — i.e. the *physiological* parameters — and which is therefore easier, and more relevant to non image-analysis specialists.

3.4.3 Methods

The Segment cardiac segmentation suite by Heiberg et al. (`http://segment.heiberg.se`) [106] was used to obtain left ventricle (LV) segmentations from a set of images, kindly provided by the University of Oxford Centre for Clinical Magnetic Resonance Research at the John Radcliffe Hospital, Oxford. This set was produced on a 3.0T Siemens Tim Trio whole-body MRI scanner using a 4D (i.e. 3D+time) TrueFISP Cine MRI protocol (as in section 3.3.2), from a patient undergoing a post-PCA investigation, following a diagnosis of an Inferior MI; the set was anonymised appropriately and no other clinical or radiological details were available. The image set consisted of 25 timeframes of 8 Short-Axis (SA) slices at a resolution of 256×176 voxels, of size $1.5625 \times 1.5625 \times 8mm$. Manual segmentations of the left ventricle were provided by an expert clinician, which were used as a gold standard; this was obtained as per-slice 2D contours, drawn at 4×4 subresolution accuracy per in-slice image voxel, using the CMR42 cardiac imaging suite [107]. Full diastole was identified in timeframe 1, and full systole at timeframe 10. Data was processed using Matlab [88] / Octave [89]; images were extracted from the DICOM files using a modified version of Laszlo Balkay's DICOM reader [87]; all other processing (including extraction of contours from CMR42 files) was performed using bespoke tools created by the author for this purpose.

A set of 100 segmentations was obtained by applying normally distributed random noise of linearly increasing standard deviation, on each of the default parameters (i.e. force modifiers; see Fig. 3.6) provided by Segment for the case of SSFP MRI; the noise was generated with mean $\mu = initial\ modifier\ value$ for each parameter, and standard deviation σ taking values linearly from 0 to μ over the 100 experiments[7]. Experiment 70, which happened to be the best outcome in

[7] This is a reasonable approach for this scenario, since all parameters represent force modifiers, and was hence chosen for simplicity. A more general approach would be to establish limits for each

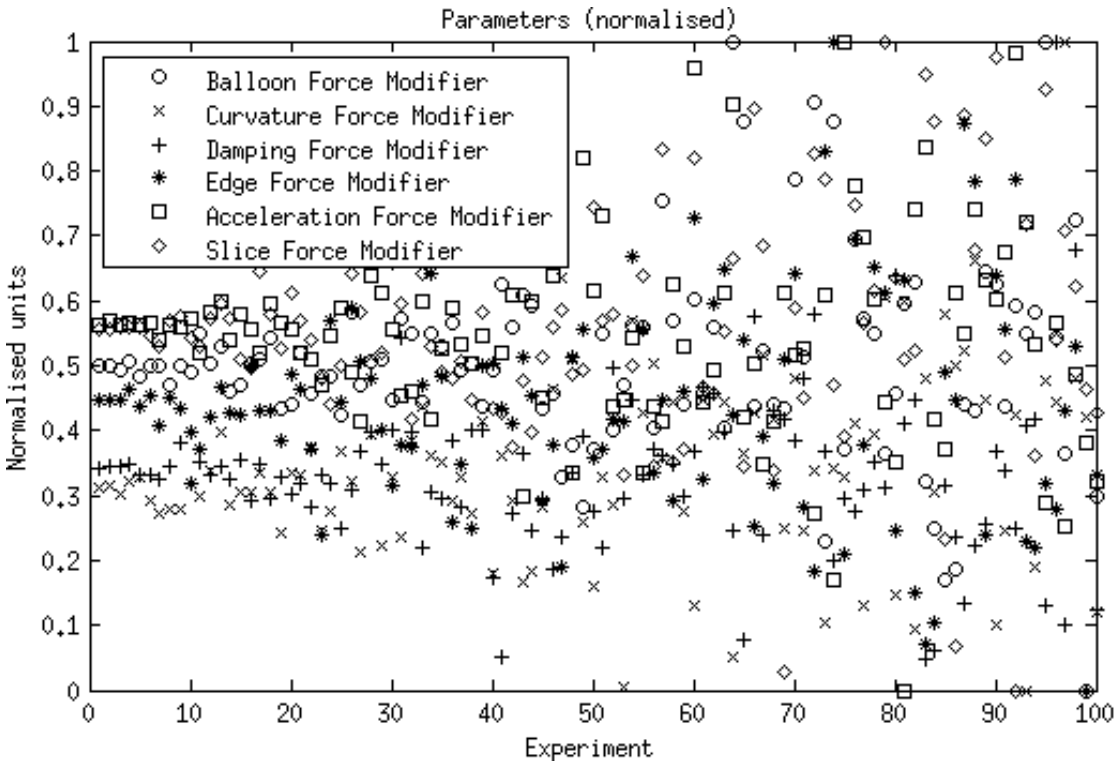


Figure 3.6: An overview of the parameter space exploration method used. Each modifier is set to the 'software manufacturer' recommended value for experiment 1, and then each subsequent experiment shifts this value randomly based on normally distributed random noise of increasing standard deviation per experiment. Modifiers are 'normalised' here (i.e. *min* and *max* values mapped to 0 and 1 respectively) for ease of comparison.

this set was retained as a reference to the best segmentation obtained with this algorithm for this particular image. For each of the resulting segmentations, the following physiological parameters were derived:

- Volumes in systole (V_s) and diastole (V_d), defined as the number of voxels in the set of LV-labeled voxels in systole (LV_s) and diastole (LV_d) respectively
- Stroke Volume (SV) $= V_d - V_s$
- Ejection Fraction (EF) $= SV/V_d$
- Centre of mass (systole): A 3D-coordinate vector $C_s = [\bar{x_i}, \bar{y_i}, \bar{z_i}]^T, \forall\ i \in LV_s$
- Centre of mass (diastole): $C_d = [\bar{x_i}, \bar{y_i}, \bar{z_i}]^T, \forall\ i \in LV_d$
- Combined centre of mass: $C = (C_d + C_s)/2$

parameter first, such that segmentation outputs at those limits results in a particular reduction in accuracy on validation (with other parameters kept constant).

Figure 3.7 shows a simultaneous graph of 'normalised' Stroke Volume and the accuracy for each of the 100 experiments, giving a graphical intuition for premise 1 above.

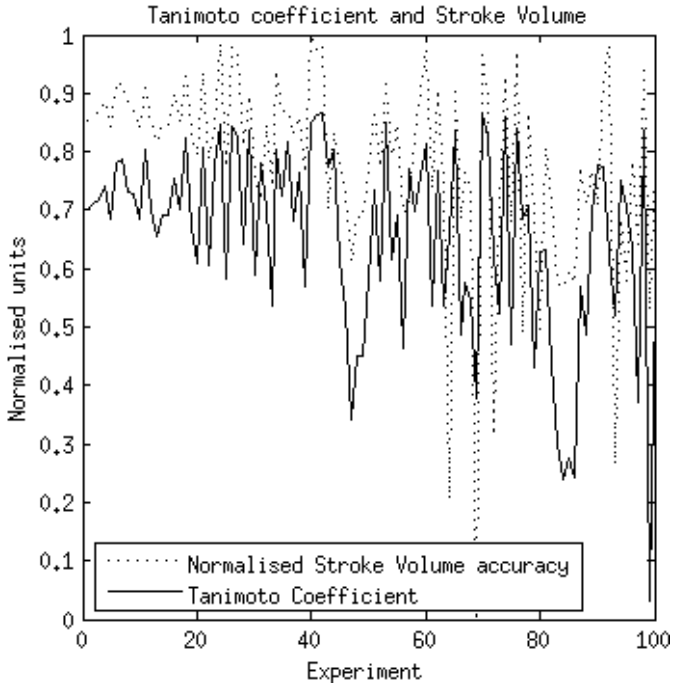


Figure 3.7: Association between a single clinical parameter (Stroke Volume) and segmentation accuracy. To make visual comparison easier, Stroke Volume is normalised to the range [0,1], such that the value that matches that of the Gold Standard maps to 1, and the value with the maximal absolute difference from the Gold Standard (in the scope of these experiments) maps to 0, with in-between values varying linearly between these two extremes. It is clear from the graph that the two curves follow similar trends.

The weight each segmentation carries within the fusion process is determined by a measure of how close each of their physiological parameters is to a reference value; in particular, a suitable range around this reference value acts as a fuzzy constraint, that prevents bad segmentations, from a physiological-estimates point of view, from exerting much influence on the fused end-result. In practice, such values might be already available clinically, i.e. from a previous / preliminary investigation (e.g. a quick echocardiogram performed during admission), or from known physiological values / clinical guidelines. However, for the purposes of this study, three different types of constraints were generated:

'Default': Reference range derived from the 'default' segmentation (i.e. the segmentation resulting from the 'default' parameter set), using the median and

inter-quartile range to define lower (l) central (c) and upper (u) reference values: This should produce the fuzzy analogue closest to the default case.

'Manual': Reference values derived from a very quick and crude initialisation process, where the user draws rough squares outside and inside the blood pool; thereby defining lower (l) and upper (u) constraint values for the reference range, with their average representing the central reference value (c).

'Optimal': Reference values optimally derived from the known Gold Standard. This should produce the best outcome which is possible from the algorithm, with respect to the known gold standard. Lower (l) and upper (u) constraint values for this case were set as $\pm$ 10% of the central value (c) for all physiological parameters, except for the distance from the centroid, which was set at the range of 0–10 voxels apart.

Weights were then calculated for each of n segmentations (S_n) from these reference values, by evaluating a fuzzy membership function on each of the estimated physiological parameters. We found that a good membership function was a normalised Gaussian membership function[8], with mean $\mu = c$ and standard deviation $\sigma = (u-l)/2$, and normalised such that the peak of the Gaussian has a value of 1. A total weight (w) was then obtained by fuzzy conjunction of all the weights; this was evaluated separately for two triangular norms: Product (involving multiplication of all terms), and Gödel (involving selecting the minimum of the set as the weight, i.e. the "weakest link"). Segmentations were then fused by a simple weighted averaging[9] process: $S_{fuzzy} = \sum_{1}^{n} w_n S_n$. If we wanted to use this result in an uncertainty-based fusion scheme as per section 3.3, we could also produce simple uncertainty maps by taking a squared-weighted average, or a 'weighted' variance, etc. For

[8]This implies a single optimal value; if an optimal *range* is indicated instead, then an appropriately constructed double sigmoid membership function could be used instead

[9]note that, in theory, a weighted average is simply a special case of a 'weighted' fuzzy union, for the case of probabilistic union of mutually exclusive and exhaustive states. We chose this fusion method here for simplicity; however, in theory more general weighted fuzzy unions could be explored.

	Default	Manual	Optimal
Product t-norm	0.6745	0.7953	0.8183
Gödel t-norm	0.6768	0.7536	0.8102

Table 3.1: Tanimoto coefficient of resulting fuzzy segmentations at diastole for the three types of reference ranges used, after thresholding at 0.5 to obtain a binary result, and as compared against the gold standard (i.e. manual segmentations), for Product and Gödel t-norms.

the purposes of validation against a binary gold standard, the fused result was thresholded at 0.5; the resulting binary segmentation mask S was validated using the Tanimoto Coefficient (eq. 2.2, p. 30).

3.4.4 Results and discussion

The accuracy of the different segmentations is shown in Table 3.1. For comparison, the Tanimoto coefficient of the original algorithm with default parameters was 0.7016; experiment 70, whose randomly selected parameter set gave the best outcome over all other experiments, yielded a Tanimoto coefficient of 0.8671. While the Fuzzy equivalent of the default parameter set seems to be a bit lower for both Product and Gödel cases, it is very close (and indeed this is also the case visually), and it is in fact a better fuzzy equivalent than the simple averaging of all 100 segmentations without weighting, which resulted in a Tanimoto coefficient of 0.6507. However, with respect to the best outcome, both the default case and its fuzzy analogues are poor by comparison. Figure 3.9 demonstrates the resulting contours, and shows the effect of our approach visually.

Effect of clinical constraints

Figure 3.8 demonstrates for all 100 segmentations involved, how the individual clinical parameters and resulting combined weights for each of the three scenarios, vary with segmentation accuracy. This graph shows why there is clear improvement when more appropriate physiological parameters are provided as constraints. In the 'Optimal' case, where physiological parameters were derived from the Gold Standard (which would be equivalent, for instance to having those parameters provided by the clinician, e.g. via a different investigation or from prior knowledge), this comes very

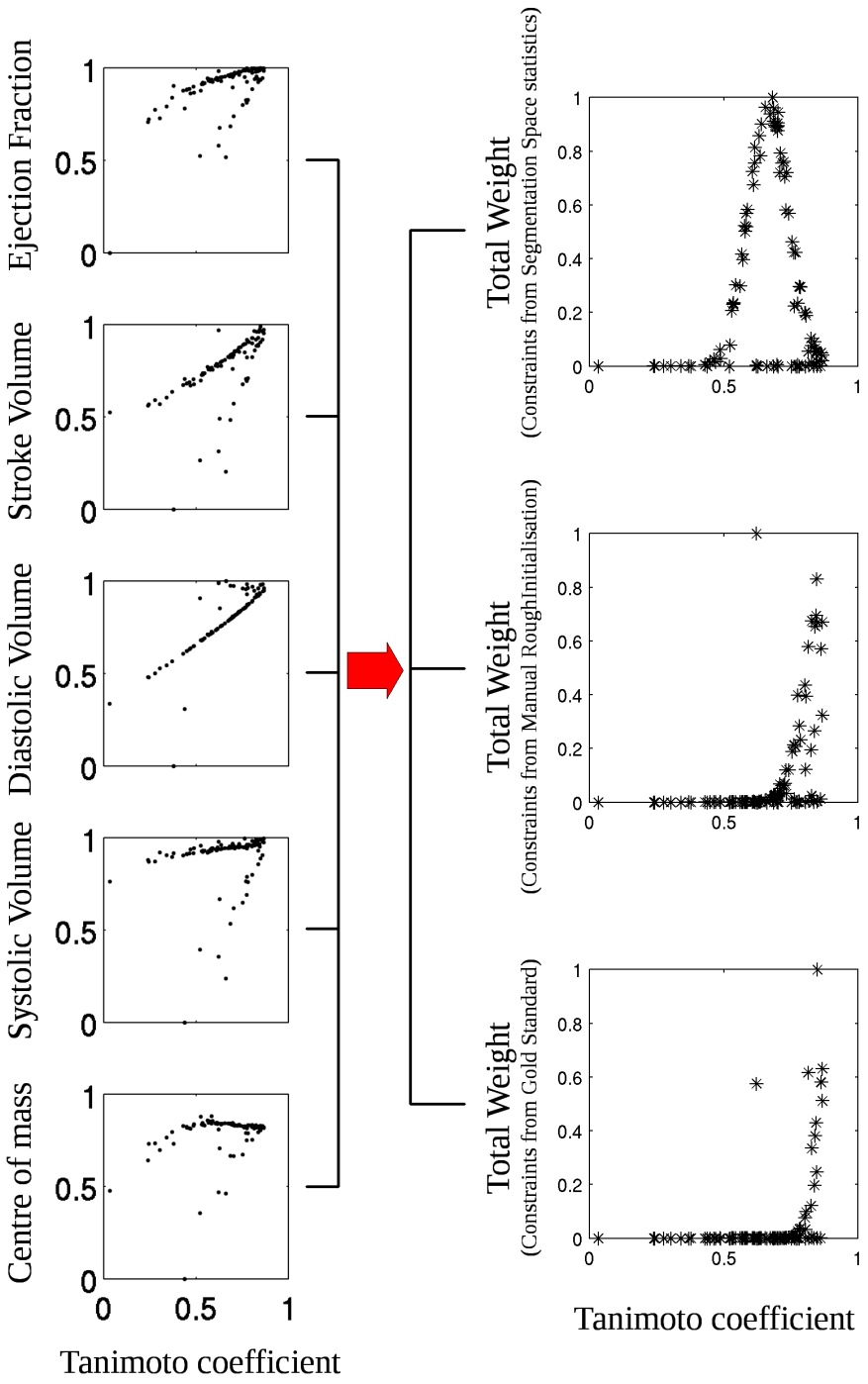


Figure 3.8: Relationship between estimated and combined clinical parameters with respect to segmentation accuracy. The left column shows how individual (normalised) clinical parameters are distributed with respect to their corresponding Tanimoto coefficient for each experiment. All clinical parameters form hysteresis curves, with a clear point of maximal accuracy, indicating that each clinical parameter contributes individually to overall accuracy, but the effect is also dependent on other clinical parameters being optimal. The right column shows how the clinical parameters combined into a single weight can favour segmentations with higher accuracy, and that the effect is improved for more appropriate choices of initial choice of clinical parameters.

close to the 'best' result of the set. Furthermore, the rough manual initialisation is not far behind in terms of accuracy. In other words, even in the absence of perfect physiological parameters, a quick, rough estimate can still lead to a markedly better result. It is worth restating that it was only possible to identify the 'best' result of the original algorithm (i.e. experiment 70) here, because validation against a gold standard was available, but in general an optimal parameter set is not something that would be readily available in real-life scenarios; therefore, in the absence of a gold standard, it would be very difficult to confidently identify any parameter set or its corresponding output as 'optimal'. Our results demonstrate that by using the more intuitively generated physiological constraints in this fashion, we can achieve similarly good results as the best possible segmentation obtained through such an optimal parameter set.

Selection of constraints in practice

In this study, we tried to demonstrate how a clinician could use clinical information, provided in the form of 'constraints', to intentionally bias an algorithm's output towards more clinically relevant results with respect to that clinical information. For the purposes of this demonstration, we simulated worst-case, reasonable-case, and best-case clinical-constraint scenarios, to demonstrate the effect clearly.

As mentioned earlier, however, in practice, there are multiple sources where these constraints could come from; we give here some example scenarios:

- *Reference values may exist in the literature*: For example, cardiomegaly is defined clinically via the cardiothoracic ratio (CTR)[10]: overt cardiomegaly is defined as a CTR $\geqslant 0.5$, with a fuzzy 'normal to overt' range (defined in terms of moderately increased clinical risk) of 0.42 to 0.5 [108], defining a complete membership function for this clinical variable. A patient with known overt cardiomegaly undergoing pre- and post-intervention scans, could

[10] defined as the ratio of the heart's widest diameter over the ratio of the ribcage at the level of the diaphragm.

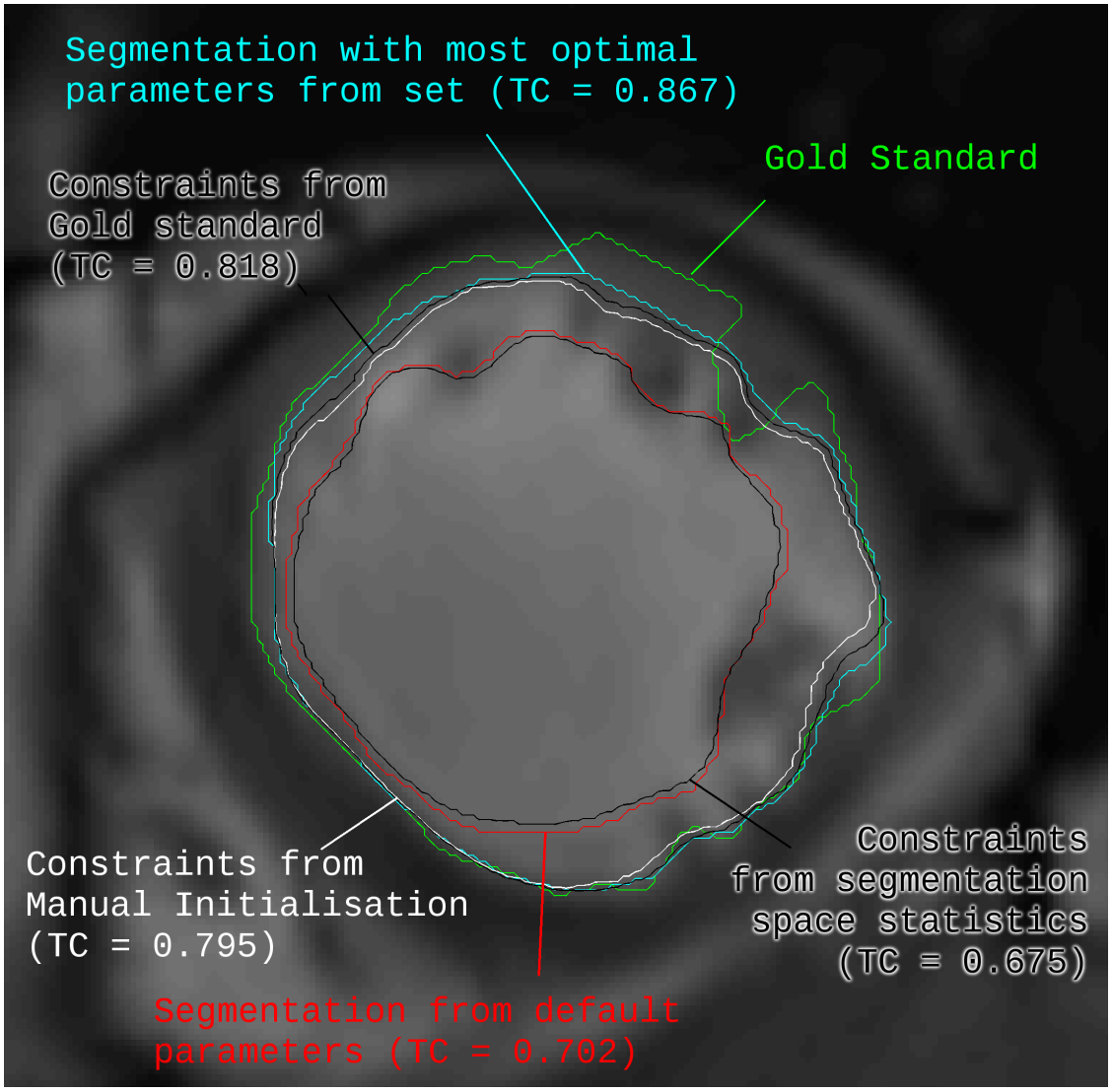


Figure 3.9: An indicative slice, demonstrating the segmentations obtained and their accuracy (TC = Tanimoto coefficient). Contours for fuzzy results correspond to the 0.5 threshold

have segmentations 'biased' towards outputs that are more consistent with a cardiomegalic picture.

- *Reference values may exist from other preliminary investigations*: There is a number of clinical variables that could be available at the time of scanning from standard, preliminary investigations performed at admission. For instance, the cardiac axis is available from a standard Electrocardiograph (ECG); preliminary Stroke Volume and Ejection Fraction measurements can be obtained from a screening bedside Echocardiogram; heart size estimates

from plain chest radiographs. All these clinical variables could be used to influence fusion towards segmentations that more closely correspond to already known clinical findings.

- *Reference values may be given by a clinician as a quick initialisation step*: Even in the absence of such pre-existing clinical information, we have demonstrated above that even *approximate* clinical information can positively influence the fusion outcome. Therefore, it is not unreasonable to ask a clinician to perform a rough initialisation from which rough estimates of clinical variables can be made, such as a rough outline of the ventricles during systole and diastole, leading to rough estimates of heart position, Stroke Volume, Ejection Fraction, etc; this is particularly the case when such initialisations may well be required by the segmentation algorithm itself.
- *Reference values may relate to simple anatomical relations*: A large amount of clinical information effectively exists as anatomical knowledge; e.g. the various compartments of the heart have intimate spatial relationships to other anatomical structures, such as the lobes of the liver, ribs and vertebrae, aorta, etc. If these landmarks can be identified (either by another automated process, or by rough initialisation), then these anatomical relationships could be expressed as constraints, and the degree to which a segmentation deviates from these anatomical relationships could be quantified. We defer the discussion of how such anatomical constraints could be defined and used for assessing segmentation quality to chapter 6.

Limitations, and global versus local constraints

Rather unsurprisingly, we note that the algorithm seems to retain its shape properties; in other words, if all segmentations in the set share common shape characteristics, the fused result is unlikely to produce a result which is structurally very different than the best result in the set. Nevertheless, since the resulting surface is biased towards having similar physiological parameters as the gold standard, the final outcome should favour surfaces that are generally closer to it.

Note, however, that this is probably also partly due to the 'global' nature of how the clinical weights were applied in this particular experiment; one could expect that more 'localised' weight-maps (i.e. at the pixel, rather than at the segmentation level) would result in more flexible fusion, as observed in the case of the Multi-Atlas Segmentation literature where such localised priors are intrinsically available due to the presence of atlases per fusion component [62]; future work could thus focus on expressing the clinical constraints in a local fashion so as to take advantage of locally-weighted fusion strategies. One such approach could be to obtain *saliency maps* (e.g. by re-evaluating a clinical variable in the presence of *partial occlusion*), a technique which is more familiar in the context of visualising convolutional neural networks [109]; that is, quantify the effect a region or single pixel has on the end clinical variable, by computing the change in the variable when that region or pixel is occluded. This would result in corresponding pixelwise weights (both positive and negative), showing the extent to which each pixel improves or worsens the clinical estimate, with respect to the provided constraint; furthermore, transforming such maps in the [0,1] range could also then enable them to be used as 'probabilistic priors' in more Bayesian-like fusion frameworks.

3.4.5 Conclusion

We have demonstrated a framework for producing a fuzzy equivalent segmentation from an existing algorithm, by exploring its parameter space to produce a segmentation space. This can then be fused in a weighted scheme, constrained by physiological parameters which can either be introduced by a clinician, much more intuitively than intrinsic algorithm parameters, or can be approximated by rough initialisation. The concept and framework can be generalised to any algorithm, such that instead of focusing on optimising intrinsic parameter sets for general cases, one would only need to explore the parameter space appropriately, and provide appropriate physiological constraints, which can be more intuitively defined, to produce better segmentations. The framework is particularly suited for medical images where the object in question has particular physiological properties

that can then be represented via a fuzzy membership function and incorporated as a constraint; heart segmentation lends itself naturally to this problem, as it offers both physiological and anatomical constraints. Further work could focus on collection of localised weight-maps as opposed to global weights; automating initialisation further, such as by using Haar features; introducing further types of physiological constraints, such as correctness of anatomical position based on other landmarks (e.g. defined as fuzzy spatial relationships of being "below the lung", "above the diaphragm" etc); improving efficiency through parallelisation or a convergent approach to the acquisition of segmentation weights.

Summary

- The Partial Volume Effect (PVE) denotes the underlying mixture of tissues within a pixel, as a result of the limited image resolution. While this is often treated as a source of noise, if modelled appropriately it can lead to more accurate models and estimates. A soft segmentation can represent partial volume effect for a particular tissue very naturally as a fuzzy value, denoting the proportion of tissue in the pixel.
- There are semantic differences between the concepts of fuzziness, probability, and uncertainty in general, in terms what kind of concept one is trying to convey (e.g. vagueness, amount of information, ambiguity, etc). One type of uncertainty can be used to quantify other types of uncertainty; this means we can naturally obtain further measures of uncertainty over the fuzzy values of a soft segmentation mask.
- This insight can be used to combine soft segmentations via the pixelwise uncertainties over their fuzzy values; we demonstrate the concept here on two simple Cine MRI sets, and uncertainty measures based on Variance and Entropy measures. Fusion based on uncertainty measures seems to outperform naive averaging / consensus approaches, however, more semantic measures of uncertainty (i.e. uncertainty denoting localised clinical / anatomical meaningfulness over a segmentation) would likely be more effective.
- We also present a framework making use of a similar notion of clinical uncertainty, relating to the discrepancy of segmentation-derived clinical estimates from an appropriately predefined range. This is used to convert deterministic segmentation algorithms to optimised fuzzy / probabilistic equivalents; the framework allows clinicians to define suitable physiological constraints over the clinical parameters, in order to guide the segmentations to more clinically relevant results. This enables a clinician to optimise a segmentation algorithm for their particular clinical setup, using clinical intuition, without further need for specialised knowledge of the particular algorithms and intrinsic parameters involved.

"Many [algorithms] in use today are not very good. There is a tendency for people to avoid learning anything about such subroutines; quite often we find that some old method that is comparatively unsatisfactory has blindly been passed down from one programmer to another, and today's users have no understanding of its limitations."

— Donald Knuth; *The Art of Computer Programming*

"The greatest obstacle to discovering the truth is being convinced you already know it."

— Ashleigh Brilliant

4

Validation theory in the context of fuzzy and probabilistic segmentations

In this chapter we introduce the main theoretical concepts underpinning the validation of medical image segmentation algorithms, and discuss limitations of current practice with respect to the emergence of 'soft' segmentation algorithms, and recent approaches towards more 'fuzzy' validation algorithms in that respect.

We then put forward a theoretical framework based on fuzzy theory and triangular norms, underlied by a semantic interpretation of fuzziness specific to the context of fuzzy mask pixels; using this framework, we demonstrate the existence of upper and lower theoretical bounds in fuzzy validation, and show that these are violated by conventional and state of the art approaches, leading to unreliable validation (and, by extension, unreliable segmentation algorithms).

Finally we introduce the concept of a 'boundary' pixel as a homogeneous fuzzy pixel with intrinsic orientation; we use this insight to propose a directional t-norm (or d-norm for short), taking pixel orientation into account, and we motivate the use of d-norms in creating more appropriate validation operators for segmentations exhibiting PVE, which we claim should improve validation precision and accuracy with respect to state of the art and conventional (i.e. thresholding-based) approaches[1].

Contents

[1]Published as: Tasos Papastylianou, Erica Dall' Armellina, and Vicente Grau. "Orientation-Sensitive Overlap Measures for the Validation of Medical Image Segmentations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 361–369

4.1 Introduction

4.1.1 What is validation? Why do we need it?

Validation is the process of deciding whether a segmentation algorithm is producing output that corresponds well to a ground truth, as represented by a gold standard. The difference between ground truth and gold standard can be subtle, leading to

the two terms occasionally being used interchangeably in the literature. In medical imaging, the ground truth refers to the actual organ or tissue of interest (e.g. the heart in heart segmentation), whereas the gold standard is our best available representation of the ground truth; this is often in the form of manual contour delineations outlining the structure of interest, as performed by an expert clinician on the available images, and is generally subject to the same or similar resolution constraints as the image itself, as well as potentially non-trivial intra- and inter-expert variability [32] — an important distinction between gold standard and ground truth.

Specifically, the process of validation attributes a numerical value to this correspondence between segmentation and gold standard. This is useful in two ways: one, it provides a measure of how successful the segmentation is, and two, and perhaps more importantly, it allows for the objective comparison of different segmentations (and by extension the corresponding algorithms that produced them). Conceptually, it could be thought of as a quantification of the algorithm's reliability or quality; in keeping with the spirit of this thesis, another way to state this, is as a statement of (un)certainty over the quality of the *algorithms* (i.e. as opposed to, say, uncertainty stemming from the image environment in which an individual segmentation object was defined). In other words, it is a numerical answer to the question "how much can I trust this algorithm?".

Furthermore, beyond its use in assessing the accuracy and precision of "finalised" segmentation algorithms, validation is also often used during the *construction* of segmentation algorithms, when these involve a "training" phase against "known" data. In this sense, validation operators may either act as a "similarity metric" (where the distinction between the two largely depends in the context they're being used, or in the type of inputs they act on), such that they optimise the parameters of a chosen model in order to maximize performance, *or* in the context of *model selection*, in terms of choosing suitable model *hyper*parameters, or even the model type itself, to guard against under- and overfitting. This is a very important point worth making explicit: **in the presence of an unreliable validation**

operator, the unreliability of a *supervised* segmentation algorithm may be compounded even further, since not only are we unable to reliably evaluate the quality of the algorithm against a suitable gold standard, but we are also unable to guarantee a reasonable degree of optimality in the resulting parameters for the chosen model, or even confidence in the model itself.

It is important to point out that the use of the term 'validation' in the literature, may either refer to the process of assessing a segmentation's similarity to a gold standard by means of an appropriate validation operator (at which point, it simply describes the quality of the particular segmentation output), *or* to the more general process of assessing an algorithm's performance in general (whether for its own sake, or for the purpose of model selection and fine-tuning of hyperparameters), by performing one or more validation operations on outputs obtained under controlled, experimental conditions, and for the purpose of obtaining a statistic, quantifying the reliability and quality of the algorithm itself. However, the two definitions tend to be used interchangeably, the particular meaning being clear from the context.

4.1.2 What are necessary / desired features for a validation algorithm?

While validation is very important, it is important to realise that it is an arbitrary operation. In a mathematical sense, validation is not something specifically intrinsic to a segmentation algorithm, but simply a general operation over a segmentation result and a gold standard, that produces output which has certain desirable properties, allowing one to describe a measure of similarity between the two. There are many ways to express this correspondence, depending on what features one wishes to highlight.

This has important implications. The choice of validation operator depends on the operator's features and the problem one tries to address. In other words, there is no single validation operator that is 'better' or universally appropriate for all classification tasks. Consider for instance a segmentation task where one

attempts to accurately locate the margins of a carcinoma to aid with surgical planning. Clearly, if one decides to judge similarity to the gold standard, by using a metric based on overlap, then a segmentation result which misses thin spokes in the gold standard will not be penalised much; but the result for the surgeon is catastrophic, since malignant tissue is left behind. Therefore, a somewhat more appropriate choice for a validation operator in this scenario, might be the Hausdorff distance [110], which is a measure of the biggest distance between 'equivalent' points in the segmentation and the ground truth; this metric would penalise the presence of such spokes appropriately, but at the cost of heavily penalising the segmentation result in its entirety, even if there was only one spoke and the rest of the segmentation was otherwise of good quality.

Conversely, if a segmentation was used to aid with the calculation of volume — such us when segmenting the ventricles of the heart during systole and diastole to obtain an Ejection Fraction — then thin spokes should not be taken too much into account, and therefore a measure like the Hausdorff distance would be inappropriate, and one might prefer an overlap metric, like the commonly used Dice [68] or Tanimoto [67] coefficients (eqs. 2.2–2.3, p. 30); if the segmentation is anticipated to have noisy positioning but in a way that doesn't affect the segmentation volume, then a functional validation operator, such as a Bland-Altman analysis [111] against clinically obtained Ejection Fractions, might be more preferable, so that slight shifts and segmentation artifacts are not heavily penalised.

Nonetheless, there are some generally desirable features that a validation operator is expected to provide; Zhang [64] identifies the following properties:

Generality: a validation operator must be general enough to be applied meaningfully to any segmentation algorithm.

Objectivity: it should not be biased by a human operator

Consistency: given the same inputs it should return the same results every time

Quantitativity: expresses outcome as a number which can be used for comparisons

Distinguishability: it should be able to detect small variations in segmentations

Complexity: (desirable) a validation algorithm should be computationally efficient

If we think of validation in the broader sense, that is, as the process of quantitatively assessing an algorithm's performance, rather than as a specific operator which returns a single scalar value, then we would add another desirable feature to this list:

Interpretability: a validation operation should provide the user with a useful general indication as to why, where, and to what extent the algorithm is failing to achieve the necessary outcome.

In other words, for the sake of scenarios like the ones given above, rather than reject an algorithm in its entirety, it would be desirable if uncertainty could be ascribed to particular behaviours of the algorithm, such that a clinician could trust certain aspects or areas of the segmentation more, and others less. This could either be used to improve on segmentation further, or it could simply provide the clinician with an improved ability and convenience to apply their clinical judgement over the interpretation of the result. For example, if a segmentation for isolating the myocardium is known to, say, be less accurate in its evaluation of the right ventricular wall due to its proximity to the liver, but is known to be reliable with respect to the left ventricular wall and interventricular septum, then this is still useful to the clinician investigating hypertrophic cardiomyopathy, who is specifically interested in interventricular septum and left ventricular wall thickness as a marker of disease [112]. Or, as we've shown in Ch. 3, this could be used in the context of algorithm merging via the uncertainty over the *algorithm* (rather than, say, an evaluation of the uncertainty over the specific mask in question), so it is very useful information.

We will talk about how to introduce such qualifiers over validation in general in subsequent sections; however for the time being, we will examine validation in the traditional sense, and deal with such qualifiers later in Ch. 6 in the context of local performance measures, spatial relationships, and validation sweeps.

4.1.3 Challenges specific to validation in medical imaging segmentation

Moving on from the theoretical difficulties inherent in the validation process of any segmentation in general, one finds that performing appropriate validation of segmentation algorithms in medical images *specifically* can be particularly challenging for a number of reasons:

- **Medical datasets tend to be of limited size.** In many cases, it is very difficult to obtain appropriate clinical datasets, as these will typically require willing human participants. This is particularly challenging in rare diseases, but even in more common conditions, the process of appropriate recruitment, consent and anonymisation or patients can prove very costly (in the broader sense of the word).

- **Clinical datasets are not necessarily optimised for research**. In the absence of a clinical trial with the explicit aim of obtaining data suitable for *computational* analysis (which is uncommon), clinical datasets such as medical images, vital signs, ECGs, etc, are typically obtained either alongside, or as a side-effect of patient treatment. Therefore, they are likely to come in a format that is not optimised for research, since they are designed for specific clinical use, rather than collection and analysis in mind. As an example, cardiac Cine MRI sets are typically collected in the aforementioned Long-Axis / Short-Axis configuration, rather than the natural coordinate system of the MRI machine; such raw data could potentially have made calculation of certain variables (such as the axis, position, and apical contour) of the heart easier to obtain and analyse; however, if no such clinical protocol exists, it is unlikely that such a dataset will be obtained.

- **There is generally a degree of ambiguity or vagueness over the Ground Truth.** That is to say, even in the presence of suitably formatted data, there is a good chance that the Ground Truth cannot be specified in an exact manner. The old adage, “ask ten doctors and you will get ten

different opinions" is an allusion to the many forms of uncertainty inherent in making a clinical diagnosis; interpretation of signs, symptoms, and anatomy, is not always black-and-white, and many decisions are made under a degree of uncertainty in one form or another. It is a great challenge to assess a computer algorithm's performance on segmenting a lung tumour from a radiograph, if a body of radiologists themselves would be at a disagreement over the tumour's location, or even its presence.

- **Clinicians often adjust their interpretation according to external information**, such as the clinical history, physical examination findings, findings from related investigations, or even just "common sense". As a trivial example, in an acute setting it is common for buttons from patients' clothing to accidentally make it into the region of clinical interest in a radiograph, and appear as bright spots. Clearly, any radiologist worth their salt would dismiss this in an instant, but one might excuse an automated algorithm trained to look for unusually positioned round shadows, for thinking it has successfully detected a "tumour".

- **Large variability exists in agreed segmentation protocols and clinical evaluation measures.** This relates partly to the point above that datasets are collected with clinical aims in mind, rather than analysis. If the protocol for defining the Ground Truth, while clinically sensible, is mathematically ambiguous, then there is no way to define perfect success for a mathematically defined algorithm. For instance, the simple protocol that "papillary muscles are to be excluded from the segmentation" opens up a can of worms for any algorithm; unless the *specific* transition from myocardium to papillary muscle can be explicitly defined (which, visually it cannot since, anatomically, this is a gradual transition, and the distinction is more in terms of function than in terms of appearance), then the criteria for how much papillary muscle to exclude from the result becomes arbitrary, and therefore so does any assessment of the algorithm's performance to that extent.

- Finally, even in the presence of all the above, there is still relative unreliability in clinically used gold standards for medical images, as **it is generally difficult to obtain a faithful, consistent representation of the ground truth**, since the Gold Standard usually defaults to manual delineation of contours around the tissues of interest, in appropriate 2D slices, by a 'medical expert'. Even ignoring the fact that contours can only be reasonably defined in 2D slices, and that there is often ambiguity over the exact borders of clinical structures in the first place, there is generally a non-insignificant intra- and inter-rater variability; for example, in [113] intra and inter-rater variability in Right Ventricle volume estimation was as high as 7% (in terms of standard deviation by volume), and a mean inter-rater difference of 5%. It is hard to reliably make claims about an algorithm's performance with absolute certainty, when the gold standard itself is relatively uncertain and unreliable.

4.2 State of the art in the validation of medical images

It is often noted in the segmentation literature, that while research on newer segmentation methods abounds, corresponding research on appropriate evaluation methods tends to lag behind by comparison [64, 65]; as discussed above, this problem tends to be further compounded in the case of medical images.

Moreover, as mentioned earlier, many of the latest approaches in segmentation have been increasingly non-deterministic, or "fuzzy" in nature [59]; this, again, is of particular importance in medical image segmentation, due to the presence of the Partial Volume Effect (PVE) [8]. However, appropriate validation methods that take fuzziness specifically into account are rarely considered, despite the fact that gold standards are also becoming increasingly "fuzzy" (e.g. via manual expert delineations at resolutions exceeding those of their corresponding datasets, or via consensus voting methods like STAPLE [59]). On the contrary, validation-wise,

most segmentation papers treat fuzziness as a 'nuisance factor' instead, as they still tend to rely on more conventional binary validation methods established from early segmentation literature, and work around the 'problem' of fuzziness by thresholding the fuzzy pixels at a (sometimes arbitrary) threshold, so as to produce the binary sets required for traditional validation. In this section we will examine why and how this traditional approach comes about, as well as alternative, state of the art approaches to the validation of fuzzy inputs (henceforth called *fuzzy validation*).

4.2.1 Traditional set-based validation measures

There is a multitude of validation approaches and distance metrics; for instance, Deza and Deza's "The Encyclopedia of Distances" [114] spans over 300 pages of distance / similarity metrics in various fields. Furthermore, some are known in different fields by different names; for example, the *Tanimoto Coefficient* (see eq. 2.2) — a metric comparing the similarity between two (non-fuzzy) sets — is also known as the *Jaccard Index*; a similar set-comparison measure, the *Dice Coefficient* (see eq. 2.3), is also known as the *Sørensen Similarity*, or *F1 Score*. Some measures are special cases of more generalised measures, such as the parameterised *Tversky index*, which generalises both the Dice and Tanimoto coefficients for different values of its parameters.

In the medical image segmentation literature, the Dice and Tanimoto coefficients are by far the most popular overlap-based metrics and the Haussdorff distance the most popular relying on contour-distance [115]; others include the Simpson (a.k.a. 'Overlap' coefficient), Cosine similarity (a.k.a. Ochiai coefficient), Kulczynski, Blanque and Anderberg coefficients [115, 116]. More generally, measures that evaluate classification performance (e.g., typically used in the context of a *Receiver Operator Characteristic* (ROC) curve) [117] such as sensitivity (a.k.a. recall), specificity, and classification 'precision' and 'accuracy' indices are also frequently used. Table 4.1 shows the definitions of these measures.

Similarity metrics	**Set-based def.**	**Component-based def.**
Tanimoto	$\frac{\lvert S \cap G \rvert}{\lvert S \cup G \rvert}$	$\frac{\lvert T_+ \rvert}{\lvert T_+ \rvert + \lvert F_+ \rvert + \lvert F_- \rvert}$
Dice	$\frac{2\lvert S \cap G \rvert}{\lvert S \rvert + \lvert G \rvert}$	$\frac{2\lvert T_+ \rvert}{2\lvert T_+ \rvert + \lvert F_+ \rvert + \lvert F_- \rvert}$
Simpson	$\frac{\lvert S \cap G \rvert}{\min\Big(\lvert S \rvert, \lvert G \rvert\Big)}$	$\frac{\lvert T_+ \rvert}{\min\Big\{\Big(\lvert T_+ \rvert + \lvert F_+ \rvert\Big), \Big(\lvert T_+ \rvert + \lvert F_- \rvert\Big)\Big\}}$
Cosine similarity	$\frac{\lvert S \cap G \rvert}{\sqrt{\lvert S \rvert \lvert G \rvert}}$	$\frac{\lvert T_+ \rvert}{\sqrt{\Big(\lvert T_+ \rvert + \lvert F_+ \rvert\Big)\Big(\lvert T_+ \rvert + \lvert F_- \rvert\Big)}}$
Kulczynski	$\frac{\lvert S \cap G \rvert}{2}\left(\frac{1}{\lvert S \rvert} + \frac{1}{\lvert G \rvert}\right)$	$\frac{\lvert T_+ \rvert}{2}\left(\frac{1}{\lvert T_+ \rvert + \lvert F_+ \rvert} + \frac{1}{\lvert T_+ \rvert + \lvert F_- \rvert}\right)$
Blanque	$\frac{\lvert S \cap G \rvert}{\max\Big\{\lvert S \rvert, \lvert G \rvert\Big\}}$	$\frac{\lvert T_+ \rvert}{\max\Big\{\Big(\lvert T_+ \rvert + \lvert F_+ \rvert\Big), \Big(\lvert T_+ \rvert + \lvert F_- \rvert\Big)\Big\}}$
Anderberg	$\frac{\lvert S \cap G \rvert}{\lvert S \cup G \rvert + \lvert S \triangle G \rvert}$	$\frac{\lvert T_+ \rvert}{\lvert T_+ \rvert + 2\Big(\lvert F_+ \rvert + \lvert F_- \rvert\Big)}$
Performance Indices		
Sensitivity	$\frac{\lvert S \cap G \rvert}{\lvert G \rvert}$	$\frac{\lvert T_+ \rvert}{\lvert T_+ \rvert + \lvert F_- \rvert}$
Specificity	$\frac{\lvert (S \cup G)^c \rvert}{\lvert G^c \rvert}$	$\frac{\lvert T_- \rvert}{\lvert T_- \rvert + \lvert F_+ \rvert}$
Precision	$\frac{\lvert S \cap G \rvert}{\lvert S \rvert}$	$\frac{\lvert T_+ \rvert}{\lvert T_+ \rvert + \lvert F_+ \rvert}$
Accuracy	$\frac{\lvert (S \triangle G)^c \rvert}{\lvert \Omega \rvert}$	$\frac{\lvert T_+ \rvert + \lvert T_- \rvert}{\lvert T_+ \rvert + \lvert T_- \rvert + \lvert F_+ \rvert + \lvert F_- \rvert}$
Tversky index	- -	$\frac{\lvert T_+ \rvert}{\lvert T_+ \rvert + \alpha\lvert F_+ \rvert + \beta\lvert F_- \rvert}$
Distance metrics		
Hausdorff distance	$\max\Big(\max_{s_i \in S} d(s_i, G), \max_{g_j \in G} d(g_j, S)\Big)$, where $d(a, B)$ is the shortest euclidean (or other norm-based) distance from pixel a to any pixel in object B	

Table 4.1: Common similarity / distance metrics and other performance indices used in the validation of medical image segmentations.

However, all the above measures, are designed to compare *sets*, and *classical* sets in particular, i.e. sets consisting of discrete elements, and crisply defined boundaries. Recall from chapter 2 that from a mathematical point of view, an image can be interpreted as a *partially ordered set* of pixels (i.e. the *image domain*, where the ordering essentially relates to the positioning of the pixels in the image), and that a segmentation denoting a particular class is simply a discrete subset of such pixels, which can be conveniently represented as a binary mask of 'ones' and 'zeros' (or 'true' / 'false') at the corresponding pixel positions. The operators mentioned above, all expect two such binary masks as inputs, to assess the extent to which the two sets are similar: in the case of overlap measures, by making use of set operations to assess the extent to which both images contain the same subset of 'true' pixels; and in the case of distance measures like the Hausdorff distance, by relying on the notion of distance between elements of partially ordered sets (in this case, this can be conveniently represented as the physical distance between pixels), to evaluate the degree of separation between particular elements of one mask with respect to the other one.

Therefore, despite the presence of so many different metrics, these are essentially all variations of the same concept of similarity between well-defined sets, which limits their application only to segmentation masks that can be adequately represented by that paradigm. In this sense, the explosive proliferation in the literature, of segmentation algorithms, both task-specific and general, that are used for medical applications, which do not necessarily strictly follow this *binary* paradigm, is left with no appropriate validation methods by which to assess them. In other words there is a disproportionate number of publications aiming to produce ever-so-slightly more "accurate" and "precise" segmentation algorithms, by making use of non-traditional, non-deterministic masks, but there is very little literature examining the ability to suitably and objectively compare the effectiveness and efficiency of such masks — and by extension, algorithms.

4.2.2 Thresholding – the conventional approach to validating fuzzy or probabilistic sets

The benefits of fuzzy segmentations (or of 'probabilistic segmentations' in the specific case that fuzziness represents probability of some sort), over standard binary segmentations, are becoming increasingly evident and sought after [59]. As we've seen, the traditional validation techniques employed, in particular the Dice and Tanimoto coefficients, which are the most commonly used, are discrete pixel overlap-based algorithms. Specifically, from an implementation point of view, this means that the operator is defined only for binary inputs, where true pixels denote foreground (i.e. object of interest), and non-true pixels denote background. Even in the case of multiple labels, these are still discrete, and therefore represented simply as separate binary masks, with the only constraint that true pixels from all labels should be mutually exclusive (i.e. for any pixel position, if one binary mask is true then all others must be false, since no two labels are allowed to represent the same pixel).

Since both the Dice and Tanimoto coefficients are set-based, in theory they could easily be extended for fuzzy inputs by using equivalent fuzzy set operations instead of binary ones; however, as we've seen, there is no single way to generalise a standard set operation to a fuzzy one, but there are families of operators instead (i.e. based on the various t-norms), each with its own underlying semantics, and each leading to different results.

Therefore, the simplest and commonest approach adopted in the literature, which partially avoids the need for such an interpretation, is that of *thresholding* a pixel; e.g. for the single-label problem (which may also be termed as a two-class problem, in the sense that pixels are separated into foreground, i.e. label of interest, and background), if a pixel can take segmentation values from 0 to 1, signifying the degree of absence or presence of the tissue in question, then fuzzy values equal to or above 0.5 are labelled as a binary '1' label, otherwise a '0' label[2].

[2]For multi-label problems, a 'main candidate' label is chosen out of all competing (non-background) labels, with respect to some criterion first. E.g. in the case where fuzziness represents *maximum a posteriori* (MAP) probability estimates for the different labels, the label with the

Occasionally a threshold other than 0.5 may be chosen empirically, denoting a measure of leniency of strictness — i.e. making it more, or less likely to assign the class label instead of defaulting to the 'background' class — or equivalently, implying a transformation over the distribution of fuzzy values in the [0,1] range in the mask, to another distribution still within the [0,1] range. However, this is still a matter of interpretation, and can be a relatively arbitrary decision; *one common danger is that such a threshold is chosen to optimally satisfy the particular validation set available* (e.g. [18]; see section 4.2.3), *rather than because it guarantees higher validation precision and accuracy* (i.e. falling victim to the logical fallacy of *affirming the consequent*[3]).

Is thresholding a suitable approach to validation?

There are a few cases where thresholding would be a reasonable and appropriate approach. For example:

- **It would make sense, if the researcher is attempting to compare a fuzzy segmentation algorithm to an older, binary algorithm**, and / or a binary Gold Standard. While this results in less 'validation resolution' than would be possible for the fuzzy case, it does however make it more directly comparable to the older (binary) algorithm, which shares the same 'validation resolution'. Particularly since, as stated above (and will be further explained below) there are many ways to obtain a validation result using a fuzzy extension of the Dice Coefficient, all of which would lead to different

highest MAP estimate is chosen as the representative candidate; if this value is higher than the prescribed threshold, the pixel is then assigned to that class, otherwise it defaults to 'background'. Note how, for K distinct labels (excluding the 'background' class), setting a threshold equal to $1/(K+1)$ is equivalent to treating the 'background' class as a simple class, competing normally for the highest MAP value; in the single-label problem this corresponds to the 0.5 threshold.

[3] The fallacious argument goes as follows: "If our algorithm is of good quality and the threshold chosen leads to reliable validation, then our algorithm will score highly on validation. Our algorithm does indeed score highly on validation. *Therefore our algorithm is of good quality and the threshold chosen leads to reliable validation*". This conclusion is fallacious, as it does not validly follow from the premises. For instance, it is entirely possible that the algorithm is in fact *bad*, but scores highly on validation purely because the threshold chosen leads to *unreliable* validation, artificially pushing the algorithm's performance towards spuriously high values.

results, depending on which aspect of fuzziness one might want to focus on, or depending on what one felt the fuzziness represents, such that they choose the appropriate fuzzy set operators.

- **Thresholding is appropriate if the underlying ground truth does indeed correspond to an all-or-nothing event** per 'pixel' (and, respectively, if the Gold Standard was appropriately compiled according to that specification); this is particularly relevant in decision-making situations. As an example, if a segmentation output was to guide a radiotherapy treatment, such that each mask 'pixel' would determine whether that particular corresponding area would be irradiated or not, then the segmentation really does need to be an 'all-or-nothing' output, since partial irradiation would be sub-therapeutic, and therefore not of use (worse still, it produces harm for no expected benefit). In other words, it is not the case that the dose could be reduced according to the fuzzy output of the segmentation result at that pixel, therefore a fuzzy output (and by extension a validation strategy which exploits the fuzziness of the output) would be inappropriate. However, in most medical applications, and particularly those involving conventional medical imaging such as CT and MRI, the need for such an 'all-or-nothing' classification is rarely the case.

- **Simply having a fuzzy outcome doesn't guarantee anything about the semantics behind the fuzziness**. As discussed earlier, voxels in a medical image, particularly at the boundaries, can be subject to the Partial Volume Effect [75], where due to the limiting nature of a medical imaging modality's resolution, a pixel may contain a mixture of tissue classes. It is tempting to assume that a fuzzy number directly represents a probability in the frequentist sense, i.e. a representation of the pixel's class composition. However this may not necessarily be the case. For example, the underlying tissues can have very different intensity profile characteristics, and any algorithm that produces a fuzzy / probabilistic value based on class intensity distributions, does not necessarily correspond to a straightforward 'mixture percentage'

for the pixel. Clearly, it would be more advantageous to attempt to model these characteristics, and put the information contained in the fuzziness to better use; however, if there is no such information available, rather than make arbitrary, oversimplistic assumptions leading to unreliably elaborate fuzzy segmentation results, it might be more realistic to treat the pixel as a single class, and use the obtained probability to collapse the classification to a single class via a thresholding operation instead. While this can still be seen as an oversimplification, this is one where the effects of that simplified assumption are visible in the output, and therefore easier in its interpretation. In other words, it is potentially easier for the clinician to apply clinical judgement over a visually simplified output, compared to a more elaborate-looking output relying on equally oversimplified assumptions. Furthermore, and more specifically with respect to validation, if there is no guarantee that the fuzzy profile of the Segmentation and the fuzzy profile of the Gold Standard represent the same underlying phaenomenon, then it may make less sense to compare the fuzzy values directly using a fuzzy validation approach. So in the absence of such information, it may simply be best to validate by threshold, as long as it clear that a high validation score essentially translates as: "I have high confidence that there is good correspondence between Segmentation and Gold Standard, such that when the Gold Standard is more or less true, then the segmentation is also more or less true". But, the "more or less true" clause should be understood to be vague by design; if the underlying fuzziness semantics differ, there is no reason to assume that the same threshold should apply equally well to both the Segmentation and the Gold Standard inputs.

However, we argue that in many cases, the thresholding approach is still used in the literature, purely by convention and in recognition of its widespread use in the literature, or at most out of a need for backwards-consistency and comparison with older literature, rather than borne out of particular consideration if it is in fact the most appropriate validation approach for fuzzy sets, and where in

fact a fuzzy validation operator would have been more appropriate. This occurs at the cost of discarding valuable information, particularly in the case where fuzziness essentially denotes a PVE.

This clinging to a conventional approach without much thought into its suitability, brings to mind the seminal 1986 paper by Bland and Altman [111] which criticised the then rampant use of the correlation coefficient as a statistical validation measure, pointing out that it was an inappropriate metric for most applications in the literature at the time, and provided an alternative measure to deal with specific cases where this wasn't appropriate.

In the subsections that follow, we will discuss whether, when, and why a thresholding operation may be inappropriate; discuss desirable properties of the specific case of fuzzy validation; discuss and critically analyse recent approaches to this problem; and propose potential avenues to improve fuzzy validation outcomes.

4.2.3 Recent approaches to fuzzy validation

There have been few instances in the literature, where researchers identified thresholding as a suboptimal approach to the validation of fuzzy segmentations, and have proposed alternatives to make better use of the information contained in the masks' fuzziness; we present here some of the more notable examples:

- **Yi *et al.*** [18] addressed this issue by treating pure and PVE pixels separately, as if belonging to distinct binary classes for validation purposes. So, for example, if an image consisted of 'blood pool', 'myocardium', and 'background' tissues, then rather than represent this state as three *fuzzy* masks, one for each tissue label, the authors would instead represent this as six mutually-exclusive, *crisp* masks, denoting:
  - a 'pure blood pool' class,
  - a 'pure myocardium' class,
  - a 'pure background' class,
  - a 'hybrid blood pool / myocardium' class,

– a ‘hybrid blood pool / background’ class,
– and a ‘hybrid background / myocardium’ class

Under this scheme, validation can then be performed using standard binary overlap operators against a corresponding Gold Standard set of three ‘pure’ and three ‘hybrid’ binary label-masks. The authors noted that by relabelling ‘hybrid’ pixels as ‘pure’ when an estimate of their mixing fraction fell under a certain threshold value (which they empirically set at 0.1), then this improved the reported classification accuracy at validation, though it was not clear whether the same treatment was applied to the Gold Standard masks as well.

While they demonstrated that this approach led to higher scores on validation for their dataset, compared to when classifying pixels only in terms of ‘pure’ labels, there was no discussion as to whether this genuinely produces a more accurate, precise, and reliable result; furthermore, it is still wasteful of information contained in pixel fuzziness, since all degrees of fuzziness / PVE within the defined threshold would be treated as a single label.

- **Chang *et al.*** [115] proposed the following framework for extending traditional validation coefficients to fuzzy inputs. Recall that the *Tanimoto coefficient* (abbr. T_c, see eq. 2.2, p. 30) can also be expressed in *classification components* form (eq. 2.15, p. 39). Therefore, Chang *et al.*’s approach in creating a fuzzy-segmentation compatible Tanimoto coefficient was to replace the binary mask definitions of $\mathbf{t}_+$, $\mathbf{t}_-$, $\mathbf{f}_+$, and $\mathbf{f}_-$ (eqs. 2.11–2.13, p. 38) with custom ‘fuzzy’ masks, derived from the fuzzy mask inputs $\mathbf{s}$ (segmentation mask) and $\mathbf{g}$ (gold standard mask). They chose to redefine the four component masks as follows:

$$\mathbf{t}_+ \equiv \begin{cases} t_{+i} = 0 & \forall i : s_i = 0, g_i = 0 \\ t_{+i} = 1 - |s_i - g_i| & \text{otherwise} \end{cases}$$

$$\mathbf{t}_- \equiv \begin{cases} t_{-i} = 1 & \forall i : s_i = 0, g_i = 0 \\ t_{-i} = 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_+ \equiv \begin{cases} f_{+i} = s_i - g_i & \forall i : s_i > g_i \\ f_{+i} = 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{-} \equiv \begin{cases} f_{-i} = g_i - s_i & \forall i : s_i < g_i \\ f_{-i} = 0 & \text{otherwise} \end{cases}$$

In other words, two pixels of equal fuzzy value are *always* given an intersection value of 1, regardless of their *actual* fuzzy value; e.g. both segmentation / Gold Standard pairs of (0.4, 0.4) and (0.8, 0.8) would be given an intersection value of 1.

However, while the authors suggested that this was a natural way to interpret fuzzy sets, they did not provide convincing theoretical justification for this, nor did they formally assess validation performance against their binary counterparts to justify their choice. We would argue that this is a rather limited interpretation of fuzziness, for a number of reasons: Firstly, these definitions seem like an arbitrary, yet highly specific interpretation of the semantics underlying pixel fuzziness. However, they are not consistent with any formal fuzzy set / fuzzy logic framework, not least because the definitions are inconsistent between themselves, if they are to be interpreted as equivalent fuzzy reimplementations of the binary '$\wedge$' operator. E.g., the definition for $\mathbf{t}_{+}$ implies that '$\wedge$' is defined as '$a \wedge b \equiv 1 - |a - b|$'. If the definition was consistent, then $\mathbf{f}_{+}$ should have been:

$$\forall i : f_{+i} = s_i \wedge \neg g_i = 1 - |s_i - (1 - g_i)|$$

but this does not correspond to the Chang *et al.* version[4]. Secondly, attributing a true positive value of 1 for *all* equal values of s_i and g_i except $s_i = g_i = 0$ seems counter-intuitive; consider for instance the case where s_i and g_i only differ from a $t_{-i} = 0$ case by a very small amount, e.g. $s_i = g_i = 0.001$; it seems very counter-intuitive that the latter case, only differing from the former case by 0.001 would be attributed a t_{+i} value of 1 rather than 0.

[4]Note that, as pointed out in section 2.4.1 (p. 37) another common binary formulation for $\mathbf{f}_{+}$ is $\mathbf{s} - \mathbf{t}_{+}$, which is generally not identical to $\mathbf{s} \wedge \neg\mathbf{g}$ for fuzzy inputs; however this formulation is also inconsistent with Chang *et al.*

We will show later that this definition is also inconsistent with PVE or geometric interpretations of fuzziness, when we discuss the geometric interpretation of t-norms.

- **Crum *et al.*** [118] proposed similar fuzzy generalisations to the Tanimoto coefficient however, they used existing axioms of fuzzy logic and fuzzy set theory to do so.

  Furthermore, Crum *et al.* specifically evaluated the performance of their Generalised Tanimoto Coefficient(s) (abbreviated as GTCs) against the traditional thresholding approach. They assessed their operator by means of a synthetic 2D 'petal' object, designed in the authors' words to be *"simple enough that an analytic expression for the overlap could be obtained, but to also feature a non-trivial border which would exhibit partial volume effects in an image of the object"*.

  The petal object is defined by the following analytical expression, expressed in polar coordinates (r, θ) as:

$$r(\theta) = r_0 + a\, sin(n\theta + \delta) \tag{4.1}$$

  where r_0 and a are constant lengths ($r_0 \geqslant a$), n is the number of petals, and δ is an angular offset.

  A set of fuzzy segmentation and gold-standard image masks were generated by creating quantized versions of the petal on an image grid, where each pixel's intensity represented the proportion of 'petal tissue'-to-'background' resulting from the quantization. Each gold-standard mask in the set had a consistent angular offset $\delta = 0$, whereas the fuzzy segmentation masks were generated for a linearly spaced range of increasing δ offsets. In other words, the overlap of a 'rotated' petal was compared to that of a 'stationary' one. For each fuzzy overlap in the set, the 'Ground Truth' overlap was calculated from the analytical expressions, and the accuracy of the GTC and the traditional threshold-based Tanimoto coefficient were compared.

The Generalised Tanimoto Coefficients were derived by reinterpreting eq. 2.2 (p. 30) *"in the light of established results from fuzzy set theory for the intersection and union of fuzzy sets (e.g., [119])"*. Namely, they stipulated that:

- for a given set of 'validation pairs' P (i.e. if we wanted to evaluate a segmentation mask $\mathbf{s}$ against a number of disparate gold-standard masks $\mathbf{g}$, such as when one has a number of manual segmentations of variable quality to one's disposal, to use for validation purposes),
- a given set of class labels L (e.g. "Blood pool", "Myocardium", and "Background")
- and given that for a particular 'validation pair' and a particular label, 'fractional amounts' A and B can be defined for that label at voxel position i of each mask, such that $\forall i : A_i, B_i \in [0, 1]$

then

- *"the fuzzy intersection is the amount of label L in common at each voxel* ***and is therefore equal to*** $\min(A_i, B_i)$*"*. [emphasis ours]
- *"the fuzzy union is the total label L at each voxel (counting the shared component only once),* ***and is, therefore, equal to*** $\max(A_i, B_i)$*"*.

and went on to define the following operators:

$$\begin{aligned} TC_F &= \frac{\sum_i \min(A_i, B_i)}{\sum_i \max(A_i, B_i)} \\ TC_{MF} &= \frac{\sum_{l\in L} \alpha_l \left[\sum_i \min(A_i, B_i)\right]}{\sum_{l\in L} \alpha_l \left[\sum_i \max(A_i, B_i)\right]} \\ TC_{PMF} &= \frac{\sum_{p\in P} \beta_p \left[\sum_{l\in L} \alpha_l \left[\sum_i \min(A_i, B_i)\right]\right]}{\sum_{p\in P} \beta_p \left[\sum_{l\in L} \alpha_l \left[\sum_i \max(A_i, B_i)\right]\right]} \end{aligned} \tag{4.2}$$

TC_F is the basic fuzzy version of the T_c for a single validation pair and label; TC_{MF} is a multi-label extension of the TC_F, where the contribution of each label to the overall coefficient value may optionally be weighted by a label-specific constant α_l (allowing, for instance, a label's contribution to be normalised by the volume of its union for that validation pair, or by some other criterion)
TC_{PMF} is a "multiple validation pairs" extension of the TC_{MF}, where the contribution of each pair to the overall coefficient value is weighted by β_p, allowing e.g. for image quality or rater ability to be taken into account.

We note a few things on the above:

- the 'min / max' formulation chosen here for "fractional" labels, corresponds to the Gödel t-norm and t-conorm (i.e. $\cap_G$ and $\cup_G$, see eq. 2.22, p. 45). Therefore, the TC_F is essentially the Tanimoto Coefficient (eq. 2.2) formulation using the Gödel norms as the intersection and union operators, and could therefore be rewritten as:

$$T_c = \frac{\left|A \cap_G B\right|}{\left|A \cup_G B\right|} \tag{4.3}$$

- the statements *"**and is therefore equal to**"* highlighted above, are only valid in a 'fractional' sense, i.e. if one does not consider how a label is *distributed* inside a pixel, but treats one rather as if it were a container one could 'pour' labels into, in order to compare label 'fractions'. However, this is a somewhat limited interpretation of partial label occupation within pixels, which disregards a label's potential distribution pattern within a pixel as a (multi-) dimensional entity.
- The TC_{MF} and TC_{PMF}, used with $\forall l : \alpha_l = 1$ and $\forall p : \beta_p = 1$, simply correspond to the sum of all intersections, over the sum of all unions in our set of pairs and labels; if we particularly use the 'components' definition for the T_c (i.e. eq. 2.15, p. 39), then we note that for $\forall l : \alpha_l = 1$ and $\forall p : \beta_p = 1$ this extension is rather trivial, as it is simply a case of

counting all T_+ pixels over the sum of all T_+, F_+, and F_- pixels in the set, i.e. there is no actual change to the formal definition of eq. 2.15, meaning this is exactly how one might calculate the T_c over such a collection in the first place. Furthermore, while this expression generalises the coefficient to multiple pairs and labels, it is not a *fuzzy* extension as such, as the only step involving processing of 'fuzzy inputs' is the TC_F subcomponent.

– For values of α_l, β_p *not necessarily equal to* 1, such that different labels and different pairs would contribute differently to the overall sum, both TC_{MF} and TC_{PMF} can be thought of as special cases of the TC_F, for appropriately adjusted *membership functions* of the constituent elements. Remember from chapter 3 that fuzziness is a largely abstract concept, and its underlying interpretation could entail, for instance, a mapping to or from the fuzzy values to, say, a probability distribution specific to each label, or, as in this case, a relative weighting between them. We consider this approach to be therefore subsumed into the whole fuzzy framework, rather than separate or independent of it, and will therefore only focus on the 'fuzzy' component of Crum *et al.*'s GTC proposal, namely the TC_F.

Crum *et al.* demonstrated increased validation accuracy for their approach compared to thresholding (although they did not consider evaluating for validation precision as well). However, other than generally stating that this captured the concept of label 'fractionality' in its entirety, they did not expand on the significance or implications of this particular interpretation of pixel fuzziness in the context of medical images, or allude to the fact that this choice was one of many possible choices from fuzzy theory. Furthermore, no evaluation of the GTC was offered for real medical datasets, despite the fact that the paper was focused on medical imaging applications. Despite these limitations, however, the paper still stands out, both for its appropriate

use of established fuzzy set theory with respect to the *particular* semantics they considered (i.e. the interpretation of fuzziness as 'fractionality'), but also particularly for its systematic approach and focus on the importance of evaluating in a quantitative manner the reliability and performance of validation operators as *algorithms* themselves, against a *latent* — in their case, analytical — ground truth, rather than merely suggesting a new validation approach purely on empirical or theoretical grounds.

4.3 Semantics of fuzziness in validation and medical image segmentation

4.3.1 What does it mean for a pixel to be fuzzy?

In classic segmentation literature, a segmentation mask $\mathbf{s}$ — and similarly, a gold-standard mask $\mathbf{g}$ — is a binary image (i.e. pixels take values in the set $\{0,1\}$) of the same resolution as the input image, where the values denote the absence or presence in the pixel, of the tissue of interest. In a fuzzy mask, pixels can instead take any value in the interval [0,1]. The underlying semantics of such a value are open to interpretation; however, perhaps the most intuitive and useful interpretation would be if pixels represented quantized portions from a higher-resolution (or an analog, or analytical) ground truth, such that there would exist a mapping from the extent to which a fuzzy pixel is occupied by the tissue in question, to a value in the range [0,1] (or possibly to a confidence interval around such a value). For example, in the simplest case of a linear mapping, a pixel with a fuzzy value of 0.56, could be interpreted as consisting of the tissue in question by 56%, and 44% background. Figure 4.1 demonstrates this graphically.

4.3.2 What does it mean for two fuzzy pixels to 'overlap'?

As fig 4.1 suggests, there is no single configuration that corresponds to a single fuzzy pixel value. Therefore, when we talk about 'overlap' between two fuzzy pixels, we are really talking about a distribution of possible overlap scenarios (and

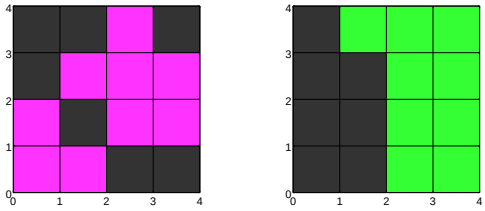


Figure 4.1: Two 'fuzzy' pixels, both represented as a 4×4 subpixel grid representing the underlying 'higher-resolution' *latent truth* in each case. Both pixels have the same resulting fuzzy value of 0.5625 (i.e. $\frac{9}{16}$), but different underlying configurations (i.e. tissue distribution inside the pixel). Purple and green subpixels correspond to foregrounds (i.e. a subpixel value of '1') respectively, and dark gray subpixels correspond to background (i.e. '0' value). The pixel's resulting fuzzy value is the average of its constituent subpixels. Note the pixel on the right is more homogeneous.

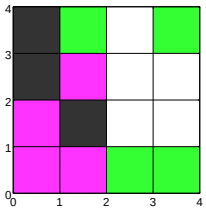


Figure 4.2: Overlap for the two pixels of fuzzy value equal to 0.5625 from fig. 4.1. While the overlap of each pixel to itself is still 0.5625 the overlap of two pixels with differing configuration is invariably less, in this case 0.3125. (Purple and green subpixels correspond to foregrounds of their respective sources, and dark gray as background class, as per fig 4.1; white subpixels signify subpixel locations where the foreground class is present in both pixels.

by extension, corresponding fuzzy values) that an 'intersection' operation could produce, depending on the number of all possible configurations consistent with the fuzzy values of the inputs. Fig. 4.3 demonstrates graphically an exploration of the resulting intersection configurations and their associated fuzzy values, in the case of all input pixel configurations possible as defined on a 3×3 subpixel grid.

Therefore, when we commit to a particular value for the intersection between two fuzzy pixels, we are implicitly making an assumption as to the particular underlying valid configurations for those pixels. For instance, the pixel that results from the intersection of two fuzzy pixels with identical configurations, will have the same fuzzy value as the two contributing pixels. This, however, is the most *optimistic*

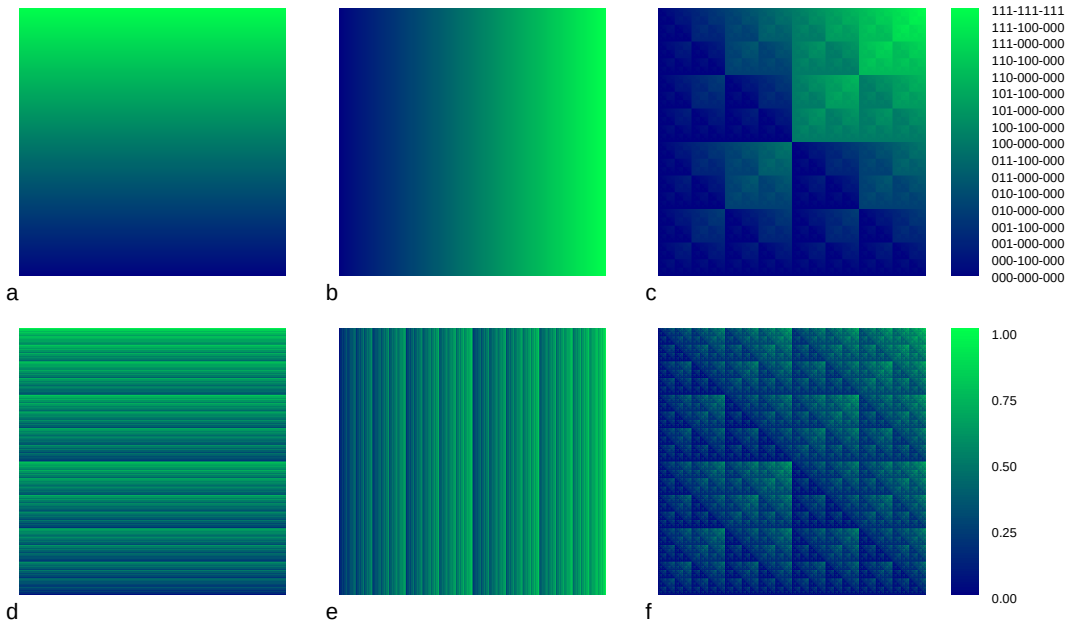


Figure 4.3: Exploration of all possible binary subpixel configurations and corresponding fuzzy values resulting from the intersection of two 'fuzzy' pixels, represented as parent pixels subdivided into 3×3 binary subpixel masks: **a)** 512 (i.e. 2^9) possible configurations of a pixel A denoting a parent pixel that consists of 3×3 subpixels, each taking a value of either '0' or '1'. Each of the 512 unique configurations is represented via a unique 'configuration index', derived from its subpixel mask. The configuration indices in A are shown increasing along the y-axis. **b)** Same for a pixel B, increasing along the x-axis instead, such that A and B define a 'reference grid' covering all possible unique configuration pairs. **c)** The resulting configuration corresponding to the intersection of A and B at each position of the above reference grid, represented using the same 'configuration indices' scheme as A and B (i.e. identical colours correspond to identical subpixel configurations in all 3 plots). **d–f)** The fuzzy values corresponding to the above configurations (obtained as the average of the constituent binary subpixels in each case)

overlap scenario, (since by definition if the two pixel configurations are not identical then their overlap will contain false negatives or false positives). Therefore, any two pixels whose underlying configuration is not identical, even when they both have the same fuzzy value, will result in an intersection fuzzy value that is lower than this optimal value (e.g. fig. 4.2).

Similarly, for any two fuzzy pixels, there is potentially one or more possible intersection configurations which correspond to the most *pessimistic* overlap outcome (in terms of the resulting fuzzy value, i.e. the number of false positives and false negatives is maximized).

These optimal and pessimal fuzzy values form *absolute upper and lower bounds* that an intersection pixel can take as its fuzzy value, while still being consistent with a "fuzziness as quantization" interpretation of fuzziness[5].

4.4 T-norms as models of tissue distribution in fuzzy pixel overlap

In section 2.4.2 we gave the definitions for three special t-norms and their dual t-conorms: the *Gödel*, *Product* and *Łukasiewicz* norms. When dealing with the overlap between fuzzy pixels, where fuzziness corresponds to the distribution of tissue within the pixel, as in the examples in the previous section, then these three t-norms are of special significance. In this section, we discuss why that is, and what useful properties and information they convey, and how this can be used to better characterise fuzzy pixels, particularly for the case of pixels at the boundary.

4.4.1 The intersection of two fuzzy pixels is a function of the amount and distribution of tissues represented within them

We remind ourselves that in this context, fuzzy pixels are defined in the frequentist sense, i.e. a pixel's fuzzy value corresponds to (or may otherwise be mapped from or to) the percentage of tissue distributed within it. Given two fuzzy pixels (for instance, one belonging to a segmentation mask, and one to a gold-standard mask) with fixed known fuzzy values (i.e. the *amount* of tissue contained in each pixel does not vary, but its distribution within the pixel may vary freely), then if we

[5]We reiterate that the 'linear' model discussed above, i.e. the notion that the fuzzy value maps *linearly* to an estimate of tissue percentage, is somewhat simplistic and will not always be appropriate, depending on the particular PVE model being considered, and how the segmentation masks were obtained. However, we will restrict the discussion that follows to the linear model, as it serves as a good baseline to explore the concepts outlined in the remainder of this thesis. In the presence of more elaborate PVE models, which could be expressed as different fuzzy mappings, with or without associated uncertainty (e.g. expressed as confidence limits or as probability distributions over the mapped fuzzy values), application of the following concepts would simply involve a pre-processing step first, applying the particular model to first obtain suitable estimates of tissue percentage (with appropriately propagated uncertainty) before proceeding with validation.

could gain *precise knowledge* of the tissue distribution inside the two pixels, we would also be able to precisely calculate the extent to which these overlap when calculating the intersection between the two.

However in general, such detail is unknown. We might therefore ask different questions instead; for example, "Given any two fuzzy pixels, then allowing for all possible configurations (i.e. the manner in which tissue is distributed within each pixel), which are consistent with the pixels' fuzzy values, what is:

- the maximum (fuzzy) intersection value possible between the two?"
- the minimum intersection value possible between the two?"
- the expected intersection value (i.e. the average intersection value one would expect over all possible configurations)?"

Two fuzzy pixels are said to achieve maximal / optimal overlap, when the tissues within them are optimally distributed with respect to one-another, such that their overlap is maximised. Equally, two fuzzy pixels are said to achieve minimal / pessimal overlap, when tissues within them are pessimally distributed with respect to each other, such that their overlap is minimized. But how can we model this, so as to attempt an answer to the above questions?

4.4.2 Modelling latent truth as superresolution

We can model tissue distribution within a pixel, by first subdividing the pixel into a grid (i.e. create a set Ω consisting of N subpixel elements), and then denoting the presence and distribution of tissue within the pixel by labelling the resulting subpixels appropriately, (i.e. define an appropriate subset of Ω representing tissue presence). In other words, we define a binary mask of a particular resolution over the pixel. For such a grid of $|\Omega| = N$ elements, there are 2^N different underlying configurations that can be represented via this grid, each yielding one of $N+1$ possible fuzzy values (i.e. $\{0, {}^1\!/_N, \ldots, {}^N\!/_N\}$) that can be represented at this level of 'inner' resolution; as $N \to \infty$ the configuration of *true* subpixels could theoretically capture the true tissue distribution and amount inside the pixel with infinite precision.

Another way of expressing this is saying that for a known *absolute latent truth* (see section 2.4.1) — i.e. when the exact tissue distribution within a pixel is known — such a superresolution approach can only express an approximate representation L_Ω of the latent truth, as limited by the choice of N, but such that as N increases this representation increases in fidelity, and becomes a faithful representation L_∞ of the absolute latent truth in the limit $N \to \infty$ (i.e. tissue distribution within the pixel can be represented in an exact / precise manner).

4.4.3 Establishing theoretical upper / lower bounds, and expectation in fuzzy intersection

We will now use the above model to provide proofs that the answers to the questions asked earlier, are in fact given by the three special norms mentioned previously.

Theorem 1: The Gödel t-norm ($\cap_G$) represents the *maximal* / *optimal* intersection between two fuzzy pixels

Proof:

Let Γ_A and Γ_B be two fuzzy pixels, with values $\mu(\Gamma_A)$ and $\mu(\Gamma_B)$ respectively denoting tissue 'coverage', where the tissue is assumed to be distributed according to some underlying configuration within the pixel.

Let Ω be a classical set with cardinality N, representing a grid subdivision of N elements as described above, and A and B be two arbitrary subsets of Ω, each modelling an underlying configuration for Γ_A and Γ_B respectively, such that it applies that $\mu(\Gamma_A) = \frac{|A|}{N}$ and $\mu(\Gamma_B) = \frac{|B|}{N}$.

The *maximal* fuzzy intersection between Γ_A and Γ_B (denoted $\Gamma_A \cap_{Opt} \Gamma_B$), as modelled by A and B, is equivalent to choosing suitable (i.e. *optimal*) underlying configurations for A and B, such that the cardinality of their intersection (i.e. $|A \cap B|$) is maximal.

We start first by examining the case where $|A| \leqslant |B|$. Optimality occurs when A is a strict subset of B, (i.e. $A \subseteq B$), since any element in A that is not also an element in B would necessarily result in a reduction in the number of elements contained in

their intersection. Therefore, under these conditions, $A \cap B = A$. Since A, B, and N were arbitrarily chosen subject to the constraints above, this result holds generally for any optimal configuration of A and B, or value of N (including $N \to \infty$), for which the constraint $\mu(\Gamma_A) = \frac{|A|}{N},\ \mu(\Gamma_B) = \frac{|B|}{N}$ holds. Therefore, under conditions of optimality, the intersection of two fuzzy pixels Γ_A and Γ_B given $\mu(\Gamma_A) \leqslant \mu(\Gamma_B)$, is identical to Γ_A, and therefore shares the same fuzzy value. Formally:

$$\mu(\Gamma_A \cap_{Opt} \Gamma_B) = \mu(\Gamma_A), \quad \text{for } \ \mu(\Gamma_A) \leqslant \mu(\Gamma_B)$$

It is clear from the symmetry of the proof that for $|B| \leqslant |A|$, the reverse holds, i.e.

$$\mu(\Gamma_A \cap_{Opt} \Gamma_B) = \mu(\Gamma_B), \quad \text{for } \ \mu(\Gamma_B) \leqslant \mu(\Gamma_A).$$

In other words, under conditions of optimality, the fuzzy value of the intersection is equal to the fuzzy value of the pixel with the smallest fuzzy value, and therefore equivalent to the application of the Gödel t-norm, i.e.:

$$\mu(\Gamma_A \cap_{Opt} \Gamma_B) = \min\Big(\mu(\Gamma_A), \mu(\Gamma_B)\Big)$$

and therefore

$$\Gamma_A \cap_{Opt} \Gamma_B = \Gamma_A \cap_G \Gamma_B \ \blacksquare$$

We immediately note, that for certain fuzzy pairs (e.g. 0.7 / 0.6) the intersection of two fuzzy pixels post-thresholding can result in overlap values that exceed this theoretical maximum.

Theorem 2: The Łukasiewicz t-norm ($\cap_L$) represents the *minimal* / *pessimal* intersection between two fuzzy pixels.

Proof:

Let Γ_A, Γ_B, Ω, N, A and B be defined as in Theorem 1 above.

We start by first examining the case where $|A| + |B| < N$. Since there are more subpixel 'slots' in Ω than all the elements of A and B combined, it is possible for A

and B to be configured in such a manner such that there is no overlap at all. In other words, under these conditions, a minimal overlap scenario occurs whenever the elements of A and B are pesimally distributed in such a manner so as to be *mutually exclusive* (i.e. their intersection is zero). In the special case where $|A| + |B| = N$, any mutually exclusive configuration of A and B will also be *mutually exhaustive* (i.e. $A \cup B = \Omega$). With respect to the corresponding fuzzy values, we have that

$$\mu(\Gamma_A \cap_{Pes} \Gamma_B) = 0 \quad \text{for } \mu(\Gamma_A) + \mu(\Gamma_B) \leqslant 1 \tag{4.4}$$

Next we consider the case where we have two mutually exclusive and exhaustive sets A and B as above, and also a modified set $\widetilde{A}$, formed from A but with the addition of m extra elements (i.e. $|\widetilde{A}| = |A| + m$), and similarly a modified set $\widetilde{B}$ formed from B with the addition of n extra elements. Because A and B are mutually exclusive and exhaustive, any extra elements added to A will necessarily be a subset of B, and any extra elements added to B will necessarily be a subset of A. Therefore starting from the pessimal overlap scenario of A and B, the intersection of any set $\widetilde{A}$ and any set $\widetilde{B}$ formed from A and B in this manner, will result in a necessary overlap formed exclusively by the union of these extra elements introduced from either set (i.e. $\widetilde{A} - A$ and $\widetilde{B} - B$). Therefore:

$$\widetilde{A} \cap \widetilde{B} = (\widetilde{A} - A) \cup (\widetilde{B} - B)$$

We note that, $(\widetilde{A} - A)$ and $(\widetilde{B} - B)$ are also mutually exclusive, from the way they were obtained. Therefore:

$$(\widetilde{A} - A) \cup (\widetilde{B} - B) = \widetilde{A} + \widetilde{B} - (A \cup B) = \widetilde{A} + \widetilde{B} - \Omega$$

and therefore

$$\mu(\Gamma_{\widetilde{A}} \cap_{Pes} \Gamma_{\widetilde{B}}) = \frac{|\widetilde{A}|}{N} + \frac{|\widetilde{B}|}{N} - \frac{|\Omega|}{N} = \mu(\Gamma_{\widetilde{A}}) + \mu(\Gamma_{\widetilde{B}}) - 1 \tag{4.5}$$

From equations 4.4 and 4.5 we have:

$$\begin{aligned} &\mu(\Gamma_A \cap_{Pes} \Gamma_B) = 0 && \text{for } \mu(\Gamma_A) + \mu(\Gamma_B) - 1 \leqslant 0, \quad \textit{and} \\ &\mu(\Gamma_A \cap_{Pes} \Gamma_B) = \mu(\Gamma_A) + \mu(\Gamma_B) - 1 && \text{for } \mu(\Gamma_A) + \mu(\Gamma_B) - 1 > 0 \end{aligned}$$

which is equivalent to:

$$\mu(\Gamma_A \cap_{Pes} \Gamma_B) = \max\Big(0,\ \mu(\Gamma_A) + \mu(\Gamma_B) - 1\Big)$$

and therefore

$$\Gamma_A \cap_{Pes} \Gamma_B = \Gamma_A \cap_L \Gamma_B \ \blacksquare$$

We immediately note, that for certain fuzzy pairs (e.g. 0.7 / 0.4) the intersection of two fuzzy pixels post-thresholding can result in overlap values that fall below this theoretical minimum.

Theorem 3: The Product t-norm ($\cap_P$) represents the *expected* intersection, i.e. the average intersection between two pixels over all possible overlaps from all underlying configurations consistent with their fuzzy values

Proof:

Let Γ_A, Γ_B, Ω, N, A and B be defined as in Theorems 1 & 2 above, $\mathbf{a}$, $\mathbf{b}$ and $\boldsymbol{\omega}$ be the binary masks corresponding to A, B, and Ω, and x be an arbitrary element in Ω.

For any specified value $\mu(\Gamma_A)$ — and similarly for $\mu(\Gamma_B)$ — the number of compatible configurations for A such that $\frac{|A|}{N} = \mu(\Gamma_A)$, is equal to the number of different ways of choosing $|A|$ out of the N elements in Ω, which is given by the binomial coefficient, i.e. $\binom{N}{|A|} = \frac{N!}{|A|!\,(N-|A|)!}$.

Similarly, the number of compatible configurations containing a *particular element* x, is equal to the number of different ways of choosing the remaining $|A|-1$ elements out of the remaining $N-1$ elements in Ω, i.e. $\binom{N-1}{|A|-1} = \frac{(N-1)!}{(|A|-1)!\,(N-A)!}$.

Since A and B are independent, there are $\binom{N}{|A|}\binom{N}{|B|}$ unique configurations consistent with the intersection $A \cap B$, where $\binom{N-1}{|A|-1}\binom{N-1}{|B|-1}$ of these configurations contain x in the intersection, for any arbitrary $x \in \Omega$. This means that for any element $x \in \Omega$ and fixed configuration of A and B, the expected value for subpixel x (e.g.

in terms of the corresponding binary mask $\mathbf{a} \wedge \mathbf{b}$) is:

$$\begin{aligned}\mathbb{E}[x] &= 0 \cdot P(x \notin A \cap B) + 1 \cdot P(x \in A \cap B) \\ &= 0 + \frac{\binom{N-1}{|A|-1}\binom{N-1}{|B|-1}}{\binom{N}{|A|}\binom{N}{|B|}} \\ &= \frac{|A|}{N} \cdot \frac{|B|}{N}\end{aligned}$$

Since $\mu(\Gamma_A \cap \Gamma_B)$ is equal to $\frac{|A \cap B|}{N}$, and all subpixels have the same expected value $\mathbb{E}[x]$ we have

$$\begin{aligned}\mu(\Gamma_A \cap_{Exp} \Gamma_B) &= \frac{\mathbb{E}\Big[|A \cap B|\Big]}{N} \\ &= \frac{\sum \mathbb{E}[x]}{N} \\ &= \frac{\cancel{N} \cdot \mathbb{E}[x]}{\cancel{N}} \\ &= \frac{|A|}{N} \cdot \frac{|B|}{N} \\ &= \mu(\Gamma_A) \cdot \mu(\Gamma_B)\end{aligned}$$

and therefore:

$$\Gamma_A \cap_{Exp} \Gamma_B = \Gamma_A \cap_P \Gamma_B \blacksquare$$

4.5 Fuzzy pixels at the boundary: a geometric interpretation

Pixels at object boundaries commonly exhibit PVE, which can manifest in their corresponding segmentation masks as fuzziness. Assuming an interpretation of fuzziness as quantization (and with particular respect to a base, 'linear' model, as discussed in the previous chapter), we made the case earlier, that for any pixel with a particular fuzzy value, there is a high number of compatible configurations that can yield that fuzzy value. For example, for a fuzzy pixel represented as a 4×4 subdivision and a fuzzy value of $\frac{9}{16}$, as per figs 4.1 and 4.2 (p. 125), there are $\binom{16}{9} = 11440$ unique configurations that are consistent with that fuzzy value for such a grid.

When we talk about fuzzy pixels in general, then all these configurations are equally likely to occur. However, in the specific case of fuzzy pixels at the boundary, this is not the case, as one generally expects these pixels to be homogeneous rather than heterogeneous (i.e. there is a single, clear separation boundary of foreground and background classes within the pixel), and therefore there are certain configurations that are less likely to occur in the context of a boundary pixel. For instance, for a pixel with a value of 0.5625, the pixel on the right in fig 4.1 is far more likely to be the case at the boundary than the pixel on the left.

If necessary, we can quantify the extent to which a particular fuzzy pixel configuration is compatible with a mask pixel occuring at the boundary, by defining an appropriate measure of "heterogeneity", e.g. based on a simple application of the Markov Property (fig. 4.4). A similar "neighbour-compatibility" measure could also be expressed for a shared edge between any two fuzzy pixels, quantifying the degree to which two fuzzy pixel configurations are compatible with being neighbouring pixels within the context of a continuous object boundary. Such measures could be used, for example, in an optimization framework to produce superresolution masks that are reasonable models of the latent truth[6].

[6]The term 'Fuzzy methods' in the context of the medical image segmentation literature, tends to be associated with a very particular 'bag' of segmentation algorithms, particularly those making use of concepts like *fuzzy affinity*, *fuzzy connectedness*, *fuzzy c-means* etc. So far, we have been using the term 'fuzzy' throughout this thesis in a much more general way. However, it is worth distinguishing between fuzzy affinity and something like the 'neighbour-compatibility' measure proposed here, for clarity. Fuzzy affinity is often used as a metric guiding a segmentation algorithm, relating to what is often termed the "hanging-togetherness" of two pixels, i.e. the extent to which two neighbouring pixels in a grayscale image belong in the same context (e.g. tissue). In the context of fuzzy affinity, the term homogeneity refers to region homogeneity, i.e. the extent to which the two pixels in question have grayscale intensities that are comparable to the average intensity of the object in question. In other words, fuzzy affinity is a score of the extent to which the two pixels belong to the same fuzzy segmentation object, given the difference in intensity relative to each other, and relative to their broader neighbourhood, and is intended to assist in the creation of a fuzzy segmentation mask, but without making any assumptions about the underlying pixel configurations of the latent model represented by that mask. By way of contrast, the 'homogeneity' and 'neighbour-compatibility' measures proposed here score just that: the potential configurations (which are an attempt to model / represent the *latent truth*) that can be generated from an existing fuzzy mask, in terms of how likely they are, collectively, to correspond to a valid latent truth, under the assumption that individual pixels are more likely to be (internally) homogeneous, and that transitions between pixels is more likely to be smooth.

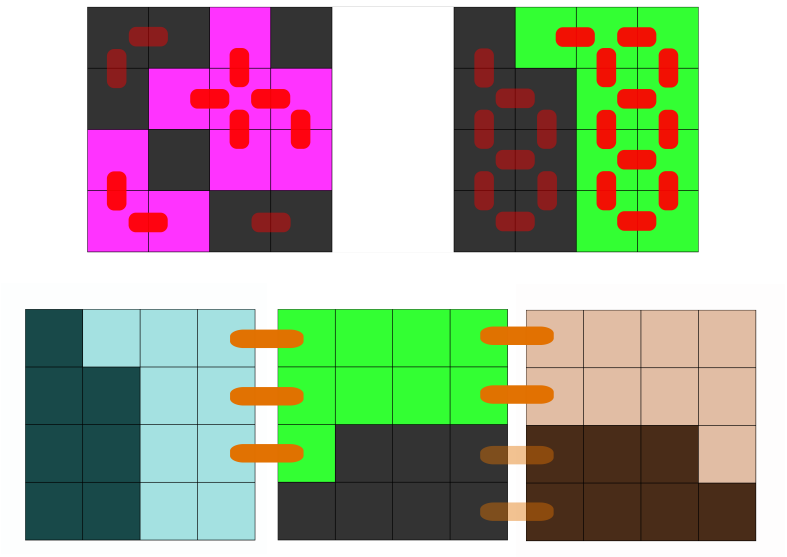


Figure 4.4: Homogeneity, and neighbour-compatibility in fuzzy pixels. Different colours signify different pixels (brighter colours signify foreground, darker colours signify background). **Top:** A simple way to quantify the 'homogeneity' of a fuzzy pixel configuration, as the number of valid 2-pixel *cliques* between 'compatible' subpixels (shown here via the red connecting lines). The pixel on the right is more homogeneous than the left, manifesting in a larger number of valid cliques (i.e. 19 cliques) compared to the pixel on the left (10 cliques). **Bottom:** A simple way to quantify 'neighbour compatibility' between fuzzy pixel configurations, as the number of valid 2-pixel cliques between compatible subpixels occurring only at the pixel edge (shown here as orange lines). An algorithm seeking to obtain meaningful boundary pixel configurations could seek to maximize these two measures.

4.5.1 Boundary pixels are homogeneous fuzzy pixels exhibiting a particular 'orientation'

This homogeneity of fuzzy pixels at object boundaries, now offers a geometric interpretation of fuzziness. A *boundary pixel* can be thought of as a homogeneous fuzzy pixel, where there exists a line or curve representing the ground truth boundary, cutting through the pixel, and cleanly separating it into foreground and background components, giving it a fuzzy value equal to the area of the foreground component relative to the total pixel area.

The simplest geometric representation of such a pixel is that of one divided by a *straight* line which best matches the general orientation of the object boundary at that point, splitting it into foreground and background regions. Therefore, we can now speak of a fuzzy, boundary pixel as having an intrinsic orientation; we

define the *boundary pixel orientation* for this specific case of a line-separated pixel, to be the outward direction perpendicular to this straight line (in other words, the negative ideal local image-gradient). Conceptually, this can be thought of as the main direction in which the foreground would be "expanding" towards the background within that pixel. This kind of representation allows us to move past the limitations of having to represent a fuzzy pixel as a grid subdivision (which is generally a very computationally expensive approach), and instead represent a boundary pixel as a geometric object that can be wholly described by a magnitude and a direction, i.e. a simple vector quantity (fig 4.5).

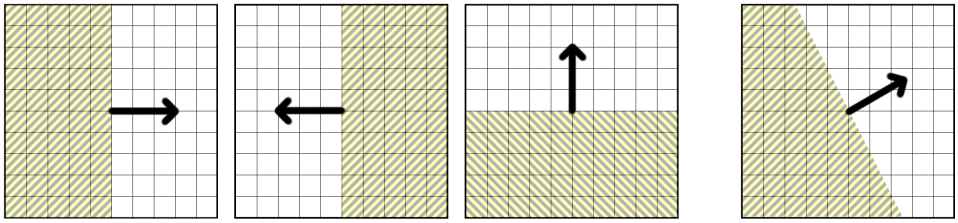


Figure 4.5: Boundary pixels with a magnitude (i.e. fuzzy value) of 0.5, representing the total foreground area relative to the total pixel area, for varying orientations. The arrows signify that this can be described as a vector.

4.5.2 The extent of overlap between two boundary pixels is a function of their relative orientations

It is easy to see that, given a segmentation mask and a gold-standard mask, fuzzy pixels at the boundary will display optimal overlap when their corresponding orientations are congruent (i.e. the foreground 'leading fronts' within the pixel are parallel, and 'expanding' towards the background in the same direction) as this ensures maximal coverage. Similarly, pessimal overlap occurs when they are completely *in*congruent (i.e. parallel, but coming from opposite directions).

In general, in the context of boundary pixels and this particular vector representation of fuzziness, for two mask pixels with fixed fuzzy values, the fuzzy value of the intersection between these two pixels is a function of the relative angle of their orientations, which is monotonically non-increasing from the optimal / upper bound

— given by the Gödel t-norm (see section 4.4.3) — towards the pessimal / lower bound — given by the Łukasiewicz t-norm — as the absolute angle difference between them increases from 0° through to 180°. Fig. 4.6 demonstrates this for the particular case of a 2D square pixel (also making the case for the special geometric significance of the Product t-norm in this case).

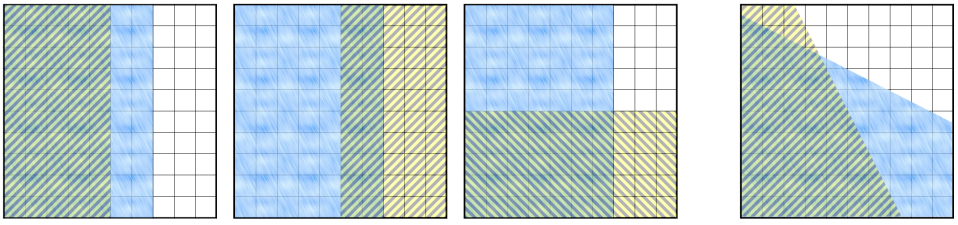


Figure 4.6: Overlap at the boundary, for the particular case of square (2D) pixels: Foreground corresponding to the gold-standard mask **g** is shaded using thin blue lines, and foreground corresponding to the segmentation mask **s** is shaded using coarse yellow lines (same as fig 4.5); background is unshaded. In all cases, gold-standard foreground area covers 70% of the pixel (i.e. a fuzzy value of 0.7), and the segmentation foreground covers 50% (i.e. a fuzzy value of 0.5). From Left to right: **g** and **s** have the same orientation; **g** and **s** have opposite orientations; **g** and **s** have perpendicular orientations; **g** and **s** exhibit arbitrary orientations. Their intersections have pixel coverages of 50%, 20%, 35%, and 46.2% respectively. Note that the first three cases correspond to $\mathbf{s} \cap_G \mathbf{g}$, $\mathbf{s} \cap_L \mathbf{g}$, and $\mathbf{s} \cap_P \mathbf{g}$ respectively.

We can deduce from this that in the context of boundary pixels as described above, for any two overlapping boundary pixels whose absolute orientation angle difference is between 0° and 180°, there exists a suitable intersection operator which returns a value between the most optimal (i.e. $\cap_G$) and most pessimal (i.e. $\cap_L$) value, and which decreases monotonically between these two limits as this absolute angle difference increases within that range. Furthermore, as fig 4.6 suggests, for a particular pixel of known shape and dimensionality, this can be calculated exactly in an analytical fashion. We define such an operator as a *Directional t-norm* ($\cap_D$), and its dual as a *Directional t-conorm* ($\cup_D$), or simply *d-norm* and *d-conorm* for short. We will demonstrate two implementations of such a norm in the next section: a *context-specific* formulation with respect to known pixel shape and dimensionality, and a *generalised* formulation, independent of it.

4.6 'Directional' t-norms: modelling overlap in oriented boundary pixels

We distinguish between two kinds of d-norms:

- **context-specific d-norms**, which aim to model the intersection between two boundary pixels in an exact manner, with respect to a known geometry (e.g. dimensionality and shape) for the containing pixel
- **generalised d-norms**, which aim to model the relationship between relative angle difference and optimality / pessimality in a more general fashion, independent of any particular pixel context.

4.6.1 A context-specific directional t-norm

As illustrated by the shaded area in fig. 4.6, for any two boundary pixels (modelled as 'line-separated' pixels as described in the previous section), a d-norm can be calculated *exactly*, if we assume the exact shape and dimensionality of a particular pixel to be known and relevant to the problem. In other words, given a fuzzy value and orientation for a pixel, its shape dictates the exact manner in which such homogeneous, intrinsically-oriented tissue is distributed within it. An intersection operation between two boundary pixels, whose fuzzy values and orientations are both known, can therefore be calculated *exactly* in an analytical manner, but the result is specific to that particular pixel shape. Table 4.2 gives an algorithmic implementation for such an exact d-norm for the specific context of a 2D isotropic (i.e. square) pixel — a corresponding Octave / Matlab implementation is provided in appendix A.1. Fig. 4.7 shows the profile of this particular 2D square-pixel d-norm for a range of fuzzy inputs and angle differences.

4.6.2 A generalised directional t-norm

Clearly, while context-specific d-norms could be reasonably expected to be more accurate, this dependence on exact pixel shape adds an extra layer of complexity, which may be undesirable. To a large extent, the main benefit of a d-norm is

Algorithm: 2D isotropic boundary pixel directional t-norm

Inputs:

- A 2D isotropic pixel $\mathbf{s}$ with fuzzy value f_s and orientation unit vector γ_s
- A 2D isotropic pixel $\mathbf{g}$ with fuzzy value f_g and orientation unit vector γ_g

Output: A fuzzy value corresponding to the d-norm $\cap_D$ of the two inputs

Steps:

1: **if** either of f_s or $f_g \in \{0, 1\}$, *or* either of γ_s or γ_g is zero:
return $f_s \cap_G f_g$ and then **exit**. (where $\cap_G$ is the Gödel t-norm)

2: For pixel $\mathbf{s}$, identify points s_1, s_2, and s_0, such that:
s_1 and s_2 are points on the pixel's perimeter, forming a line segment $\{s_1, s_2\}$ that dissects the pixel into two components A_s and $A_{s'}$, such that γ_s is normal to this line, and the area of A_s (i.e. the component to the opposite side to the one pointed to by γ_s) is equal to f_s.
s_0 is one of the pixel's four corners (i.e. $s_0 \in C$), corresponding to the 'origin of expansion', which is the first corner that would be encountered if a linear-front parallel to $s_1 s_2$ was moving from outside towards the pixel, along the direction of γ_s.
For pixel $\mathbf{g}$, likewise identify equivalent points g_1, g_2, and g_0

3: **Let** C be the set $\{c_{00}, c_{10}, c_{01}, c_{11}\}$, corresponding to the four corner points of the pixel, located at coordinates $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ respectively.
Let c_{sg} be the centroid of $\{s_1, s_2, g_1, g_2\}$
Let K be initialised to an empty set '$\{\}$' of 'contributing points'

4: Find whether / how lines $\{s_1, s_2\}$ and $\{g_1, g_2\}$ intersect:
Case 1 — the two lines coincide:
if $\gamma_s = \gamma_g$: $K \leftarrow \{s_1, s_2, \forall x_i \in C \mid x_i \in A_s\}$ (where $\leftarrow$ denotes the *append* operation)
else (implying $\gamma_s = -\gamma_g$): **return** 0 and **exit**
Case 2 — line segments do *not* intersect:
if line segment $\{s_1, s_2\}$ is farther from s_0 than c_{sg}: $K \leftarrow \{g_1, g_2\}$
if line segment $\{g_1, g_2\}$ is farther from g_0 than c_{sg}: $K \leftarrow \{s_1, s_2\}$
if $K = \{\}$: **return** 0 and **exit**
else: $K \leftarrow \{\forall x_i \in C \mid x_i \in A_s \cap A_g\}$
Case 3 — line segments intersect inside the pixel:
Let g_{in} be one of g_1 and g_2, such that g_{in} would be encountered first for a line parallel to $\{s_1, s_2\}$ sweeping in the direction γ_s.
Let s_{in} be similarly defined as for g_{in}, with respect to a line parallel to $\{g_1, g_2\}$ sweeping in the direction γ_g.
$K \leftarrow \{s_{in}, g_{in}, c_{sg}, \forall x_i \in C \mid x_i \in A_s \cap A_g\}$

5: Obtain $\cap_D$ as the final area bound by the set of 'contributing points' K:
- find the centroid c of all points in K, and sort K in terms of increasing angle from c acting as a 'polar' origin
- for each pair of consecutive points, estimate the area of the triangle formed by these two points and c
- **return** the sum of the resulting triangle areas and **exit**.

Table 4.2: Algorithm for the calculation of an exact directional t-norm, in the specific context of 2D isotropic pixels.

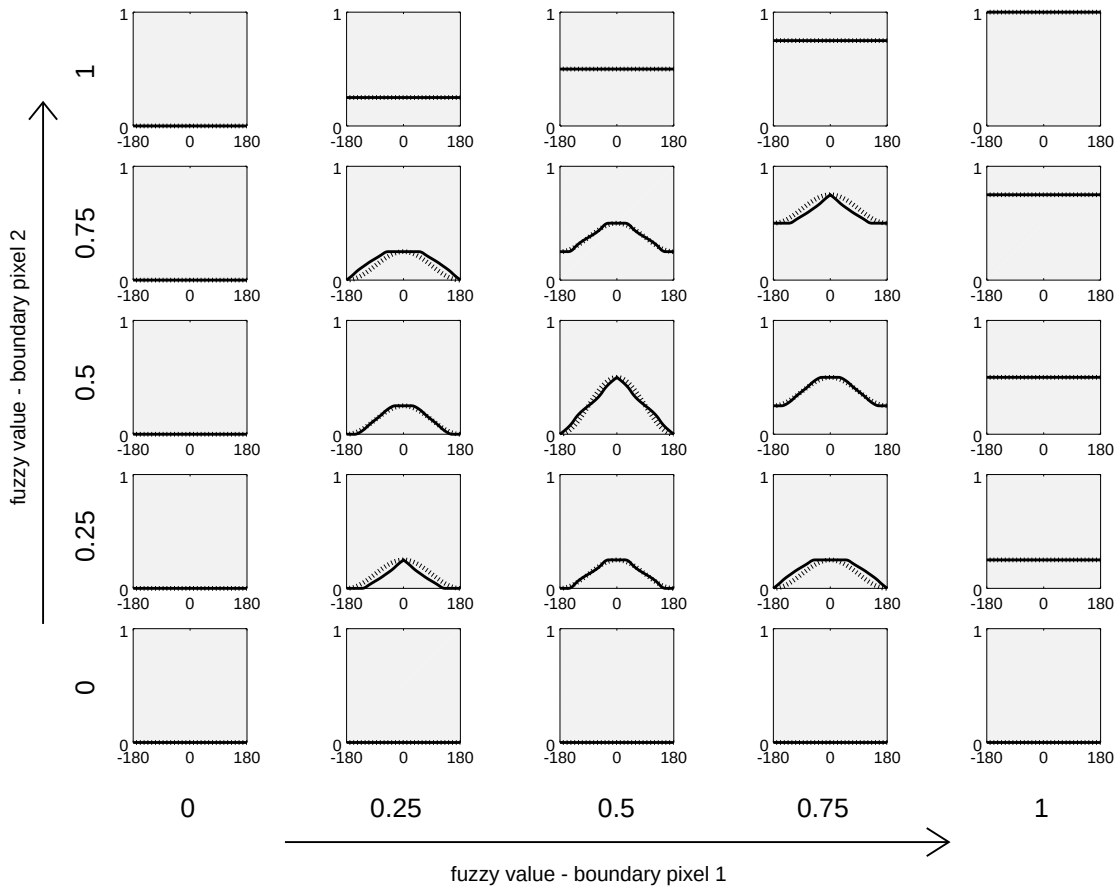


Figure 4.7: Directional t-norms: A d-norm specific to the context of 2D isotropic pixels as per section 4.6.1 (solid line), and the generalised d-norm described in section 4.6.2 (dashed line), evaluated between two boundary pixels for a range of fuzzy values and relative angle differences. Each subplot demonstrates the intersection between the two pixels at a fixed fuzzy value for both pixels; subplots from left to right correspond to increasing fixed fuzzy value for pixel 1, and from bottom to top correspond to increasing fixed fuzzy value for pixel 2. Within each subplot, the x-axis corresponds to the angle difference between the two pixels, and the y-axis to the fuzzy intersection value output given by the d-norm.

its property of outputting a suitably non-increasing value between the theoretical upper and lower bounds for the fuzzy intersection value, as imposed by the Gödel and Łukasiewicz t-norms; therefore, any function that adheres to this general scheme, irrespective of the exact pixel context, should be able to provide most of the benefits afforded by a context-specific d-norm, while having significantly less computational overhead, i.e. being of comparable computational complexity to that of standard t-norm implementations.

We provide here one such suitable implementation of a generalised d-norm to

demonstrate the concept.

Let:

- Γ_A and Γ_B be two homogeneous fuzzy pixels with intrinsic orientation, such as is the case for boundary pixels
- $\mathcal{G}$ be the Gödel t-norm of Γ_A and Γ_B, i.e. $\mathcal{G} = \Gamma_A \cap_G \Gamma_B$
- $\mathcal{L}$ be the Łukasiewicz t-norm of Γ_A and Γ_B, i.e. $\mathcal{L} = \Gamma_A \cap_L \Gamma_B$

We define a generalised d-norm $\cap_D$ to be the following sinusoidal function:

$$\Gamma_A \cap_D \Gamma_B = \Big(\frac{1 + cos\ \theta}{2}\Big)\mathcal{G} + \Big(\frac{1 - cos\ \theta}{2}\Big)\mathcal{L} \tag{4.6}$$

where θ signifies the discrepancy angle between Γ_A and Γ_B orientations. Our particular choice of a sinusoidal implementation here aims at providing a good fit to the '2d square pixel' context-specific version specified above (see fig 4.7), while keeping the analytical expression simple and thus also being straightforwardly generalisable to pixels of higher dimensions. However, we reiterate that this is simply one of many valid formulations, chosen as proof of concept; in theory, any valid formulation, i.e. any monotonically non-increasing function for increasing θ, even if this does not provide as good a fit to the context-specific case as our proposed generalised d-norm, will still provide a more accurate result than the Łukasiewicz ($\cap_L$) or Gödel ($\cap_G$) t-norms by themselves (which are already considered state of the art in a validation context).

4.7 Evaluating the reliability of fuzzy validation operators

Armed with the directional t-norms and t-conorms, we can now create a new class of fuzzy (overlap-based) validation operators, which are more appropriate for the validation of segmentation masks exhibiting Partial Volume Effect (PVE), and manifesting as fuzzy mask pixels at the object boundary. We remind ourselves that most overlap-based validation operators (such as the Tanimoto and Dice coefficients),

can be expressed as compound set operators obtained from a combination of 'base' set operations (like intersection and union), applied to segmentation and gold-standard sets (see table 4.1, p. 111). We also remind ourselves, e.g. from comparing eqs. 2.2 (p. 30) and 4.3 (p. 122) that generalising such an operator to fuzzy sets can be as simple as substituting the binary 'base' operators for fuzzy ones. Therefore we can create equivalent fuzzy validation operators (e.g. fuzzy variants of the Tanimoto coefficient) which are appropriate for PVE pixels, by generalising an existing binary validation operator using fuzzy norms — such as the directional norms — instead.

However, while it is reasonable to expect that, given their theoretical formulation, directional t-norms should lead to more reliable validation of fuzzy pixels at the boundary compared to the conventional (i.e. threshold-based) approach, in practice such a claim needs to be formally evaluated. Unfortunately, this is not as simple as assessing a validation set (i.e. a segmentation mask against a corresponding gold-standard mask) for each fuzzy validation operator, and then comparing their respective scores, since a higher score on validation does not necessarily imply a better validation *operator* per se — indeed, a validation operator could yield unnaturally high scores to the extent that, and as a result of it being *un*reliable (e.g. see footnote p. 114).

Therefore, if we are to assess the performance of the d-norm based Tanimoto operator, against the conventional and state of the art approaches, we require a suitable "external" validation score that can be trusted to be accurate, and representative of the "true" Tanimoto score for that particular validation set, so as to act as a credible *reference standard*. This may sound like a 'chicken-and-egg problem' at first; however, if such a suitable reference standard *can* be obtained, we can then easily devise an assessment scheme to meaningfully compare the performance of the validation operators more generally as algorithms *themselves*, rather than simply with respect to the specific mask inputs they act on. We describe below one way of how such a credible reference standard could be obtained; we will be using this technique in chapter 5 to compare the performance of the d-norm based

Tanimoro operators to that of the conventional and state of the art approaches, as evaluated on a synthetic and a clinical dataset.

4.7.1 Assessing fuzzy validation operator performance using a latent set

We discussed in section 4.4.2 how each fuzzy pixel in a fuzzy mask, essentially acts as some sort of summary statistic of the *latent truth* over that pixel, (i.e. the underlying, unknown, true tissue distribution within the pixel), and how such a latent truth can be modelled using a *superresolution* approach, by further subdividing each pixel into a finite (or infinite, in the theoretical limit) grid of constituent subpixels, giving rise to a finite (/ infinite) number of compatible binary masks of a higher (/ infinite) resolution compared to the initial fuzzy mask, such that subsequent 'fuzzification' of any such resultant binary mask would yield back the original fuzzy mask.

From a clinical perspective, only *one* of those binary masks is the *actual* latent truth alluded to by the experimentally obtained fuzzy mask — or, to be more exact, there is only one such mask that is the best representation of the latent truth at that particular higher resolution. While this principle applies straightforwardly to the obtained fuzzy gold standard mask, it could conceptually also be said to apply to the fuzzy segmentation candidate mask, such that this mask would have a corresponding latent segmentation candidate mask, since the underlying objective of the segmentation is to extract, characterise, and represent the *actual* object of interest (at least with respect to the latent mask's resolution limits), rather than a 'fuzzy object at low resolution' in absolute terms. Therefore we can talk about a latent validation set, consisting of this latent segmentation and a latent gold standard. If one is to assess the accuracy and precision of a fuzzy validation operator with respect to its latent validation set, then for both the segmentation and gold-standard masks, one needs to know exactly which binary mask among all the compatible masks generated at superresolution, is in fact the *one* mask that corresponds more faithfully to the latent one. One approach would be to select the best possible candidate for the latent truth from the population (or a

representative sample) of all possible, compatible binary masks, via some sort of optimisation procedure (e.g. using the heterogeneity and neighbour-compatibility scores we described earlier in section 4.5), but in practice there's no guarantee that this would be the most faithful representation of the actual latent truth.

However, the superresolution principle provides us with another approach to assess a fuzzy validation operator's performance; rather than *start* from a fuzzy set (i.e. a fuzzy segmentation mask and a fuzzy gold standard mask) and attempt to *model* the underlying latent truth for *both* masks as two binary masks of higher resolution, we can reverse this logic, and start from known *binary* masks at high resolution, from which we produce corresponding *fuzzy* masks.

More specifically we start with the following *latent set*:

- a 'high' resolution binary mask representing a very *precise* (but potentially inaccurate) *segmentation candidate*, which we denote as S_L ('L' for 'latent', and
- a 'high' resolution binary mask representing a very precise *gold standard*, acting as the best (i.e. most accurate) available representation of the ground truth for the object in question, which we denote as G_L

From this latent set, we can produce a *fuzzy set*, i.e. a 'low' resolution fuzzy mask S corresponding to the same segmentation as S_L (and similarly G to G_L), where each fuzzy pixel in S represents a distinct subset of binary pixels from S_L, obtained, for example by simple block averaging. E.g. if S_L is a 100×100 binary mask, its corresponding fuzzy mask downsampled by a factor of four would be a 25×25 *fuzzy* mask, where each fuzzy pixel has value equal to the average value of the 4×4 pixel block it represents in S_L.

We can now assess the performance of a validation operator on the fuzzy set, by comparing it against the corresponding validation on the latent set, (or *latent validation* for short). We remind ourselves that whereas in fuzzy sets, the output of

a validation operation is dependent on the particular implementation used for the intersection and union subcomponents (i.e. the choice of t-norm and t-conorm used), for binary sets there is no such distinction, as there is only one possible intersection and union outcome possible between binary masks irrespective of implementation (see section 2.4.1 eq. 2.16, p. 39), and therefore only one possible validation output.

Therefore, we can now talk about the *accuracy* of a fuzzy validation operator with respect to its corresponding *latent validation*, meaning we have a way of comparing fuzzy validation operators for performance. Furthermore, if we assess each operator's performance for a *range* of validation sets, we can also obtain a measure of *precision* for each operator.

4.8 Conclusion

We have discussed how the presence of a 'Partial Volume Effect', that is, the scenario where a pixel may contain a mixture of tissue classes, where the exact underlying distribution of the various tissues inside the pixel is unknown, may be represented using 'soft' segmentation masks, where the fuzzy pixel values act as a summary statistic of the underlying true tissue distribution inside each pixel.

We then showed that given such 'soft' masks, a fuzzy mask corresponding to the intersection (i.e. overlap) between a segmentation and a gold standard, cannot simply be obtained directly from the fuzzy values involved, but is a function of the respective presumed underlying tissue distributions in each mask, with clearly defined minimum (i.e. pessimal) overlap, average (i.e. expected) overlap, and maximum (i.e. optimal) overlap values, and we provided mathematical proofs that these correspond to the application of the Łukasiewicz, Product, and Gödel t-norms respectively.

We then discussed how PVE pixels in such soft segmentations and gold standard masks, are best represented as 'boundary' pixels, defined as homogeneous mask-pixels exhibiting an intrinsic orientation as well as a fuzzy value. We have shown

that the intersection (i.e. overlap) between two such pixels is a function of their respective orientations as well as their fuzzy values, where the lowest possible value is given by the Łukasiewicz t-norm, the highest possible value by the Gödel t-norm, and more generally, the intersection of two such pixels is a function of their difference in orientation angle, such that, at 0° difference, the intersection takes the value of the highest bound, and as this angle difference increases, the intersection takes monotonically non-increasing values, until it reaches the lowest bound at 180° angle difference.

Based on the above, we proposed the concept of a directional t-norm (or d-norm), taking boundary pixel orientation into account, and by extension, a fuzzy generalisation of the Tanimoto coefficient using directional t-norms and co-norms as the underlying fuzzy intersection and union operations. We would expect such a validation operator to be more reliable in terms of validation precision and accuracy compared to formulations which do not take this information into account, and particularly with respect to the conventional approach of 'thresholding', which is further shown to violate the theoretical pessimal and optimal theoretical bounds described above.

We reiterate that these are novel concepts, and that [22] *represents the first publication of such an idea.*

Finally, based on the principle that tissue distribution inside a pixel can be modelled as successive levels of superresolution, we detailed an approach for evaluating fuzzy validation operators acting on 'soft' sets, by providing a simple way to generate such sets together with a known, higher-resolution 'latent' validation set that can be used as the ground truth for operator comparison.

In the next chapter, we will be using this *latent validation* scheme, to evaluate the accuracy and precision of directional-norm fuzzy validation on both synthetic and real clinical sets, against the conventional and state of the art approaches; we will

also be using this to explore how the choice of validation operator may affect the comparison of performance between two or more algorithms.

Summary

- Appropriate validation is crucial in ensuring the quality and reliability of segmentation algorithms, both in terms of performance evaluation, as well as in terms of reliable training.
- Validation in medical images presents many specific challenges, some practical, and some technical in nature.
- Most validation approaches for classical deterministic segmentation — particularly those relating to overlap — make use of classical set operations.
- The conventional approach to the validation of 'soft' segmentations in the literature is to convert them to 'classical' sets, typically by thresholding; despite the odd concern expressed in the literature, adequate investigation into the reliability or suitability of this approach has been lacking.
- 'Soft' segmentation masks may be used to model tissue distribution in Partial Volume pixels; in this context, the degree of overlap between two fuzzy pixels is a function of their respective underlying tissue distributions.
- We have provided proofs for the fact that there is an upper and lower bound that this overlap can take, and that these correspond mathematically to the Gödel and Łukasiewicz t-norms respectively.
- We demonstrated that thresholding-based intersection violates these absolute bounds, rendering thresholding-based validation operators unreliable.
- Pixels at the object boundary can be modelled as homogeneous fuzzy pixels with an intrinsic orientation.
- The degree of overlap between two boundary pixels is a function of their intrinsic orientations as well as fuzzy values. This naturally leads to the concept of *directional t-norms and t-conorms*, as fuzzy generalisations of the intersection and union operators for 'boundary pixels', defined as monotonic functions of the orientation difference between two pixels, constrained within the absolute bounds described above.
- We differentiate between **context-specific** and **generalised** d-norms, depending on whether specific pixel shape is taken into account or not.
- Tissue distribution within a 'soft' pixel can be modelled as successive levels of superresolution; this process can be used in reverse to generate fuzzy validation sets from a known 'latent' set, such that the reliability of fuzzy validation operators can be assessed against a latent truth.

"A sail with no direction knows no favourable wind"

— Lucius Annaeus Seneca (Roman stoic philosopher, c. 4 BC – 65 AD);
Letter LXXI: On the supreme good

*"In theory, there is no difference between theory and practice.
In practice, there is."*

— Yogi Berra

5

Directional t-norms for fuzzy validation

In this chapter, we apply the theoretical concepts introduced in the previous chapter, and demonstrate the application of a d-norm based validation operator on a synthetic, and on a real clinical dataset[1].

Furthermore, we demonstrate conclusively the unreliability of the conventional, thresholding-based operator, and show the dangers this entails when comparing the performance of individual algorithms.

The chapter ends with a plea to the medical imaging community to stop using this as a de facto standard for validation, and highlights the importance of rigorous research and evaluation of validation algorithms to the same degree and high standards as segmentation algorithms themselves.

Contents

[1] As published in: Tasos Papastylianou, Erica Dall' Armellina, and Vicente Grau. "Orientation-Sensitive Overlap Measures for the Validation of Medical Image Segmentations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 361–369

In the previous chapter, we motivated the case for directional norms (or *d-norms* for short), and their use in the context of fuzzy-generalised validation operators that are more appropriate for the case of soft masks exhibiting PVE in the form of 'boundary pixels'.

In this chapter, we demonstrate the application of such d-norm based validation operators on a synthetic and on a real clinical dataset, and evaluate their accuracy and precision using the latent validation scheme described in section 4.7.1, compared to the conventional (i.e. thresholding-based) and state of the art approaches.

For the purposes of demonstration in this thesis we focus explicitly on the Tanimoto coefficient (T_c), as this is a straightforward operator to use, though we reiterate the point made in section 2.4.1 that the concepts we discuss here apply equally to the Dice coefficient, or any other set-based operator (and that these all tend to be largely interrelated anyway).

5.1 Comparison to state of the art on a synthetic set

In this section, we evaluate the directional-norm operators, the traditional thresholding-based approach described in section 4.2.2, and the state of the art operators described in section 4.2.3, using a slightly modified version of the synthetic 'petal' set by Crum *et al.* [118] (see eq. 4.1, p. 120).

5.1.1 Methods

The original petal experiment in Crum *et al.* involves comparing an analytical shape to a quantised version of that shape. In our experiments, instead of calculating petal overlap analytically at each rotation step like in Crum *et al.*, the latent truth is obtained by simply replicating the petal set as a 'high resolution' binary image (100×100), like the one shown in fig. 5.1. More than just easing calculations, this

slight modification confers a few more advantages in our particular context, while still comparable and consistent with the spirit of the Crum *et al.* experiment:

- it allows for more consistent and comparable calculation of gradients between the two sets with respect to the particular 'blocks' involved, and also allows for the effect that the 'level' of subpixel resolution (i.e. in terms of number of 'grid subdivisions') has in terms of operator accuracy to be quantified.
- it is more consistent with clinical practice, since 'soft' gold-standards are usually obtained as manual contours drawn at a particular subpixel resolution, rather than as a truly continuous contour
- it enables better consistency with, and better comparison to the clinical set later on, which is only available as a high-resolution mask rather than as a continuous or analytical contour.

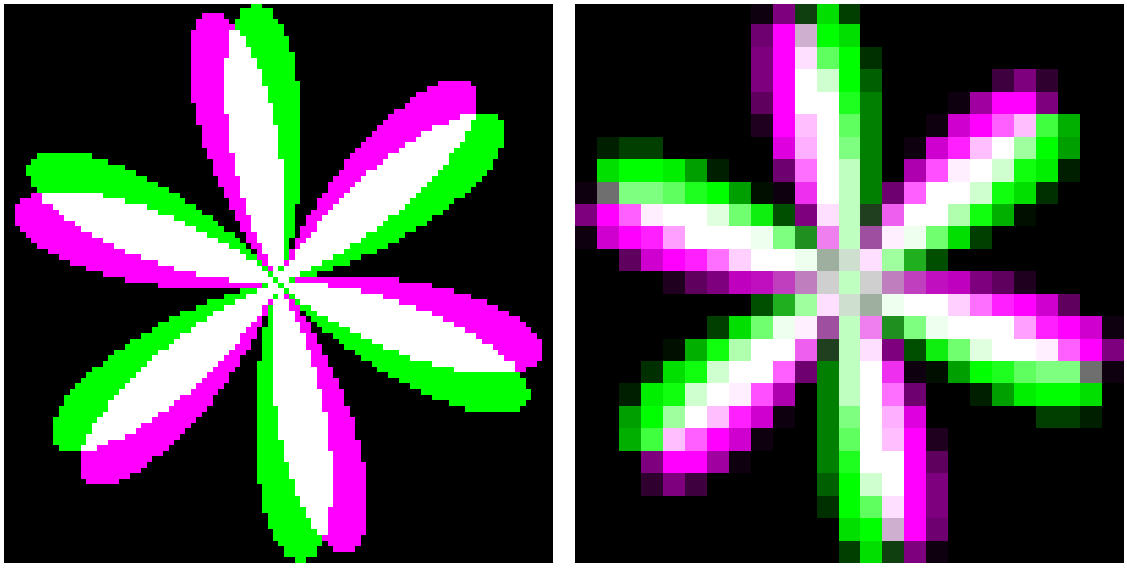


Figure 5.1: Segmentation vs gold-standard mask fusion images from the petal set. Left: Latent truth L as a 100×100 high-resolution representation of the petal structure. Right: Fuzzy data, obtained by downsampling the latent truth down to a lower-resolution (25×25). The segmentation mask is shown as violet, and the gold-standard mask as green; colour-fusion (corresponding to areas with overlap) produces white colour.

'Latent validation' data is obtained by rotating one copy of the high-resolution petal image (acting as the segmentation binary mask S_L), onto another, stationary high-resolution petal image acting as the gold-standard binary mask G_L. For each

S_L and G_L set, we calculate the standard Tanimoto coefficient, to evaluate the quality of their overlap, (i.e. the accuracy of the segmentation candidate with respect to the gold standard) at various angles of rotation. At each rotation angle, corresponding fuzzy segmentation and gold-standard masks are also produced from their corresponding latent truth masks, using simple 4×4 block-averaging, resulting in a fuzzy mask with 25×25 resolution (i.e., each 4×4 block in the high-resolution masks becomes a single pixel in the fuzzy, low-resolution masks). For each angle, we obtain a fuzzy validation output for each of the following methods: traditional (i.e. binary validation post-thresholding at the 0.5 threshold), Yi [18], Chang [115], Crum [118] (i.e. the Gödel norms), Product norms, Łukasiewicz norms, and finally the directional norms — both the generalised and the context-specific one (as implemented in appendix A); a discussion on the choice of gradient responses used in the calculation of orientation angles is deferred to section 5.1.2 (p. 156).

5.1.2 Results

We compare the validation outputs obtained from the various fuzzy operators to the validation output of the latent truth; fig 5.2 shows both the absolute outputs, and the difference between each method and the latent validation output at each rotation angle.

Figures 5.3 and 5.4, similarly show the distribution of these differences over all angles, demonstrating the accuracy and precision of each operator more clearly; table 5.1 expresses the same data as fig. 5.4 in tabular form.

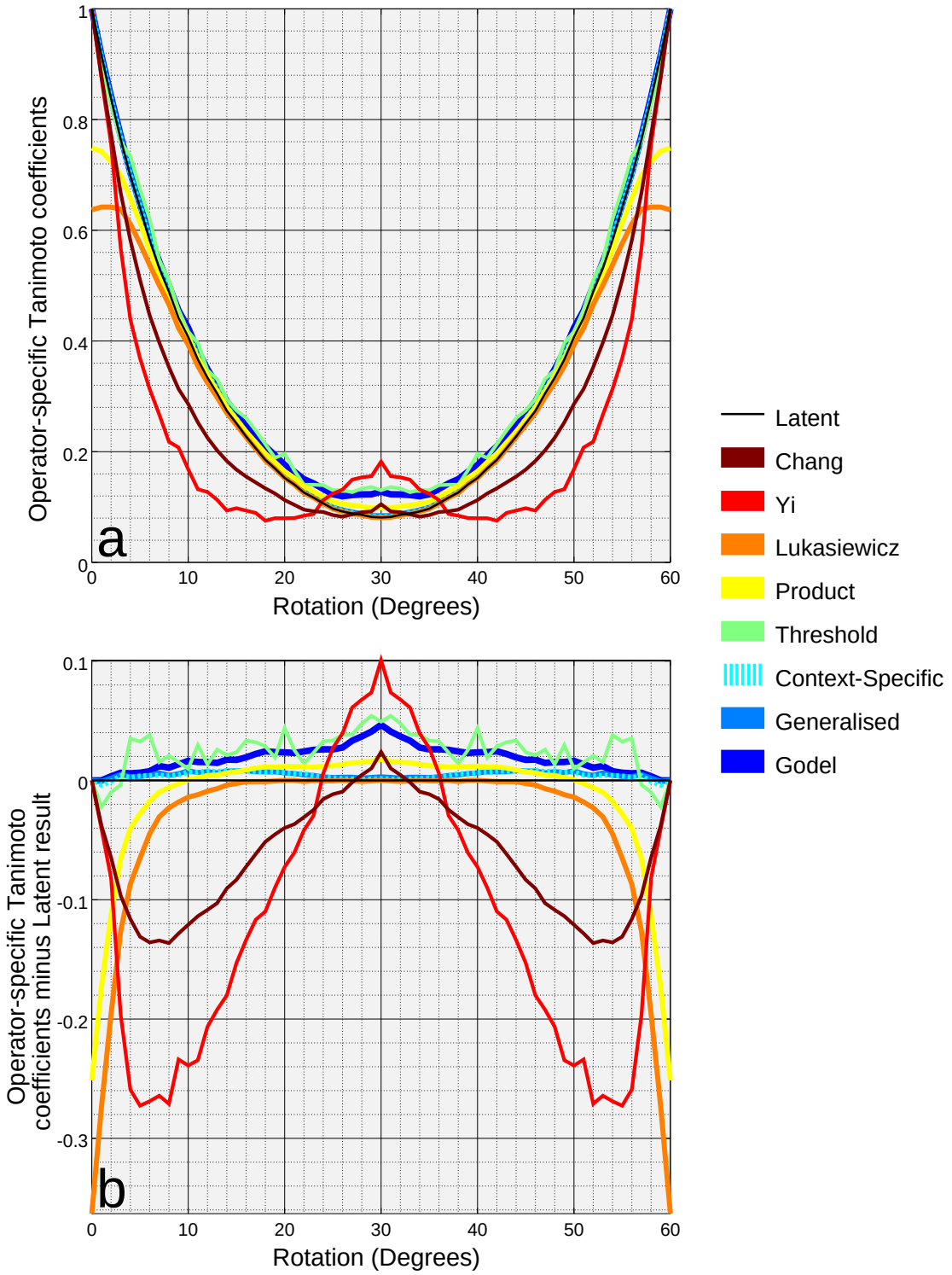


Figure 5.2: Fuzzy-operator dependent Tanimoto coefficients compared to latent validation truth: **a.** Actual Tanimoto coefficient values for all operators. Only rotations from 0° to 60° are considered due to the petal's symmetry. **b.** Outputs expressed in terms of their difference from the latent validation (i.e. fuzzy operator minus latent validation). The *context-specific* and *generalised* d-norms produce similar results, (effectively overlapping at this scale), and are the outputs with the least variability, and closest to the latent validation truth (i.e. the closest to the 'zero-difference' line in the bottom plot).

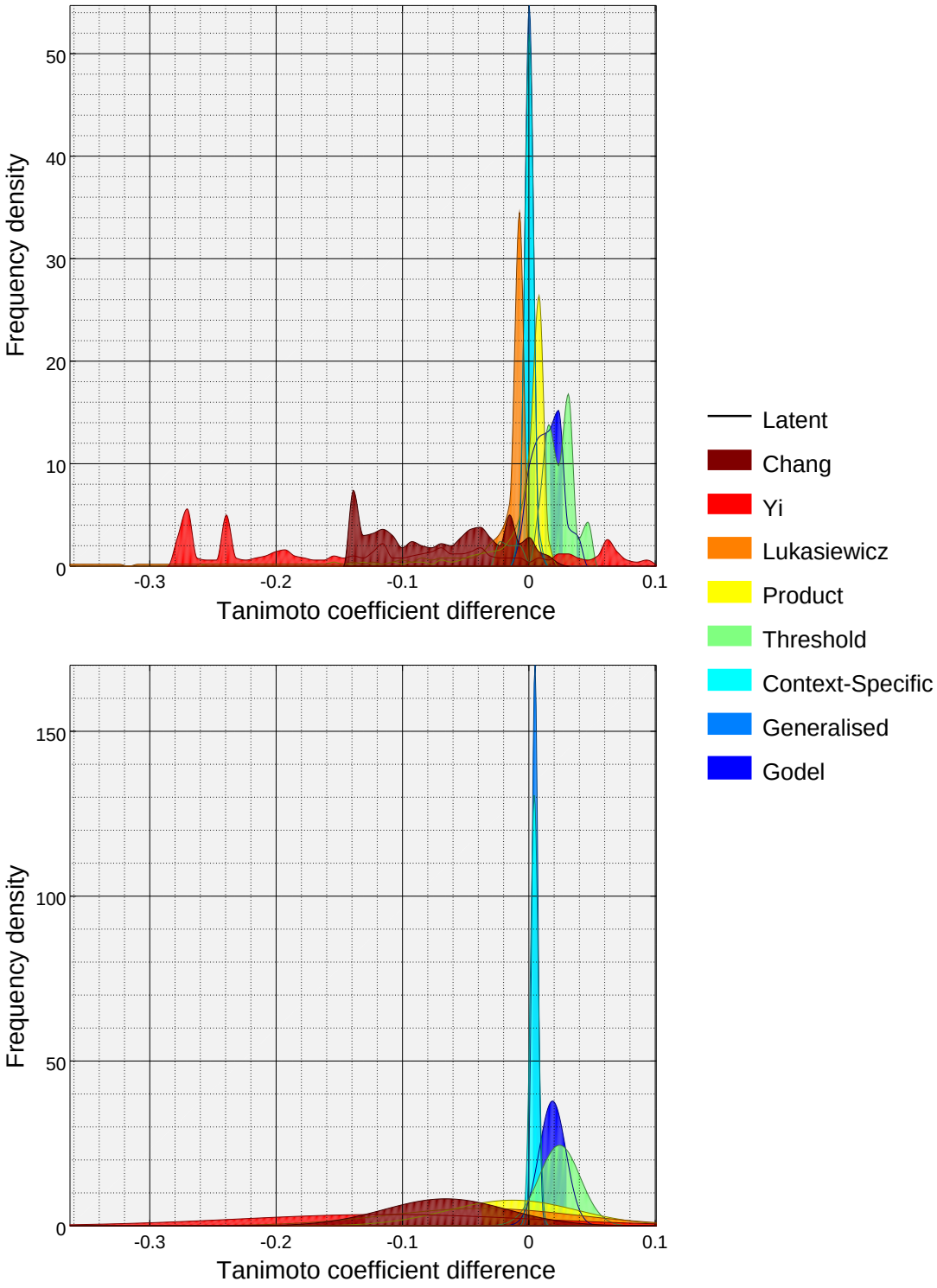


Figure 5.3: Distributions of Tanimoto coefficient differences with respect to latent truth, for all fuzzy operators. **a.** Actual distribution of differences. **b.** Differences represented as Gaussian curves with mean and standard deviation derived from sample mean and standard deviations from the actual distributions above, making the expected value (accuracy) and spread (precision) between the different operators easier to compare visually.

The *context-specific* and *generalised* d-norms produce similar results, (effectively overlapping at this scale), demonstrating the least variability (as evident by their tall, sharp peaks) and highest accuracy (as evident from having the mean closest to zero).

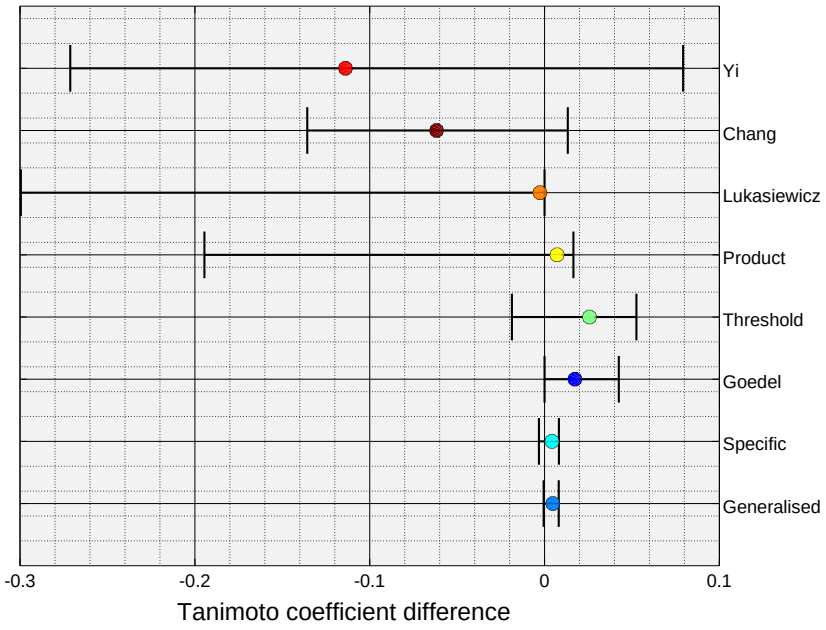


Figure 5.4: Boxplot comparison of fuzzy operators for the synthetic dataset. Each boxplot shows the median in a coloured marker, and 95% confidence intervals.

	Petal set					
	Median	95% confidence range (interval)				
Yi	-0.114	0.351	(	-0.271	–	0.079)
Chang	-0.062	0.149	(	-0.136	–	0.013)
Lukasiewicz	-0.003	0.300	(	-0.300	–	0.000)
Product	0.007	0.211	(	-0.195	–	0.017)
Threshold	0.026	0.071	(	-0.019	–	0.053)
Goedel	0.017	0.042	(	0.000	–	0.042)
Specific	0.004	0.011	(	-0.003	–	0.008)
Generalised	0.005	0.009	(	-0.001	–	0.008)

Table 5.1: Summary of differences from the latent truth, evaluated for each fuzzy validation operator on the synthetic set — this is the same information presented in fig. 5.4 but in tabulated form.

Unprocessed vs ideal gradient for fuzzy pixel orientation

We mentioned in section 4.5.1 that a boundary pixel can be modeled as a homogeneous, linearly-separated pixel of a particular orientation, where the orientation corresponds to the ideal negative local gradient for the object at that pixel. In practice, this gradient needs to be calculated from the fuzzy mask, and there are several ways to calculate such a gradient. Therefore, the effect that the choice of gradient algorithm has on the accuracy and precision of the d-norm based validation operator needs to be quantified.

Methods: We compared two gradient calculation methods. The first is the standard Sobel gradient operator, applied directly on the fuzzy mask (implemented via the matlab / octave `imgradientxy` command. The Sobel operator is considered one of the most basic gradient responses that can be obtained from a mask; it is usually very susceptible to noise, which is why pre- and post-processing steps are usually required to improve the outcome (e.g. a typical pre-processing step involves blurring the mask to reduce the contribution of 'noisy' pixels).

Since we choose to focus simply on demonstrating *whether* a more appropriate gradient response makes the d-norm based validation operators more accurate and precise, and if so to what extent, we compare the output of the Sobel operator with no further pre- or post-processing, against an "*ideal*" gradient response; similar to how we obtain the latent set, rather than try to *model* the ideal gradient response using sophisticated methods from the literature, we obtain the *actual* latent gradient response directly from the latent masks, and 'fuzzify' these using *block-averaging* in order to obtain a gradient vector for each pixel in the fuzzy mask; the resulting mask was also convolved with a 'neighbourhood averaging' kernel[2], to ensure 'smoother' gradient transitions, offsetting potential 'step' changes introduced during quantization.

[2]specifically, the kernel used in our implementation was $\begin{pmatrix} 1/16 & 1/16 & 1/16 \\ 1/16 & 1/2 & 1/16 \\ 1/16 & 1/16 & 1/16 \end{pmatrix}$.

Results: Table 5.2 summarizes the performance of the two d-norm based validation operators, as obtained for an unprocessed and an 'ideal' gradient.

Petal Set				
	Unprocessed Gradient		Ideal Gradient	
	Median	95% confidence range (interval)	Median	95% confidence range (interval)
Specific	0.004	0.011 (-0.003 – 0.008)	0.002	0.008 (-0.003 – 0.005)
Generalised	0.005	0.009 (-0.001 – 0.008)	0.002	0.006 (-0.001 – 0.005)

Table 5.2: Effect of gradient response on d-norm based operator precision and accuracy. The gradients are used to estimate the boundary pixel orientations. A basic, unprocessed gradient is compared to an ideal gradient, obtained directly from the latent set

Effect of effective fuzzy mask resolution on validation

It is reasonable to expect that the effective resolution of a fuzzy mask, i.e. the degree of quantization inherent to the fuzzy mask with respect to its latent truth, is an important factor affecting the accuracy and precision of validation on such a fuzzy set. For instance, a fuzzy mask which is of only slightly lower resolution than its generative 'latent' mask, could be reasonably expected to give a validation outcome which is much closer to the latent truth, than a fuzzy mask of much lower resolution, regardless of the validation operator used. What might not be straightforward to predict, however, is the extent to which individual validation operators are more or less susceptible to this effect.

Methods: To test this, we repeated the experiments for multiple levels of fuzzification, namely 2×2, 4×4 (default), and 8×8 block averaging.

Results: Table 5.3 shows the effect of 'high' (50×50), 'default' (25×25), and 'low' (13×13) fuzzy mask resolutions on the output of the various validation operators, with respect to an original latent resolution of (100×100).

	'Low' resolution median (95% range)	'Medium' resolution median (95% range)	'High' resolution median (95% range)
Yi	0.085 (0.318)	-0.114 (0.351)	-0.084 (0.158)
Chang	-0.002 (0.125)	-0.062 (0.149)	-0.088 (0.130)
Lukasiewicz	-0.023 (0.455)	-0.003 (0.300)	0.000 (0.117)
Product	0.016 (0.389)	0.007 (0.211)	0.002 (0.068)
Threshold	0.028 (0.098)	0.026 (0.071)	0.029 (0.041)
Gödel	0.063 (0.091)	0.017 (0.042)	0.005 (0.008)
Specific	0.032 (0.037)	0.004 (0.011)	0.001 (0.005)
Generalised	0.031 (0.034)	0.005 (0.009)	0.000 (0.005)

Table 5.3: Effect of effective fuzzy mask resolution on validation accuracy. Fuzzy sets of 'low' (13×13), 'medium' (25×25), and 'high' (50×50) resolution are validated using the above fuzzy validation operators. Outputs correspond to validation outputs minus the validation output of the latent truth set (100×100).

5.1.3 Discussion and analysis.

The results obtained from this synthetic experiment demonstrate that both the context-specific and generalised d-norm validation operators are robust and reliable, significantly outperforming both conventional and other state of the art methods in terms of validation precision and accuracy. Furthermore they support the claim that the conventional approach to validation by thresholding is unreliable, in that it is both inaccurate and imprecise with respect to the validation reference standard. In this section, we discuss some of these findings in more detail.

Performance of the directional t-norms

Both the generalised and context-specific d-norms seem to perform equally well. Furthermore, as predicted from the theoretical analysis in this chapter, both d-norm types are much more accurate and precise than all the other methods investigated.

Effect of gradient estimation method in d-norm accuracy / precision.

Arguably, one of the simplest gradient calculation methods that can be applied to represent the orientation at each pixel, is the direct application of the Sobel operator on the fuzzy mask values, with no further pre- or post-processing; such an approach does not necessarily give the most useful edge response since it is

susceptible to noisy pixels, and does not attempt to model continuity with respect to the transition in orientation angles from one boundary pixel to the next along the object's contour in any meaningful manner.

There are many gradient estimation approaches we could take instead of this simple one. In general, we might consider the choice of a gradient method to be purely a question of implementation, so such a choice is left to the user, to be decided with respect to the image substrate and particular problem at hand. However, from the point of view of investigating the effect of using a 'better' gradient method in general, to see how much difference it makes, the optimal gradient response that would lead to the most accurate validation, is the gradient response that corresponds to the latent object. While this is generally not known in straightforward fuzzy validation, it is easy to construct from our latent set in the context of the experiment above, designed to evaluate the fuzzy operators against a *known* "latent" truth.

From the results in table 5.2 we can see that a better gradient does indeed make the directional operator both more accurate and precise. However, interestingly, we see that despite the simplicity of a Sobel approach, the d-norm based validation operator still leads to remarkable improvement in accuracy and precision over the other methods. We could conclude from this that, although better quality orientation masks will clearly lead to more reliable validation in the context of a d-norm, even lesser quality orientation information will lead to marked improvement, compared to not taking this information into account at all.

Performance of fuzzy validation operators based on standard t-norms

Out of all the methods compared, after the directional norms, the approach by Crum *et al.* (i.e. a Tanimoto coefficient implemented via the Gödel norms) seems to be the next most accurate / precise overall, despite its known *optimism* bias (i.e. the difference from the 'true' validation score is always positive).

The Łukasiewicz-based operator, which has the reverse bias, while very accurate at times of very *bad* overlap, seems to be extremely *in*accurate at times of decent /

good overlap, and as such is an imprecise and unreliable operator overall. In this experiment, due to the fact that bad overlaps are *over*represented (since there is only a short range of angles for which overlap is good, and the remaining angles of rotation lead to relatively bad overlap), the median value for the Łukasiewicz operator makes it seem like a rather accurate operator (more accurate than the Gödel-based operator in fact, see fig. 5.4); however, the long negative tail is a tell-tale sign of outputs that are extremely inaccurate, even if those instances were relatively few in this particular dataset, and therefore this operator is unreliable, even for this dataset where it is accurate more often than not.

As might be expected from the Product-based operator, which corresponds to an 'average-overlap' scenario, it always returns validation scores between the pessimistic, Łukasiewicz-based operator, and the optimistic, Gödel-based operator; in the context of this experiment, given the fact that the medians for both the Gödel and Łukasiewicz operators were fairly accurate, this results in a slightly more accurate result for the Product operator than either of the two; however, we note that while relatively accurate, it has a long negative tail like the Łukasiewicz result, and therefore could not be expected to be a reliable validation operator in general.

Performance of the conventional / thresholding approach

The thresholding approach is shown to be unreliable with respect to the 'true' validation score. This finding is of particular importance, given the fact that by convention, validation by thresholding seems to be the method of choice in the segmentation literature. In terms of its absolute output, it often violates the theoretical upper and lower bounds defined by the Gödel and Łukasiewicz norms respectively, particularly the upper bound (fig. 5.2). In terms of its overall distribution, it resulted in a median value that was even *higher* than the corresponding Gödel distribution, implying that it tended to produce outputs that were artificially overoptimistic (at least in this experiment), beyond what is semantically meaningful in the context of fuzzy pixels.

Furthermore, it also has wide 95% confidence intervals ($\sim 7\%$ of the Tanimoto operator 0–1 range in our sets), compared to the Gödel norm ($\sim 4\%$) and the directional norms ($\sim 1\%$), showing it is imprecise as a validation operator. This is presumably due to the high degree of 'quantization' and information lost through the act of thresholding, resulting in a validation operator with reduced 'validation resolution', so to speak.

We could also make the point that, while the Gödel-based operator could be said to suffer from a *systematic bias* (i.e. it is biased towards optimistic values by design), which, while unwanted, could nevertheless potentially be taken into consideration when interpreting the result, and otherwise results in a relatively more precise operator, the Threshold-based operator is completely unpredictable, as can be seen from the large, unpredictable zigzag deviations around the latent truth in our experiments. As a result, the use of a threshold-based validation operator could even lead to false conclusions regarding both the absolute performance of a segmentation algorithm / result, but also, more importantly, regarding the relative quality between two segmentation algorithms.

Performance of the remaining state of the art methods investigated

Neither the Yi nor the Chang operators seem particularly accurate / precise compared to the other operators. In this experiment, they both seem to be over-pessimistic at times of good overlap, and over-optimistic at times of bad overlap.

We note a few things about them that might explain this observation:

- Both algorithms are based on re-interpretations of fuzziness, which have no basis in established fuzzy set theory.
- Both the Yi and Chang algorithms also violate the theoretical lower (Łukasiewicz) and upper (Gödel) bounds discussed earlier, much like the Threshold-based operator.

The original papers never evaluated their operators in terms of validation accuracy or precision, but simply presented them as theoretical generalisations of existing operators, under the assumption that it therefore made sense that they would work better than a simple thresholding approach. However, it turns out that this is not necessarily the case.

Effect of decreasing the resolution of the generated fuzzy set relative to the latent set

As might be expected, lowering the resulting resolution of the fuzzy mask with respect to the latent truth, affects all methods negatively in terms of overall reliability (table 5.3). However, the relative strengths between the various operators seem to be largely preserved, and the directional t-norms still perform significantly better than the other approaches at any resolution level.

Interestingly, even though precision was lower as expected, the Threshold-based operator's accuracy did not deviate for lower resolutions in this set. This may simply be an artefact of this set, or it might indicate that the accuracy bias demonstrated by the Threshold-based operator is rather systematic from a resolution point of view.

5.2 Comparison to state of the art on a retinal set

While the directional norms should in principle prove to be more accurate and precise in any clinical set in the same manner as the synthetic set, cardiac MRI segmentation (e.g. of the blood pool) may not be the best example to demonstrate their superiority, as the objects involved have a relatively low surface-to-volume ratio (in other words, a disproportionately low number of boundary pixels compared to core pixels) compared to delicate structures such as blood vessels, and therefore the special significance of fuzzy validation, although still important, will be downplayed and harder to demonstrate visually (we elaborate on this point further in the analysis below). Therefore, for the purposes of a visually clear and convincing

clinical example we demonstrate the above concepts here on a medical dataset involving thin structures instead.

5.2.1 Methods

The STARE (STructured Analysis of the REtina) Project [23] provides a clinical dataset of 20 images of human retinae, freely available online. For each image, it also provides a triplet of binary masks (700×605): two manual delineations of retinal blood vessels from two medical experts and one automated method. For the purposes of this assessment, one of the manual sets is treated as the gold-standard mask, and the other two are treated as normal segmentation masks (i.e. comparing a human rater versus a computer algorithm). Fig. 5.5 shows an example of the automated algorithm-derived segmentation mask, against the gold standard mask. Similar to the petal set, fuzzy versions were produced using various degrees of block-averaging, and validated using the same array of methods.

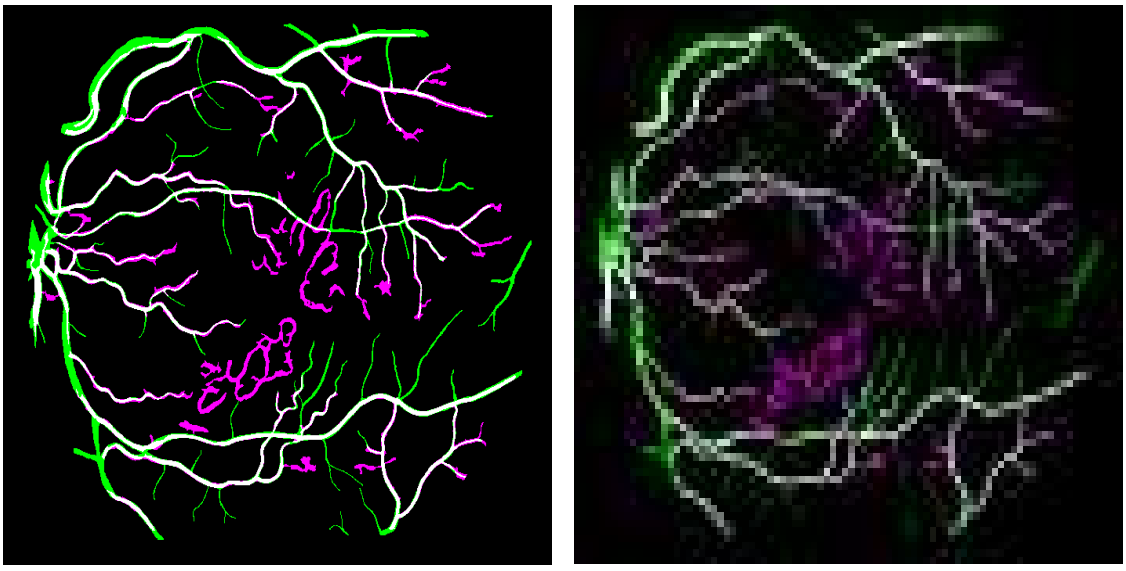


Figure 5.5: Illustrative segmentation vs gold-standard mask fusion images from the STARE clinical dataset. Left: Latent validation set as a 100×100 high-resolution representation of the retinal masks. Right: Corresponding fuzzy masks, obtained by downsampling the latent truth down to a lower-resolution (25×25). The segmentation mask is shown as violet, and the gold-standard mask as green; colour-fusion (corresponding to areas with overlap) produces white colour.

5.2.2 Results

We compare the validation outputs between the computer algorithm and human rater, for each of the 20 experiments, over all fuzzy operators. Fig. 5.6 shows both the absolute validation and the differences between the two raters (human vs algorithm). Figs 5.7–5.8, and table 5.4 show the distribution of the differences, as in the synthetic case. Fig 5.9 compares the performance difference between the human rater and the automated algorithm, as assessed by each validation operator.

	Human rater					Automated algorithm				
	Median	95% confidence range (interval)				Median	95% confidence range (interval)			
Yi	-0.081	0.171	(	-0.150	– 0.020)	-0.096	0.129	(	-0.139	– -0.010)
Chang	-0.033	0.103	(	-0.068	– 0.036)	-0.041	0.070	(	-0.074	– -0.004)
Lukasiewicz	-0.252	0.184	(	-0.340	– -0.156)	-0.166	0.111	(	-0.225	– -0.114)
Product	-0.147	0.123	(	-0.207	– -0.084)	-0.090	0.074	(	-0.132	– -0.058)
Threshold	0.024	0.088	(	-0.023	– 0.066)	0.033	0.080	(	-0.007	– 0.073)
Goedel	0.014	0.043	(	0.006	– 0.048)	0.010	0.015	(	0.005	– 0.020)
Specific	-0.004	0.019	(	-0.016	– 0.003)	-0.002	0.005	(	-0.005	– -0.000)
Generalised	-0.004	0.021	(	-0.016	– 0.005)	-0.002	0.003	(	-0.003	– 0.000)

Table 5.4: Summary of human rater / automated algorithm differences from the latent truth, evaluated for each fuzzy validation operator — this is the same information presented in fig. 5.8 but in tabulated form.

Effect of unprocessed vs ideal gradient for fuzzy pixel orientation

As noted in section 5.1.2, since the orientation of boundary pixels is estimated using the mask gradient, one would expect that the choice of gradient estimator method should matter in terms of the potential accuracy of the d-norm validation operator. In the synthetic test, the difference in accuracy and precision proved to be minimal, and still better than the other methods that we compared against.

However, the synthetic set consists of a highly symmetric object. One could have argued, therefore, that this could have explained, in part, why the choice of gradient estimator method did not have a pronounced effect; the retinal set consists of highly irregular, kinked structures. Therefore we repeat this process here, to further quantify this effect in a non-contrived object.

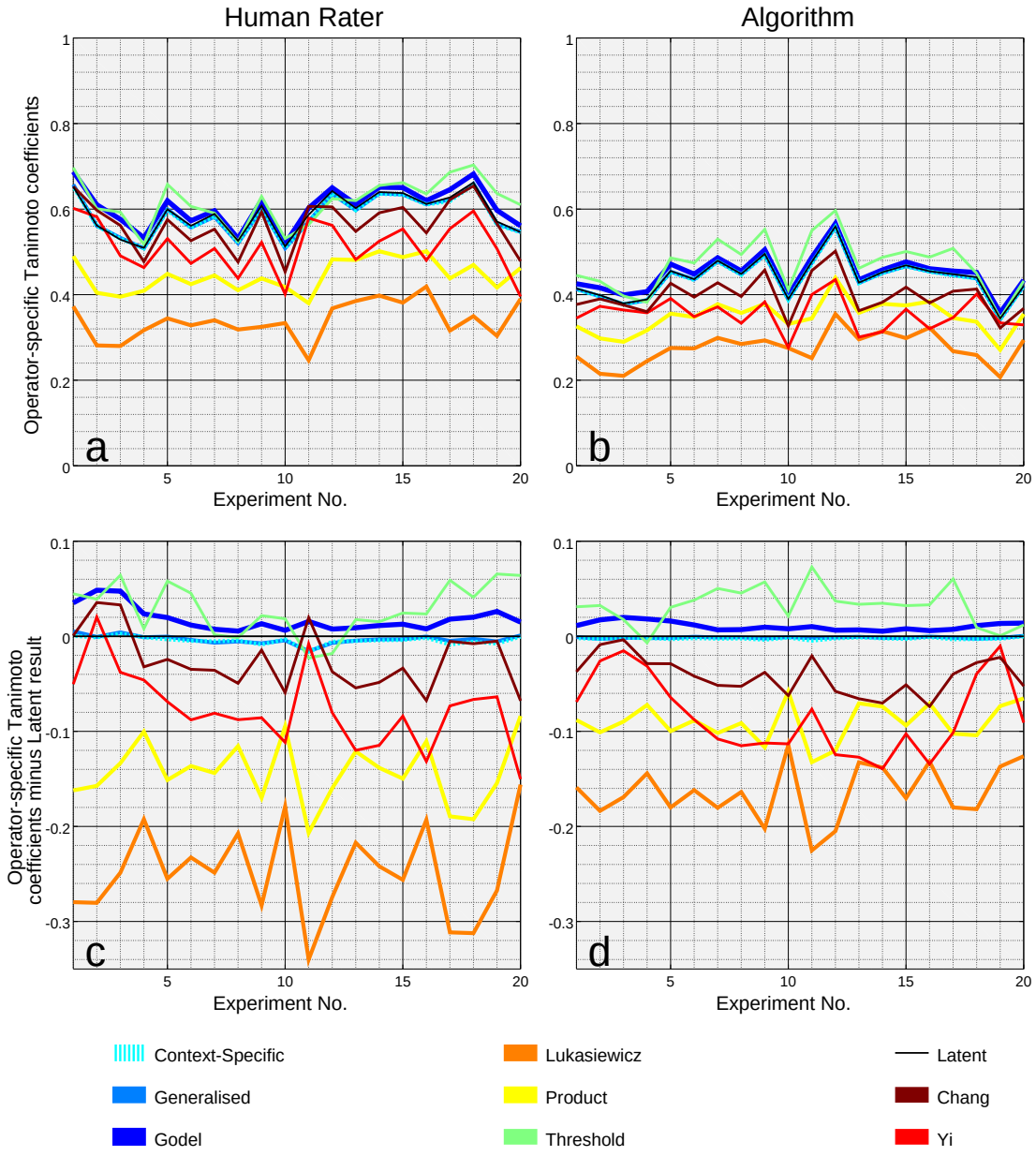


Figure 5.6: Absolute and relative outputs of fuzzy validation operators, evaluated over a clinical set of 20 experiments. **a.** Absolute validation outputs for the human rater. **b.** Absolute outputs for the automated algorithm. **c.** Validation output relative to latent truth for the human rater. **d.** Validation output relative to latent truth for the automated algorithm.

In both cases, the *context-specific* and *generalised* d-norms produce similar results, (effectively overlapping at this scale), and show the least variability, and difference from the latent validation truth (i.e. they consistently appear the closest to the 'zero-difference' line in the bottom plots).

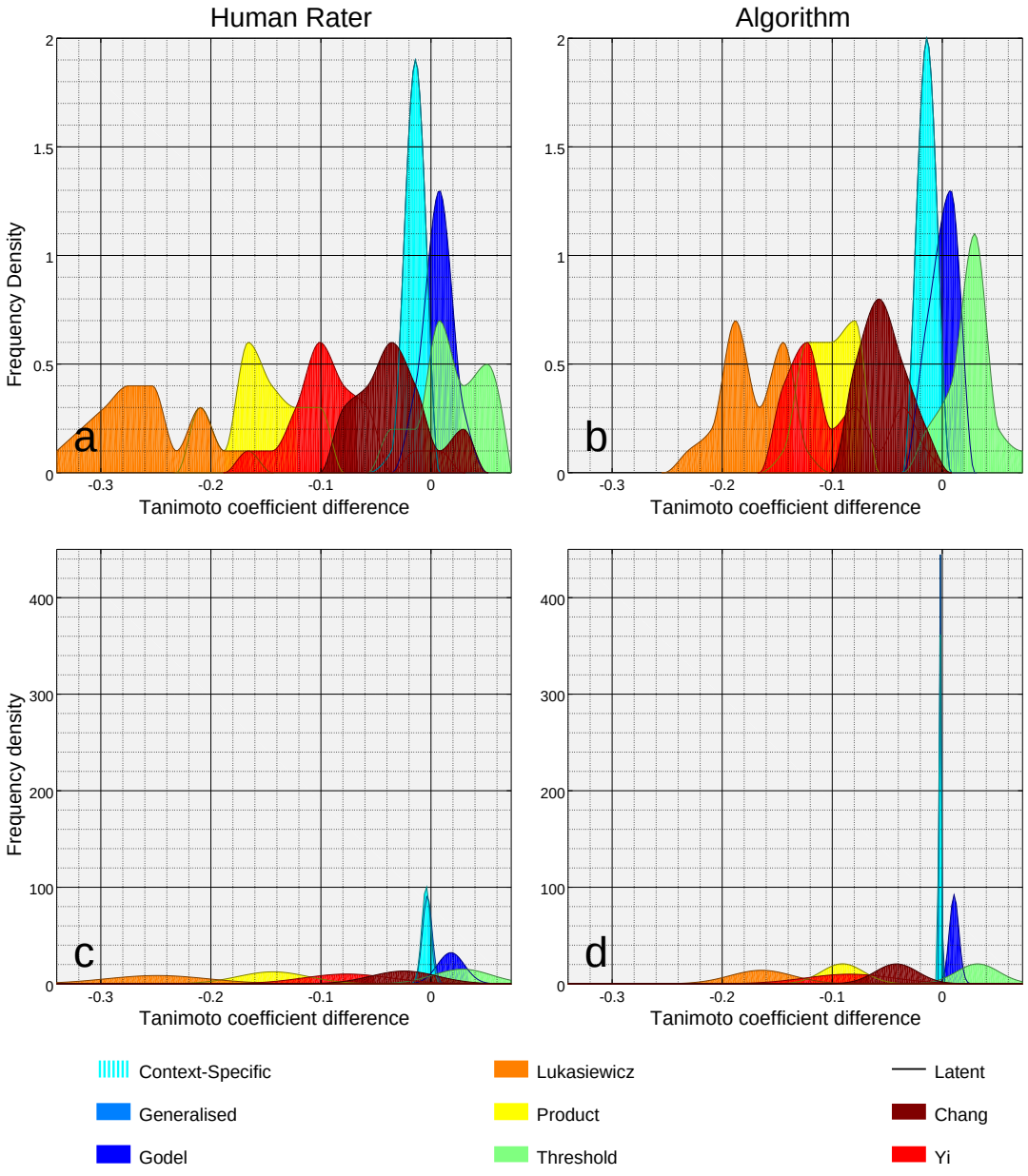


Figure 5.7: Distributions of Tanimoto coefficient differences for the clinical set with respect to latent truth, for all fuzzy operators. **a:** Actual distribution of differences for the human rater. **b:** Actual distribution of differences for the automated algorithm. **c** and **d:** Differences represented as Gaussian curves with mean and standard deviation derived from sample mean and standard deviations from the actual distributions above, making the expected value (accuracy) and spread (precision) between the different operators easier to compare visually

In both cases, the *context-specific* and *generalised* d-norms produce similar results, (effectively overlapping at this scale), demonstrating the least variability (as evident by their tall, sharp peaks compared to the other operators) and highest accuracy (as evident from having the means closest to zero). Note that, in general, the validation *operators* were more precise and accurate in the automated algorithm case compared to the human rater, even though the algorithm's validation *output* itself was generally lower than that of the human rater (see top row of fig. 5.6).

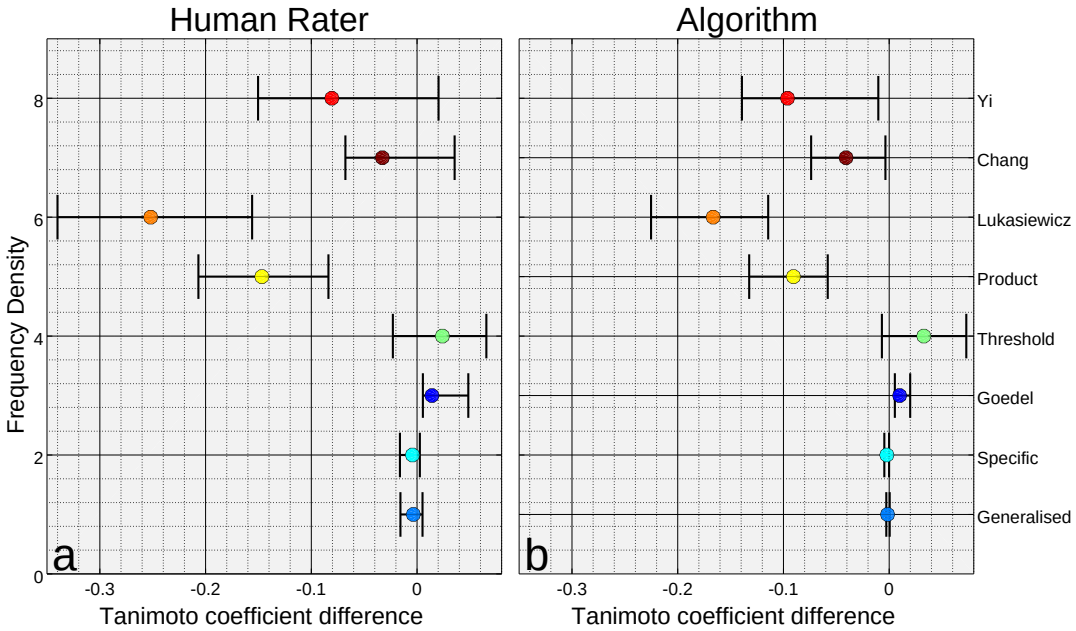


Figure 5.8: Boxplot comparison of fuzzy operators for the clinical dataset. Left: human rater. Right: automated algorithm. Each boxplot shows the median for each method in a coloured marker, and 95% confidence intervals.

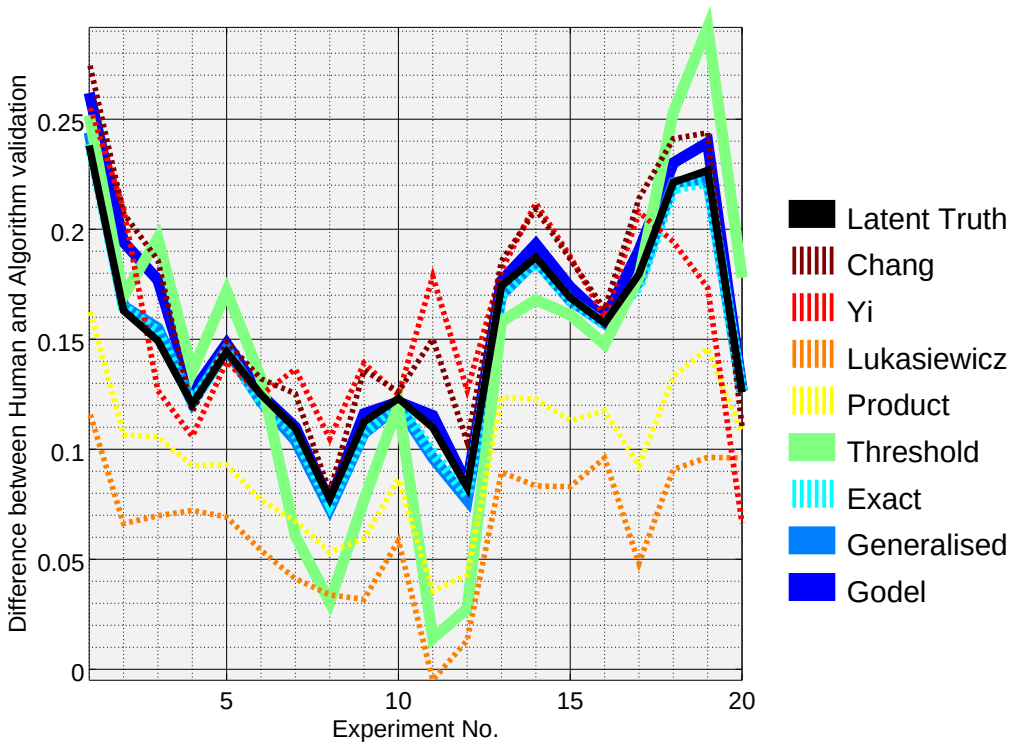


Figure 5.9: Performance difference between the human rater and the automated algorithm, as assessed by each validation operator

Methods: We compare, as before, the standard Sobel gradient operator, against an "*ideal*" mask gradient, obtained in the same manner as in section 5.1.2.

Results: Table 5.5 summarizes the performance of the two d-norm based validation operators on the STARE dataset, as obtained for an unprocessed and an 'ideal' gradient.

STARE dataset - Human Rater				
	Unprocessed Gradient		Ideal Gradient	
	Median	95% confidence range (interval)	Median	95% confidence range (interval)
Specific	-0.009	0.025 (-0.023 – 0.002)	-0.004	0.019 (-0.016 – 0.003)
Generalised	-0.008	0.026 (-0.025 – 0.001)	-0.004	0.021 (-0.016 – 0.005)

STARE dataset - Automated Algorithm				
	Unprocessed Gradient		Ideal Gradient	
	Median	95% confidence range (interval)	Median	95% confidence range (interval)
Specific	-0.006	0.011 (-0.014 – -0.003)	-0.002	0.005 (-0.005 – -0.000)
Generalised	-0.005	0.008 (-0.009 – -0.001)	-0.002	0.003 (-0.003 – 0.000)

Table 5.5: Effect of gradient response on d-norm based operator precision and accuracy. The gradients are used to estimate the boundary pixel orientations. A basic, unprocessed gradient is compared to an ideal gradient, obtained directly from the latent set

5.2.3 Discussion and analysis.

Unsurprisingly, the results from the clinical experiments above seem to be in agreement with the findings from the synthetic set. The d-norm validation operators again significantly outperformed both conventional and other state of the art methods in terms of both validation precision and accuracy. The conventional approach to validation by thresholding appears, if anything, to be even more unreliable in these experiments, leading to false conclusions regarding the true performance of algorithms, both in absolute terms, as well as in terms of segmentation-algorithm comparison. We shall not be repeating here discussion points made previously in the context of the synthetic experiments, to the extent that they are the essentially the same conclusions, but we shall instead focus only on insights that are unique to this particular dataset in the discussion below.

Validation of thin vs bulky structures

Pixels at the object boundary are generally more likely to be fuzzy than pixels at the core; indeed, one would expect true core pixels to be fully deterministic, since it is very unlikely that a pixel in the core of an object would contain more than one tissue types. In fact, one would do well to be circumspect of fuzzy segmentation masks that express fuzzy pockets in what should anatomically correspond to the 'core' of an object, if this makes no clinical sense. We discussed in chapter 3 that not all fuzziness necessarily represents the same thing, let alone specifically corresponds to tissue content, therefore care must be taken to ensure that if a fuzzy mask is intended to represent tissue content, that the segmentation algorithm has been created in such a way that the manner in which the fuzziness has been created is compatible with such an interpretation. If fuzzy pockets occur in such a mask regardless, under circumstances where it can not possibly represent partial volume effect, It may make more sense to 'fill' such pockets in as a post-processing operation, rather than perform fuzzy validation under an assumption of PVE.

We mentioned above that our choice of demonstrating on a retinal set over a standard Cine MRI segmentation was intentional in this instance, such that the relationship between fuzziness at the boundary denoting PVE, choice of validation operator, and validation precision / accuracy could be demonstrated more clearly, as this particular dataset has structures with a relatively high surface-to-volume ratio, implying a higher boundary-to-core pixel ratio, therefore making the contribution of boundary pixels towards the overall validation that much clearer.

For segmentations of large organ structures, one generally expects to see a relatively high overall-reported, absolute validation value, even with less sophisticated segmentation algorithms, due to the disproportionately large contribution of core pixels; conversely, this also explains why the latent truth validation for the clinical set in particular, even for the assessment of the human expert, was of a generally low value in this set (with a mean of the order of 0.6). More importantly, one might also expect less variability in validation values in general for the same reason, even between

algorithms of variable quality. However, a segmentation algorithm's *true* quality / superiority compared to other algorithms, generally boils down to its superior performance on exactly such boundary pixels; it is therefore even more important in such objects that appropriate validation methods are chosen that ensure accurate and precise discriminating ability at boundary pixels as well as core ones.

Performance of the directional t-norms

As with the synthetic set, both the generalised and context-specific d-norms seem to perform equally well. Oddly enough, in this particular dataset, the context specific d-norm seems to be slightly less accurate than the generalised one, even though it is more precise. The reason for this is not clear; it might reflect the simplicity of the 'linear boundary separation' model, or more generally reflect slight 'noise' in the validation process. But in general, as predicted from the theoretical analysis in the previous chapter, both d-norm types are much more accurate and precise than all the other methods investigated.

Effect of gradient estimation method in d-norm accuracy / precision.

From the results in table 5.5 we can see that, as in the synthetic set, a better gradient does indeed make the directional operator both more accurate and precise. However, interestingly, as with the synthetic set, we see that despite the simplicity of a Sobel approach, the d-norm based validation operator still leads to remarkable improvement in accuracy and precision over the other methods. Given the different characteristics of the two sets, we can safely conclude that this is a property of the validation operator itself rather than a chance occurrence due to the particular nature of a specific set. Again, this is consistent with the intuitive assumption that even lesser quality orientation information will lead to marked improvement, compared to not taking this information into account at all.

Performance of fuzzy validation operators based on standard t-norms

As with the synthetic set, the approach by Crum *et al.* (i.e. a Tanimoto coefficient implemented via the Gödel norms) performed second best, after the directional norms.

In the synthetic set, where the Łukasiewicz-based operator, gave a good median accuracy but a bad precision overall, we suspected this might represent the fact that bad overlaps were *over*represented in that set (since there was only a short range of angles for which overlap was good, and the remaining angles of rotation lead to relatively bad overlap), and that therefore it would generally be unlikely to give reliable results as an operator, given real-world examples.

We confirm this suspicion from the results of this clinical set (cf. table 5.4), where the Łukasiewicz operator is now shown to be very *in*accurate as well as imprecise, and in fact performs the worst out of all other operators. Intuitively, this makes sense, as in general one would expect 'real-life' segmentation candidates to have at least 'decent' overlap to begin with; it is very unlikely that a sophisticated segmentation algorithm will demonstrate 'terrible' overlap, such that a Łukasiewicz-based validation operator might have been a more fruitful approach.

Equally, this may be one of the reasons the Crum *et al.* approach seems to be the better choice out of the three standard t-norms from the outset, as the optimistic approach could be justified in the sense that it would rely on the assumption that the output from a sophisticated algorithm (or manual delineation) is more likely to be on the "decent-to-good overlap" side of the spectrum to begin with, rather than the "outright terrible overlap" side. In other words, it is a relatively reliable validation operator, but only under the assumption that the overlap is assumed to be of a relatively good standard in the first place (i.e. a chicken-and-egg problem).

As with the synthetic set, and as expected, the Product-based operator again returns validation estimates roughly between the pessimistic, Łukasiewicz-based operator, and the optimistic, Gödel-based operator. As such, its precision, predictably, also seems to be between the two other norms; given the large imprecision of the Łukasiewicz operator, we suspected in the previous chapter that this might make the Product operator unreliable for validation as well, and this is confirmed by the results of this experiment.

One might have expected that since both the Łukasiewicz and Gödel norms represent "extreme" scenarios of absolute pessimism and absolute optimism respectively, that the Product norm, representing a somewhat more 'moderate' view between the two, should have therefore generally resulted in a more "pragmatic", and therefore more accurate result; however, in reality this is not the case, since as we pointed out above, in most real-life scenarios we would expect the optimistic case to be closer to the truth in the first place. It makes sense, therefore, that if one of the more extreme views is actually *closer* to the truth, then a 'moderate' view would still be less accurate, to the extent that it is affected by the 'false' extreme to the same extent as by the 'true' extreme[3].

Performance of the conventional / thresholding approach

The thresholding approach is confirmed to be an unreliable operator for the clinical dataset as well, behaving in a very similar way as in the synthetic experiments:

- we see that, as in the synthetic experiments, in terms of its absolute output, it often violates the theoretical upper and lower bounds defined by the Gödel and Łukasiewicz norms respectively, particularly the upper bound (fig. 5.6 – compare to fig. 5.2).
- in terms of its overall distribution, it resulted in median values that were even *higher* than the corresponding Gödel distributions in both the human and algorithmic raters, implying that it tended to produce outputs that were artificially overoptimistic, beyond what is semantically meaningful in the context of fuzzy pixels.
- it had wider 95% confidence intervals — 8.0% / 8.8% of the Tanimoto operator 0–1 range for algorithm and human rater respectively, compared to 1.5% /

[3]In fact, this is another known fallacy, known as "the fallacy of the middle ground", or "the argument to moderation", i.e. the fallacy that the middle ground of two "extreme" positions is necessarily always closer to the truth compared to either extreme. This is a fallacious argument, since one of the extreme positions may be true, or much closer to the truth than the other, and therefore the middle ground will deviate from the truth to the extent that it deviates away from the 'true' extreme).

4.3% for the Gödel norm, and 0.4% / 2.0% for the directional norms — showing it is imprecise as a validation operator.

In the previous section we made the point that the threshold-based operator's low precision, manifesting as an unpredictable, zigzag line around the latent truth (as in fig. 5.6 – compare to fig. 5.2), which is also observed in the clinical set, could potentially lead to false conclusions regarding both the absolute performance of a segmentation algorithm / result, but also, more importantly, regarding the relative quality between two segmentation algorithms / raters.

Experiment No. 11 in the clinical set is a good example of this (see fig. 5.9): The human rater is more accurate than the automated algorithm according to the latent validation truth, with a Tanimoto value difference of 0.110. The Gödel and directional norms are both reasonably close to this value (0.115 and 0.098 respectively). However, the Threshold-based operator results in a validation output difference of only 0.014, essentially implying the automated algorithm is essentially almost as good as the human rater. Worse, if we were to take the operator's precision into account when interpreting this result (which we mentioned was around 7–9% of the Tanimoto range in our experiments) we might even conclude that our confidence interval for the real difference would also include scenarios where the automated algorithm is deemed to have outperformed the human operator. And, while it is "unfortunate" that no actual instances were observed where the Threshold-based operator resulted in an *inverse* conclusion for the comparison between two segmentation candidates in our experiments, ("unfortunate" in the sense that this would make for an even more compelling example to help drive this point home), it is highly likely that such instances would have been observed in the presence of a larger dataset, given the wide confidence intervals observed.

Performance of the remaining state of the art methods investigated

The results from the clinical dataset support the findings from the previous section with regard to the quality of either the Yi or the Chang operator, in that they do

not seem particularly accurate / precise compared to the other operators.

5.3 Conclusion

We have shown experimentally that directional norms and their respective validation operators are reliable in terms of validation precision and accuracy, as compared to a known latent truth, and perform significantly better than conventional and other state of the art methods, as predicted by the theoretical contributions of the preceding chapter.

We reiterate that these experimental findings (as well as the underlying concepts presented in chapter 4) are novel work, and that [22] *represents the first publication of the kind.*

Furthermore, we have shown conclusively that, as suspected, the conventional, thresholding-based validation approach is unreliable, and could lead to false conclusions when evaluating the performance of algorithms, both in absolute terms, but also in terms of the relative performance of two or more algorithms.

We hope this should serve as a message to the medical image segmentation community to abandon the thresholding approach as the de facto standard for validation of fuzzy and probabilistic segmentation algorithms, and also to highlight the importance of research and appropriate evaluation of validation algorithms, to the same degree and rigorous standards as the evaluation of segmentation algorithms themselves.

Summary

- We evaluate the accuracy and precision of d-norm based fuzzy validation operators, compared to the conventional and state of the art validation operators described in the previous chapter, on a synthetic and clinical set. We show that d-norm based validation operators significantly outperform their competitors.
- In particular, we emphasize the fact that the conventional threshold-based approach is both inaccurate and imprecise, and violates the theoretical bounds outlined in the previous chapter, which can lead to false conclusions; the chapter ends with a plea to the medical imaging community to abandon the widespread use of thresholding for validation simply on the basis of convention, and to consider more appropriate methods specific to the datasets used on a per case basis.

"Sometimes the truth is arrived at by adding all the little lies together and deducting them from the totality of what is known."

— Terry Pratchett; *Going Postal*

6

Beyond validation: characterising modes of segmentation failure

In this chapter we discuss ways of obtaining useful information regarding 'in what way', rather than just 'to what extent' a segmentation outcome fails to match a gold standard or not.

To this effect, we describe a number of tools and approaches, each reflecting a different kind of question that can be asked by the researcher to discern particular modes of segmentation failure, such that they can be answered in both a qualitative and a quantitative manner, and more generally be used as part of a more informative evaluation strategy for segmentation algorithms.

These consist of: a number of 'local-performance maps', each looking at a particular aspect of performance, and how it varies over the image domain; 'fuzzy direction masks' and 'fuzzy distance masks' which can be used with performance maps to query the extent to which a segmentation performs well or fails in a particular direction or at a particular distance respectively, with respect to a gold standard or other object of interest; a fuzzy generalisation of the Hausdorff distance for use with fuzzy segmentation objects based on the above; and 'validation sweeps' for examining the extent to which a segmentation fails due to the abnormal presence of a particular fuzzy feature.

Contents

Validation is essentially an attempt at quantifying the quality and performance of a segmentation algorithm, and by extension the quality and reliability of its outputs. However, validation as we have seen it so far, can tell us *whether*, and to what extent a segmentation is of good quality or not *as a whole*, but will probably tell us very little about *why*, *how*, *where*, and to what extent *locally* it is better or failing to achieve the desired outcome. Is the segmentation shifted with respect to the Gold Standard? Scaled? Deformed? Rotated? Is it more locally accurate on certain parts of the object, and less so on others?

As mentioned earlier in section 4.1.2, to an extent this can be mitigated by selecting an appropriate validation approach, e.g. evaluating using object overlap, as opposed to contour distance, or clinical parameters, or a combination of the three; however, this still gives relatively little information on the particular mode in which a segmentation succeeds or fails (whether locally or otherwise) to the

extent that it does.

This chapter therefore discusses ways in which a researcher might ask relevant 'questions' expressed in a suitable mathematical manner, in order to discern and characterise particular modes of segmentation failure in a qualitative and quantitative manner, thereby providing a more comprehensive evaluation strategy for the performance of segmentation algorithms. We start this discussion by describing ways in which various aspects of performance can be assessed at the *local* level (or in other words, how this varies, or is distributed, throughout the image domain). We then use this insight to investigate the variability or importance of particular areas in the image domain in that respect, which are spatially relevant with respect to the gold-standard (or any other point or object of interest), such as located in a particular direction or at a particular distance from such an object. We extend the discussion on distance to how this relates specifically to notions of distance from fuzzy objects, and propose on that basis a 'fuzzy' generalisation of the Hausdorff distance metric commonly used for evaluation in binary image objects. We end the chapter by introducing the concept of a 'validation sweep', and showing how this can be used to determine segmentation failures consistent with the unwanted presence of particular features in the segmentation output.

6.1 Local-performance maps for assessing spatial variability in performance

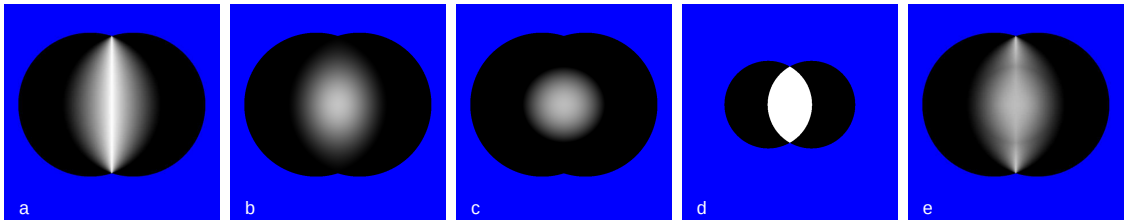
We mentioned in chapter 3 that it would be desirable to know if a particular segmentation algorithm is "stronger" in certain areas (consistently or otherwise), and "weaker" in others, as this would be useful information to put to further use, e.g. by retaining strong areas of an algorithm and discarding weaker ones.

6.1.1 Pixelwise Tanimoto Coefficient masks as measures of local overlap / accuracy

One way to get such local information in the presence of a gold standard mask, is to evaluate local (i.e. pixelwise) versions of the validation operators. A Pixelwise Tanimoto Coefficient (pT_c) can be defined as:

$$pT_c = \frac{\mathbf{s} \wedge \mathbf{g}}{\mathbf{s} \vee \mathbf{g}} = \forall i : \frac{s_i \wedge g_i}{s_i \vee g_i} \tag{6.1}$$

where $\wedge$ and $\vee$ correspond to appropriate t-norms and t-conorms respectively. Since this is defined per pixel, the output of the above operation is a fuzzy mask of the same size as the segmentation and gold-standard masks, except that it is *undefined* for any pixel i where both the intersection and union operations are exactly zero; fig. 6.1 demonstrates the pT_c mask for some of the fuzzy operators already discussed in this thesis, using the 'fuzzy discs' example set from section 2.4.2 (figs 2.3-2.4, p. 43-44).



	Gödel	*Product*	*Lukasiewicz*	*Threshold*	*Directional*
$\overline{pT_c}$:	0.195	0.122	0.078	0.249	0.159
T_c :	0.307	0.228	0.155	0.249	0.265

Figure 6.1: Pixelwise Tanimoto Coefficient (pT_c) masks for a range of fuzzy validation operators, evaluated on the validation set $(\mathbf{s}, \mathbf{g})$ from figs 2.3-2.4 (p. 43, 44). Pixels in blue denote pixels where the pT_c is undefined (i.e. intersection and union values are both zero). Average pT_c values ($\overline{pT_c}$) and Tanimoto coefficients (T_c) are reported for each operator.

Since the pT_c denotes a *distribution* rather than a single statistic, we can easily obtain summary statistics over it. For instance, If we denote this set of undefined pixels as U (and therefore the set of 'valid' pixels by its complement U^c), then an average pT_c value (denoted $\overline{pT_c}$) can be obtained as:

$$\overline{pT_c} = \frac{1}{|U^c|} \sum_{i \in U^c} \frac{s_i \wedge g_i}{s_i \vee g_i} \tag{6.2}$$

For binary intersection and union masks (and by extension for the thresholding operator at any threshold), the standard Tanimoto coefficient (T_c) and the $\overline{pT_c}$ coincide; however, for fuzzy masks and operators in general, the two are generally not equivalent, since the $\overline{pT_c}$ is constructed as a *sum of ratios*, as opposed to the T_c, which is constructed as a *ratio of sums* (eq. 2.2, p. 30). There is no straightforward analytical relationship for the discrepancy between the two, except perhaps for the casual observation that the discrepancy is likely to be smaller for pT_c masks with lower entropy.

However, arguably, any single statistic we evaluate from the pT_c mask, such as a mean or median pT_c, is unlikely to be particularly more useful as a single measurement in itself compared to the T_c. The pT_c mask is more useful to us in its mask form, in that it may serve as a measure of pixelwise certainty for the segmentation, reflecting accuracy at each pixel with respect to the gold-standard mask. A simple threshold over this pT_c mask could reveal areas where the algorithm underperforms, either in absolute terms (i.e. with respect to a specific, arbitrary threshold), or in relative terms (i.e. with respect to its median or a particular percentile, partitioning the pT_c distribution into 'weaker' and 'stronger' constituent parts in terms of performance). Fig. 6.2 demonstrates this visually.

6.1.2 'Regional' Tanimoto Coefficient masks

Sometimes, rather than obtain local accuracy in a pixelwise manner, it is more useful to describe the accuracy represented by a pixel location in terms of how accurate the 'region' surrounding that pixel is. To this purpose, a suitable fuzzy mask denoting such a region of interest when centred over a pixel, can be defined and applied[1] to both the segmentation and gold-standard masks separately for each pixel, such that a 'regional' Tanimoto coefficient can be calculated at each position. When viewed in this manner, the pT_c itself becomes a 'regional' T_c, acquired using a region-mask

[1] We discuss in more detail later in this chapter how one might 'apply' such a mask and what the relevant semantics are.

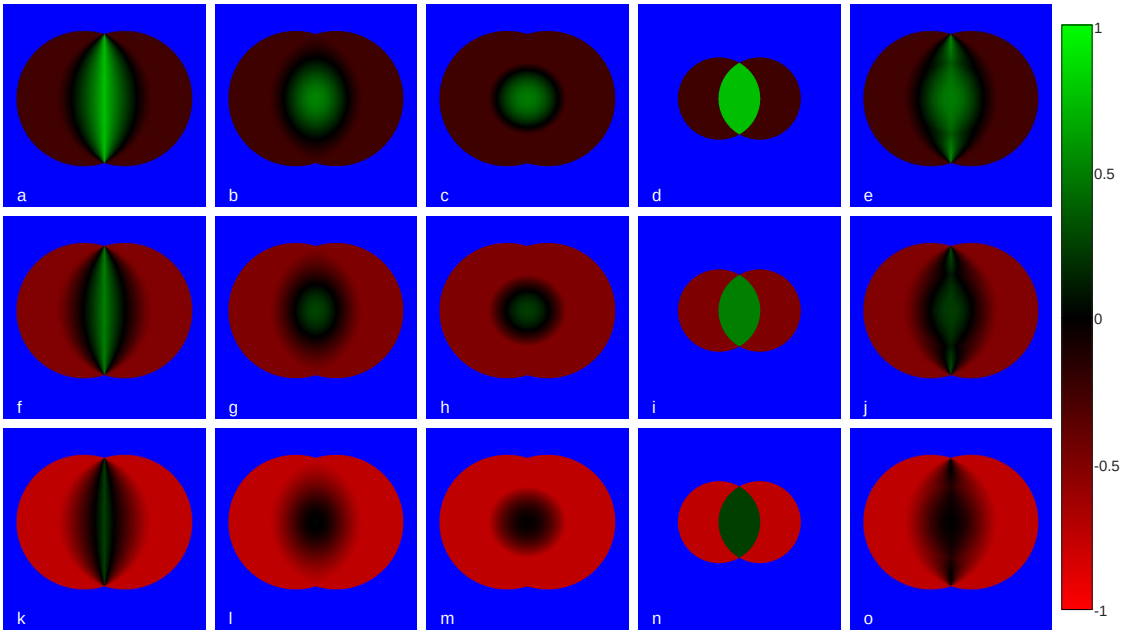


Figure 6.2: A pT_c-based map, partitioned into 'strong' (green) and 'weak' (red) areas in terms of segmentation performance with respect to a gold standard. The partitioning is effected via a threshold on the original pT_c map (see fig. 6.1); here we show three partitioning schemes corresponding to pT_c thresholds of 0.25 (top row), 0.5 (middle row), and 0.75 (bottom row) respectively. Pixels occurring *exactly* at the pT_c threshold chosen in each case appear as pure black; a positive pT_c difference with respect to the threshold (i.e. areas where the algorithm is performing *better* than this threshold, as per the original pT_c mask) is shown in increasing intensities of green (with pure green on this scale signifying the maximum possible *positive* difference value of 1); negative differences (i.e. areas where the segmentation is *under-performing*) are shown as increasing intensities of red (with pure red signifying the minimum possible *negative* difference value of -1). Finally, pure blue pixels denote areas where the pT_c is undefined.

comprising a single pixel with a value of '1'. Fig. 6.3 compares the (Gödel) pT_c map with two regional T_c maps generated using region masks of increasing size.

We have made this distinction between 'pixelwise' and 'regional' approaches for obtaining local-performance maps for the sake of completeness. However, we will make no further mention of 'regional' approaches from here on, other than to say that any pixelwise approach could be substituted for a regional one if desired; for the sake of simplicity, where we refer to local maps with respect to the remaining concepts discussed in this chapter, we will limit ourselves to

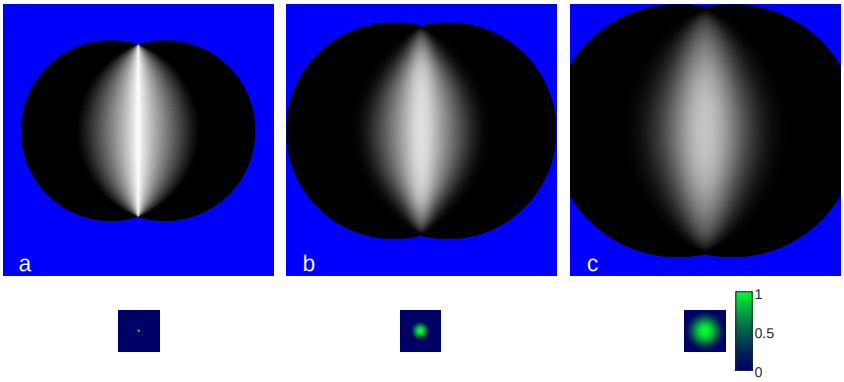


Figure 6.3: Regional Tanimoto Coefficient maps for different-sized region masks.
a. The pT_c map and its corresponding 'single-pixel' mask
b. A regional T_c map obtained for a '50-pixel radius fuzzy disc' region mask
c. A regional T_c map obtained for a '100-pixel radius fuzzy disc' region mask

demonstrating only using pixelwise methods.

6.1.3 Symmetric difference masks as measures of local misclassification

Another way to quantify the local performance of a segmentation with respect to a gold standard is to evaluate a simple measure of (pixelwise) discrepancy for the two (fuzzy) sets (i.e. as opposed to evaluating a bespoke measure of overlap accuracy). The most natural fuzzy set operation for this is the *symmetric difference* (implemented as *exclusive disjunction* in the context of fuzzy masks / pixelwise operations; see section 2.4.1, p. 29 and 32).

The symmetric difference gives somewhat different information to the pT_c, which is arguably a bit more intuitive, at least in the visual sense: instead of detecting areas of over- or under-performance in terms of *overlap*, the symmetric difference effectively visualises the degree of misclassification for that pixel, i.e. by how much the segmentation missed its target value for that pixel (whether by over- or under-estimating the 'true' value), expressed as a measure of 'distance' between the two fuzzy values, (which, as per chapters 4 and 5, could well have a geometric interpretation). Fig. 6.4 demonstrates the symmetric difference for the same 'fuzzy discs' set used with the pT_c, both in absolute terms, and partitioned into 'strong'

and 'weak' areas with respect to two different 'misclassification distance' thresholds (which could, for instance, be used as decision thresholds in appropriate decision-making schemes).

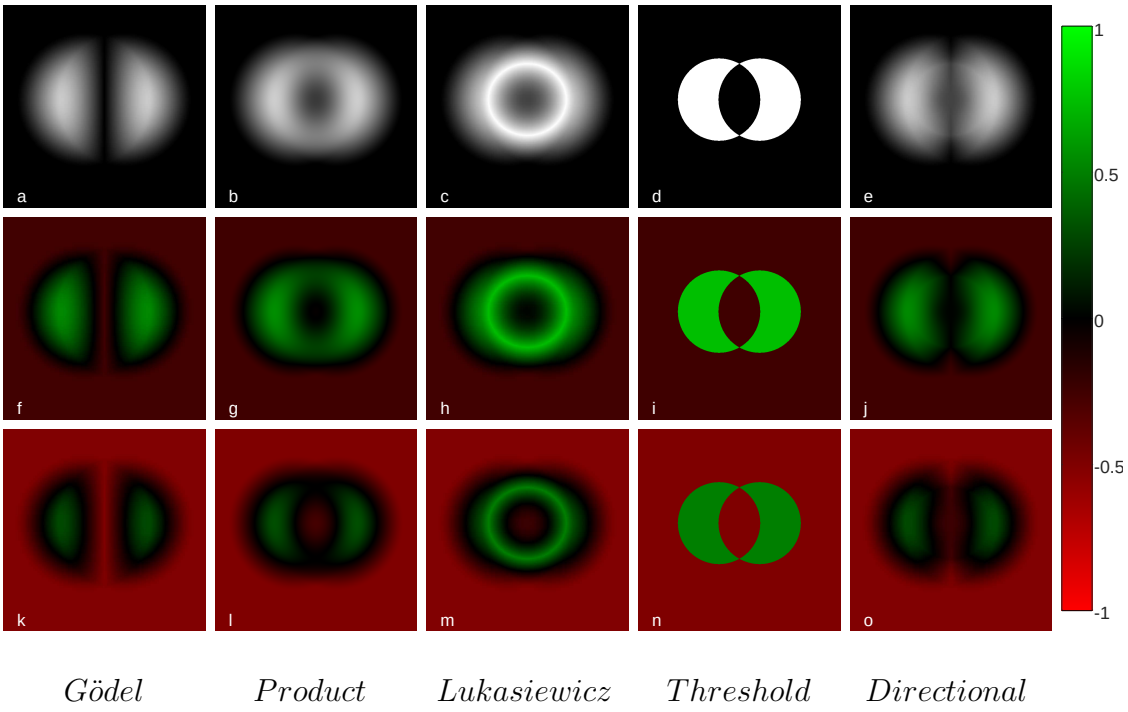


Figure 6.4: Symmetric difference for the 'fuzzy discs' set (both absolute and partitioned relative to a threshold). Top (a-e): Symmetric difference mask, shown in a standard grayscale colourmap (where pure black denotes a symmetric difference value of 0, i.e. denoting no misclassification, and pure white denotes the maximum symmetric difference value of 1, i.e. denoting full misclassification). Middle (f-j): Values above (green) or below (red) a symmetric difference threshold of 0.25 (i.e. denoting misclassification above vs below that threshold), using the same colour-scheme as with fig. 6.2. Bottom (k-o): Same as f-j but for a 0.5 threshold.

Note that, much like the calculation of 'false positives' (F_+) and 'false negatives' (F_-) (see section 2.4.1, p. 37), symmetric difference can be expressed using a combination of 'base' set operations (i.e. complementation, intersection, and union) *exclusively*, expressing an "either / or but not both" set relation as:

$$S \triangle G = (S \cup G) \cap (S \cap G)^c \tag{6.3}$$

or, alternatively, as a simpler 'distance-based' formulation between the union and

intersection of two sets, i.e.:

$$S \triangle G = (S \cup G) - (S \cap G) \tag{6.4}$$

For the purposes of this section, we focus on the latter formulation, partly due to the simpler expression involved, but more importantly as it has a more intuitive, straightforward, and relevant interpretation in the geometrical context of fuzzy pixels as discussed earlier, as the local extent to which tissue distributed inside the pixel has been misclassified; however, we note that, as with F_+ and F_- (see footnote on p. 119), for *fuzzy* sets specifically, the two formulations are not necessarily equivalent (both in a strict semantic sense, and in terms of their output).

6.1.4 Over- and under-segmentation, versus false positive and false negative fuzzy masks

We can use the symmetric difference mask to isolate *effective* 'false positive' and 'false negative' areas in the segmentation, or more accurately, areas of *over*- or *under*-segmentation at each pixel. Namely, the set representing the symmetric difference between sets S and G, can be further subdivided into two areas: one where the misclassification distance is due to *over*-segmentation (which we will denote S_+) and one where it is due to *under*-segmentation (S_-).

We make the distinction here explicit from the *actual* false positive (F_+) and false negative (F_-) fuzzy sets, evaluated via their standard set definitions (e.g. as per section 2.4.1, p. 37), in that ordinary F_+ and F_- components may coexist in the same pixel, depending on the fuzzy operator used and the particular semantics underlying it. By way of contrast, S_+ and S_- defined via the symmetric difference, are mutually exclusive by definition. Using mask notation:

$$\begin{aligned} \mathbf{s}_+ &= \mathbf{s} \veebar \mathbf{g}, \quad \forall i : s_i > g_i \\ \mathbf{s}_- &= \mathbf{s} \veebar \mathbf{g}, \quad \forall i : s_i < g_i \end{aligned} \tag{6.5}$$

Fig. 6.5 shows the distinction between S_+ / S_- and F_+ / F_- components.

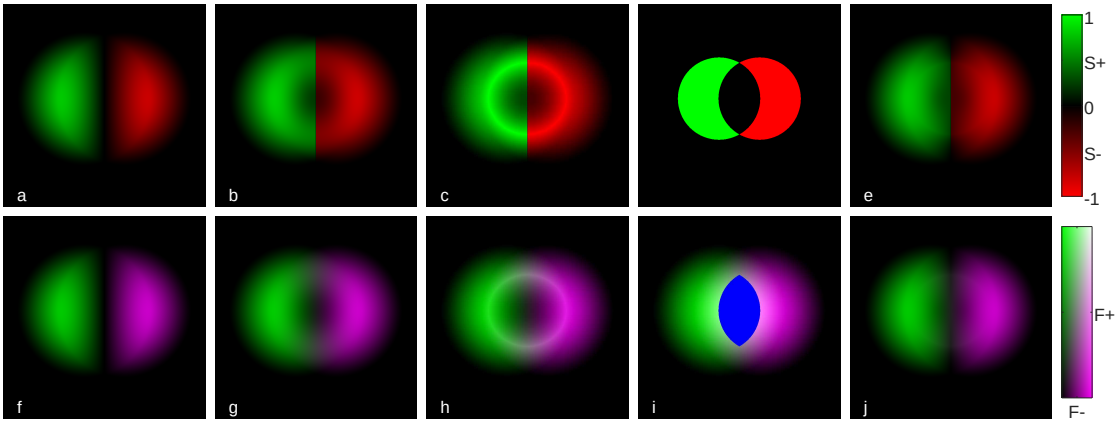


Figure 6.5: Over- / under-segmentation components versus false positive / false negative components. The top row shows the oversegmentation mask S_+ using a green scale, and the undersegmentation mask S_- using a red scale; since the two sets are mutually exclusive, they can be presented simultaneously on a single image. The bottom row is a fused RGB colour image (as described in fig. 2.2, p. 35) of the F_+ (green channel) and F_- (red and blue channels) masks. Pixels where the red, green, and blue channels all have the same value, appear grey, with brightness (from black to white) depending on the actual value from 0 to 1 (see also the 2D colourbar on the right; in particular note the gray values in the diagonal). The solid-blue area seen in the bottom row for the Threshold case marks areas of invalid F_+ / F_- values, due to the nature of the Threshold operation leading to 'intersection' values that are larger than the individual S and G sets themselves at those pixels.

6.2 Fuzzy spatial / anatomical relationship masks

We mentioned in ch. 3 (see figs 3.1 and 3.2, p. 64), that we can represent fuzzy (i.e. vague) spatial / anatomical relationships such as "to the right of", "inferolateral to", "proximal or distal to" etc, either from a single point or from a whole-object boundary, as a fuzzy set R (and corresponding mask $\mathbf{r}$), whose membership function quantifies such a relationship over the image domain (i.e. Ω), according to a suitable mathematical definition.

6.2.1 Evaluation of segmentation performance in a particular direction with respect to the gold standard (or other object of interest)

For example, one way to mathematically express the (vague) spatial relationship "left of point x" as a fuzzy set, could be via a *cosine of the angle* approach (fig. 6.6a), which involves setting the value of each pixel in the image domain Ω to the cosine of the angle difference from that pixel (constrained to $\pm 90°$) with respect to point x and the vector $(-1, 0)$ (i.e. the *exact* direction 'left'). This has the effect that pixels *directly* to the left of point x are given the value 1, with values of surrounding points gradually reducing their value as angular distance increases, such that they drop to 0 for pixels directly 'above' or 'below' x.

This of course, is just one of many possible approaches; any other approach will be equally valid to the extent that it is a reasonable and valid interpretation with respect to the relationship it represents and the context in which it is meant to be applied. In fact, such a relationship mask need not be fuzzy.

For instance, a broader, binary (i.e. crisp) interpretation of 'left' could partition the image into two parts (i.e. a 'left' one and a 'right' one). In general, such a *linear separation boundary* approach (fig. 6.6b) would divide the image domain Ω crisply into two parts (i.e. a binary 'partitioning') via a dividing line (or plane in the 3D case), passing through point x perpendicularly to the direction expressed by the relationship.

Conversely, one could create a very restrictive definition of 'left', only detecting pixels that are *exactly* left of point x in a crisp manner, i.e. all pixels falling on the horizontal line segment starting from point x and ending at the left image edge. In general, such a *line segment* approach (fig. 6.6c) would only capture pixels falling on a line segment from a point x to a single point on the image edge in the specified direction.

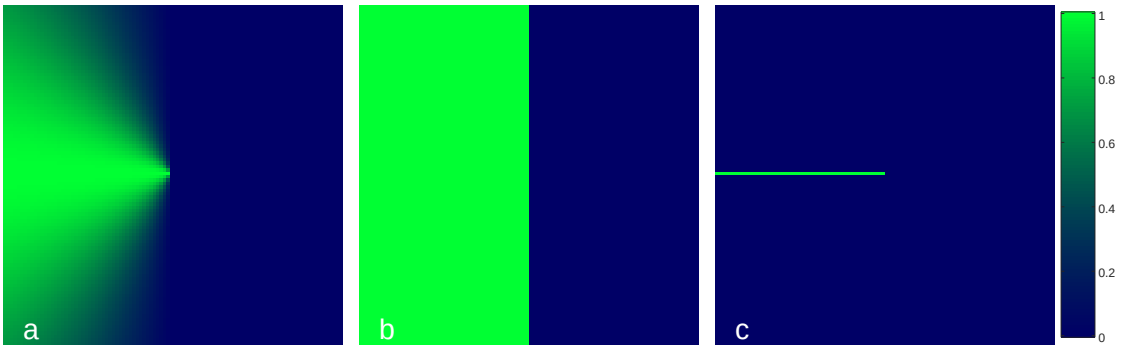


Figure 6.6: Fuzzy mask representations of the spatial relationship 'left' with reference to a single pixel located at the image centre. **a)** Spatial relationship 'left' represented using a *cosine* approach. **b)** Spatial relationship 'left' represented using a *linear separation boundary* approach. **c)** Spatial relationship 'left' represented using a *line segment* approach.

In the context of uncertainty, we discussed the use of such masks as potential measures of clinical / anatomical uncertainty, by defining relationships from prior clinical knowledge, e.g. representing statements such as "it is highly unlikely that heart tissue would be present outside the ribcage", or "the left ventricle is to the (anatomical) left of the right ventricle" via correspondingly appropriate fuzzy masks with respect to suitable landmarks.

In the context of describing local performance and modes of failure of a segmentation with respect to a gold standard, instead of defining such spatial relationship masks from 'known landmarks', we can instead define them with respect to the ground truth. This could enable them to be used to ask questions like "to what extent is a segmentation failing with respect to this particular spatial relationship"; e.g. "to what extent is the myocardial segmentation algorithm prone to false positives towards the liver?".

Local-performance measures, such as the ones discussed above, in conjunction with such spatial / anatomical relationship masks, allow us to do just that, by using a simple intersection between the two, while still maintaining the semantics of the different fuzzy operators used to generate such measures.

Fig. 6.7 demonstrates the concept for the 'fuzzy disc' set, asking the question "which parts of the segmentation object S itself are located *left* of the gold-standard G, as

assessed via the fuzzy relationship mask **r**, defined using a cosine approach, both with respect to G as a single representative point of reference (i.e. its centroid), and as a whole-object.

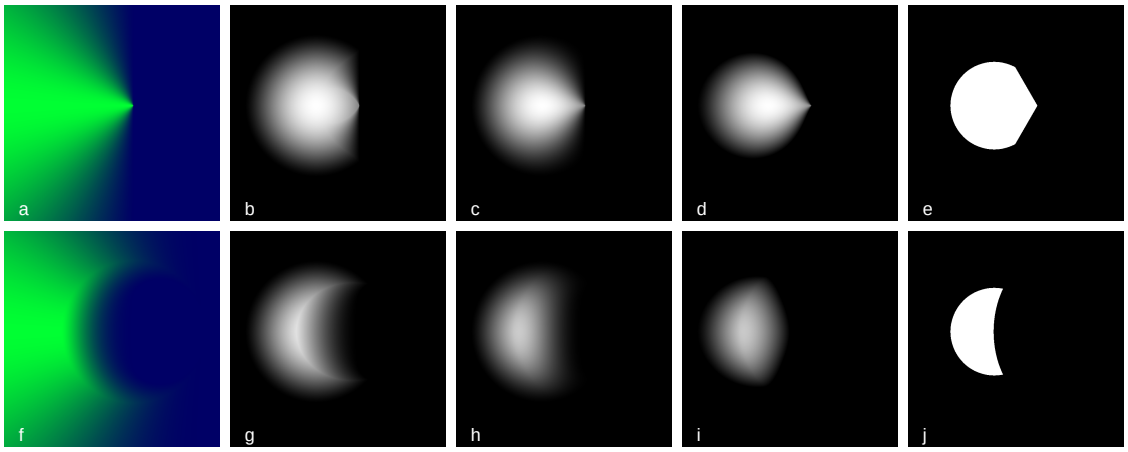


Figure 6.7: Fuzzy mask representation of the statement "Part of segmentation S that is left of the gold standard G". *Top row:* **a.** The fuzzy spatial relation mask R representing "left of G", defined using a *cosine* approach and the centroid of G (i.e. the whole of G is being represented by a single representative point). **b-e.** The intersections of R with S (as defined in fig. 2.3a, p. 43), implemented using Gödel, Product, Łukasiewicz and Threshold semantics. *Bottom row:* same as above, but for R defined via G as a whole object (i.e. the resulting mask is the union of all point-derived fuzzy masks generated for all pixels in G in a manner similar to panel 'a', but weighed by their fuzzy values, and with the original object G subsequently subtracted from this union). Note: panels **a** and **f** denoting pure spatial relationship masks use the same colourmap as fig. 6.6, to distinguish them from the remaining masks which represent the application of R on S.

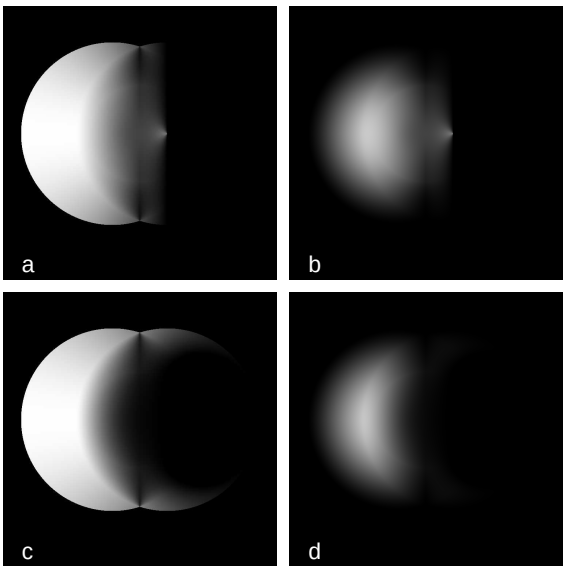
Note that since both the segmentation and relationship masks are fuzzy masks, when we talk about an 'intersection', we are again talking about an appropriate choice from a selection of t-norms, each with their own semantics. However, it is important to note here that the semantics in this case differ, and are distinct from the *spatial* semantics we attributed to the different operators in previous chapters in the context of fuzzy / PVE pixels and their overlap. In this case, rather than implying subpixel-distribution of any kind, the fuzzy value of each pixel in R represents the *strength* of the relationship represented by R for that pixel. Therefore, the semantics of the various t-norms, representing the combined statement "the

extent to which each pixel in the image domain Ω belongs in S *and* is left of G (as per R)" change, where the t-norms evaluate such a relationship as follows:

- a Gödel t-norm evaluates the combined statement using a *weakest-link* approach between sets S and R; e.g. for $s = 0.5$, and $r = 0.8$ (i.e. mostly, but not wholly / exclusively left of G), the combined statement for that pixel evaluates to 0.5 which is the minimum of the two.
- a Product t-norm evaluates the combined statement using an "S weighted by R" approach; e.g. for s and r as above, the combined statement evaluates to the product of the two, that is $0.5 \times 0.8 = 0.4$
- a Łukasiewic t-norm evaluates the combined statement as the degree to which the effect of S and R added together manages to overcome a certain threshold (which defaults to 1 but could potentially be parameterised further): e.g. for s and r as above, the combined statement evaluates to $0.5 + 0.8 - 1 = 0.3$
- a Threshold "t-norm" evaluates the combined statement in a binary manner, by selecting only pixels where both masks individually exceed a predefined threshold; e.g. for s and r as above, and for a 0.5 threshold applied to both[2], the combined statement evaluates to *true*.

Fig. 6.8 shows the pT_c and *symmetric difference* masks (obtained using Directional, i.e. *d-norm* semantics) after 'applying' the same relationship R — that is, "left of G" — again defined both with respect to G as a single representative point of reference (i.e. its centroid), and as a whole-object, and with the intersection between pT_c / symmetric difference and R calculated using Product semantics. Note that while the images shown here are two-dimensional, the concept applies naturally to three-dimensional spatial relationships as well.

[2]Note that this does not necessarily have to be the case: two (or more) sets can have distinct thresholds. Similarly for the other operators, modifiers can be introduced to weigh one set's contribution more in relation to the others, or modify all sets proportionally to produce more 'permissive' / 'strict' intersections, like we did in fig. 3.2 (p. 64) for the Łukasiewicz operator



$1 - pT_c$ Symmetric diff.

Figure 6.8: Overlap inaccuracy (panels **a**,**c**; evaluated as $1 - pT_c$) and misclassification (panels **b**,**d**; evaluated as $S \triangle G$) occurring "left of the gold-standard G". *Top:* Intersection with fuzzy mask R, where R represents the relation "left of G" as defined from the *centroid* of G using a *cosine* approach. *Bottom:* Intersection with fuzzy mask R, where R represents the relation "left of G", as defined from G as a *whole object* using a *cosine* approach (see fig. 6.7).

6.2.2 Evaluation of spatial relationship around the gold standard

For any given fuzzy mask F and a spatial relationship mask R, since their intersection is a fuzzy mask itself, the "proportion", or "degree" to which F is consistent with R, is proportional to the cardinality (i.e. the 'pixel mass') of their intersection (i.e. $|F \cap R|$). By varying the direction represented by R, this allows us to evaluate the extent of the presence of the segmentation itself — or of its "strengths and weaknesses" as represented by suitable local-performance measures — in a 360° arc around the gold standard. For example, when applied to 'overlap inaccuracy' or 'misclassification' masks as above, we could effectively evaluate the variability and the directional extent of the misclassification or overlap inaccuracy around an object, expressed as a proportion with respect to that measure over the image domain as a whole. Fig. 6.9 shows three 'polar plots' demonstrating what this might look like for the 'fuzzy disc' set for each of the three spatial relationship approaches described above (i.e. *cosine*, *linear separation boundary*, and *line segment* approaches). In the case of the *cosine* and *linear separation boundary* approaches (i.e. figs 6.9a and 6.9b) it makes sense to express the spatial extent of F in a particular direction as the ratio $\frac{|F \cap R|}{|F|}$, i.e. the proportion of F's 'pixel mass' that is present in that direction,

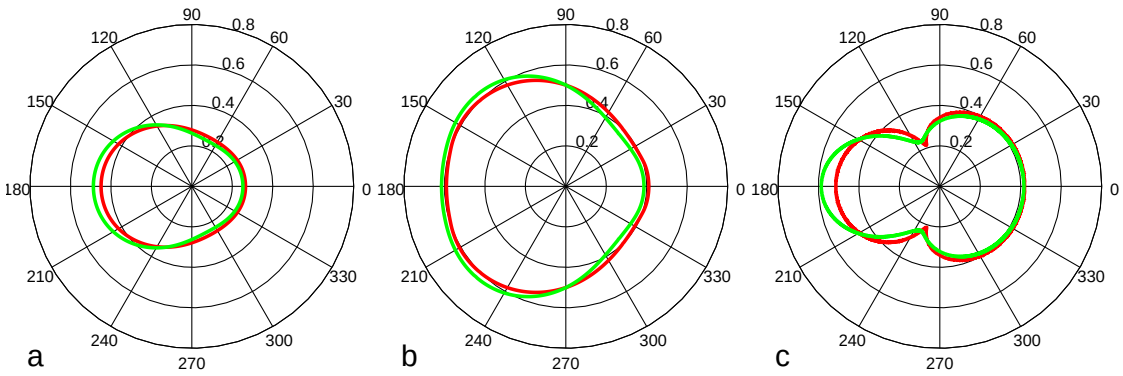


Figure 6.9: Overlap inaccuracy (red) and misclassification (green) as per fig. 6.7 for a 360° arc around the ground-truth centroid using: **a)** a *cosine* approach, **b)** a *linear separation boundary* approach, and **c)** a *line segment* approach, resulting in a *polar profile* (i.e. the approaches corresponding to the fuzzy direction masks in fig. 6.6, p. 188).

with respect to F as a whole. For example, in fig. 6.9a, 50% of the symmetric difference satisfies the relationship 'to the left of G', and 25% satisfies 'to the right of G' respectively. Note that these two numbers are *not* expected to add up to 100%, since for the *cosine* approach the fuzzy-relationship masks corresponding to 'left' and 'right' are not mutually exclusive. However, this *is* the case for the *linear separation boundary* approach (with %60 and %40 values respectively).

Polar Profiles

The *line-segment* approach (fig. 6.9c) is a bit different. With this approach, the area covered by the line-segment mask R is expected to be much smaller than F (i.e. $|F \cap R| \ll |F|$), therefore it makes less sense to use $\frac{|F \cap R|}{|F|}$ as our measure as this would lead to very small, hard to interpret values. In this case, the intersection of F with R rotated by an angle θ represents the numerical approximation for the operation where:

- Mask F is interpreted as a *continuous* function $f(x, y)$ over the image domain Ω (e.g. obtained here by interpolation of F)
- The line segment in mask R represents a linear path $C_\theta(t)$ placed at an angle θ in the image domain Ω, where:
  - t lies in the interval $[0, 1]$ (i.e. $t \mapsto [0, 1]$) representing points along this path

  - C_θ is a parameterisation, mapping from points along this linear path to their corresponding coordinates in the image domain, i.e. $C_\theta : t \mapsto (x, y)$, where $(x, y) \in \Omega$

  such that C_θ starts from the origin of interest ($t = 0$) and terminates at the image boundary ($t = 1$), and where for a given origin, the linear path C_θ in Ω is fully defined by its angle θ.

- The values of f along the path C_θ capture the distribution of mass along that particular path (and by extension, for that particular angle θ); the *total* mass r along any such path is then given by the *line integral* of f along C_θ $\Big($ i.e. $r(\theta) = \int_0^1 f(C_\theta(t))\mathrm{d}t$ $\Big)$; the polar diagram resulting from plotting r as a function of increasing θ results in a closed curve, whose area A can be evaluated via a *polar integral*, i.e. $A = \frac{1}{2}\int_0^{2\pi} r^2 \mathrm{d}\theta$
- Since the value of r represents total mass along C_θ irrespective of how it's distributed along this path, the area contained in any segment $\Delta\theta$ in the polar plot therefore also represents the mass contained in f for that equivalent segment, irrespective of how it's distributed within it; the total area A similarly corresponds to the total mass of f as a whole[3].

It makes more sense, therefore, if we need to normalise the resulting values for r, to do so such that the area of the closed curve sums up to 1, effectively treating $r(\theta)$ as the probability density function of a polar (or *circular*) distributed variable. In other words we normalise the obtained values for r by dividing with a constant $\alpha = \sqrt{A}$.

In the context of our numerical approximation on mask F, if the number of pixels at the image periphery[4] is equal to K, then:

[3]Compare this to the previous two approaches where this is not the case. In other words, the area inside the closed curve does not have any particular meaning or significance in the case of the *cosine* and *linear separation boundary* approaches

[4]i.e. the total number of pixels contained in the topmost + bottom rows and the leftmost + rightmost columns of the image domain; counting the corner pixels only once, then for an $M \times N$ image domain, there are $K = 2M + 2N - 4$ pixels at the periphery, from which to obtain angle measurements with respect to a centroid.

- $\{\theta_1, \theta_2, \ldots, \theta_K, \theta_{K+1}\}$ represents the angles for all K periphery pixels (with respect to the selected origin) from $\theta = 0$ to $\theta = 2\pi$ (inclusive).
- $\{r_1, r_2, \ldots, r_K, r_{K+1}\}$ are the respective approximations to the line integral mentioned above, evaluated numerically as the sum of pixel values along the lines corresponding to each θ_k, using interpolation where appropriate (i.e. to evaluate fuzzy values as occurring at fixed, pixel-width distances along the line).
- Area A is then approximated as the sum of K small triangles formed by each $\Delta\theta_k$ interval and the respective 'sides' r_k and r_{k+1}, i.e: $A = \sum_{k=1}^{K} r_k \; r_{k+1} \; \sin(\theta_{k+1} - \theta_k)$. The extent to which A differs from $|F|$ reflects the numerical error from this approach.
- If normalisation is desired, then the final normalised set of values $\{r'_1, r'_2, \ldots, r'_{K+1}\}$ is obtained by dividing by $\sqrt{A}$, i.e. $\forall k \in \{1 : K\} : r'_k = \frac{1}{\sqrt{A}} r_k$.

Like all probability density functions, this approach is more suitable for evaluating *intervals*, rather than for evaluating intersection mass in the direction of particular, individual angles. More generally, it is useful for describing the shape of the circular distribution as a whole; when used this way, we refer to the resulting plot therefore as the *polar* or *directional profile* of the mask, as defined with respect to a particular origin.

6.2.3 Evaluation of segmentation performance at a particular distance from the gold standard (or other object of interest)

So far, we have evaluated segmentation objects and performance measures for particular directions, using suitable fuzzy spatial relationship masks that denote 'direction' with respect to a single point or whole object. It should be straightforward to see that we can apply this concept to many other spatio-anatomical relationships, so long as we can express them as suitable fuzzy masks, or more generally via suitable mathematical functions.

For instance, fuzzy masks representing the concept of *distance* (in vague terms or otherwise) with respect to the gold-standard or some other landmark, can be straightforwardly expressed via a normalised version of the standard distance transform (described below). Such a fuzzy mask could be used to express anatomical relations such as "larger / smaller than expected" or "proximal / distal to an object of interest", enabling us to interpret a segmentation result's validity with respect to clinically-relevant questions, such as "In a myocardial segmentation, to what extent is misclassification due to inaccurate detection at the endocardial border as opposed to that of the epicardial border?", which might denote a particular weakness of the algorithm, or the need for a more consistent segmentation protocol with respect to how papillary muscles are treated / defined image-wise and dealt with, etc.

The distance transform of a binary segmentation mask gives a value to each pixel in the image domain Ω, equal to its 'distance' from the nearest pixel belonging to the segmentation; the 'distance' itself is measured via a suitable *distance metric* — typically the Euclidean distance in terms of 'number of pixels', though other metrics exist and are commonly used, such as the 'Checkerboard distance' or the 'Manhattan distance', depending on the problem at hand.

Any such distance transform can be converted to a fuzzy mask by normalising the values in the [0,1] range in a suitable manner (e.g. by dividing by the highest 'distance value' in the resulting transform). This 'base' mask can then be transformed further via a suitable fuzzy membership function, e.g. to weigh parts of the distance transform more than others, isolate particular ranges, reverse the direction, etc. Fig. 6.10 demonstrates this concept over the 'base' transform obtained from the gold standard of the petal set (see fig. 5.1, p. 151).

6.2.4 Evaluation of object mass distribution as a function of distance from the gold standard

In the same way we evaluated fuzzy *direction* masks in a 360° arc (including the special case of a *directional profile*), which gave us information on how the 'mass'

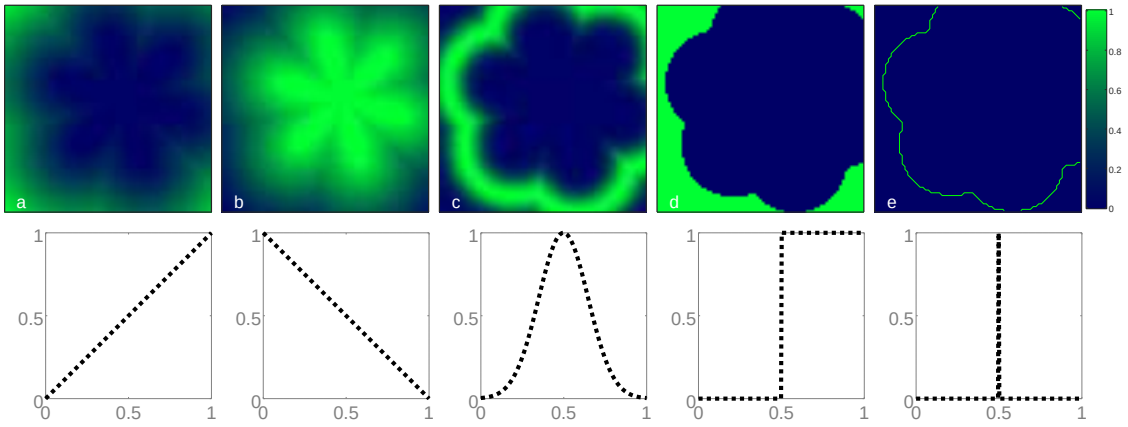


Figure 6.10: Fuzzy masks representing the notion of 'distance' with respect to an object.
a. The 'base' normalised distance transform obtained from a 'petal' mask, with no further transformation applied, representing the relation "distant"; values closer to '1' denote more "distant" pixels.
b. A 'closeness' mask, obtained as the fuzzy complement of the base mask. Values closer to '1' denote pixels that are "closer" to the gold standard.
c. A mask denoting the vague relation 'neither too close nor too distant', obtained by transforming the base mask with a Gaussian fuzzy membership function, centred here at 0.5.
d. A mask denoting the crisp relation 'points that are at a distance ℓ or more from the object', obtained by transforming the base mask with a step function (where ℓ here is set to 0.5 in normalised units).
e. A mask denoting the crisp relation 'points located *exactly* at a distance ℓ from the object', obtained by transforming the base mask with a *delta* function (where ℓ here is set to 0.5 in normalised units).

of a segmentation object or local-performance mask is distributed *around* the gold standard or any other point of interest in general, we can similarly evaluate how such mass is distributed as a function of the *distance* from the gold standard or other point of interest (where this is expressed via the parameterised range [0,1] as above).

For this purpose, fuzzy distance masks like the ones in fig. 6.10 can be 'applied' to a segmentation object or local-performance mask via fuzzy intersection, and the 'total mass' of the intersection gives the extent of the object that is consistent with the particular definition of distance expressed by that fuzzy distance mask. If the fuzzy membership function generating such masks is parameterised such that the resulting distance mask is expressed with respect to a particular distance of interest ℓ expressed in normalised units (i.e. taking values in the $[0, 1]$ range), we can then describe the object's mass distribution as a function of distance, by obtaining a

"total mass" measurement for all different values of ℓ.

As with directional masks, when dealing with 'standard' fuzzy masks having a well-defined cardinality, like the "Gaussian" (fig. 6.10c) and "step" (fig. 6.10d) fuzzy distance masks, the 'total mass' resulting from the application of a distance mask onto an object is more usefully expressed as a proportion with respect to the 'total mass' of the object as a whole (i.e. as defined over the entire image domain). When it comes to "masks" such as the one in fig. 6.10e, however, it is more useful to treat this as one of the many continuous building blocks required to describe the object's *distance profile*, in a sense analogous to the directional profile described earlier[5].

Distance Profiles

We won't go into the mathematical formalism involved in too much detail here like we did for the *directional profile*, since the concepts involved can be straightforwardly adapted here, but in general, the steps involved in creating a distance profile are:

- Interpret the fuzzy mask F whose mass distribution we are interested in (i.e. the segmentation candidate itself, or a suitable performance measure) as a *continuous* function $\mathbf{f}$ (e.g. via suitable interpolation over the image domain).
- Obtain a continuous function corresponding to a suitable 'base' distance transform, normalised in the range [0,1].
- For any *specific* distance value ℓ in the [0,1] range, the contour of that value over the 'base' transform above forms a closed curve.
- The *closed-curve line-integral* of $\mathbf{f}$ along that particular contour is a measure of the total mass m corresponding to that particular distance.

[5]The use of quotes here is to indicate that this is not in fact a 'standard' fuzzy mask, in the sense that it is not defined with respect to the discrete pixels comprising the image domain, but is a continuous closed contour instead, resulting from the application of the *delta* fuzzy membership function onto the distance transform, where this is interpreted as a continuous field *over* the image domain (that is, as opposed to *via* the image domain, i.e. the pixels comprising it). Since the resulting contour is a *curve* (i.e. it has no 'area' to speak of), the cardinality of such a "mask" with respect to the image domain, and by extension that of its intersections, is not well-defined, and the 'mass' encompassed by it cannot be measured meaningfully, except when considered as a differential component (i.e. in a larger integral operation). In practice, however, for operations involving such masks, this is approximated numerically by 'standard' fuzzy masks with a well-defined cardinality in terms of the image domain.

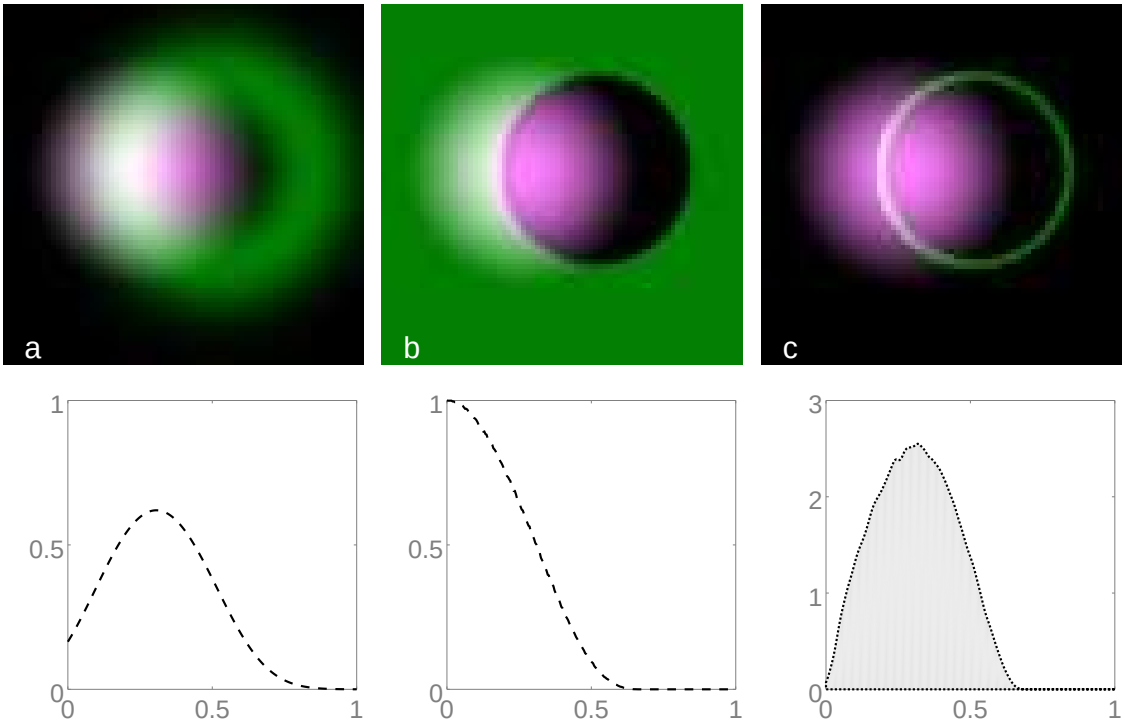


Figure 6.11: Evaluation of object mass distribution as a function of distance from the gold standard (represented here by its centroid), for three types of fuzzy distance mask.
a. Evaluation of the segmentation object using a 'Gaussian function'-based fuzzy distance mask.
b. Evaluation of the segmentation object using a 'step function'-based fuzzy distance mask.
c. Evaluation of the segmentation object's *distance profile*.
The top row for each type of mask is a fused RGB image; the segmentation object is shown in purple, and an instance of the fuzzy distance mask corresponding to $\ell = \frac{1}{3}$ is shown in green. The bottom rows show plots of the total mass m obtained as a function of the distance ℓ under evaluation. Note that 'a.' and 'b.' are normalised such that their values denote percentages with respect to the segmentation object's mass as a whole, whereas the distance profile in 'c.' is normalised such that the area under the curve sums up to '1'.

- A distance profile can be obtained by plotting m as a function of ℓ, normalised if appropriate, such that the area under this curve sums up to 1.
- A numerical approximation to such a continuous distance profile can be obtained via the sum of pixel values over a suitably interpolated path (whose width matches the increment in ℓ, e.g. one pixel wide for one-pixel increments), matching the contour in question as closely as possible.

Fig. 6.11 shows the evaluation of the "Gaussian" and "step" fuzzy distance masks

for increasing ℓ, as well as the *distance profile*, evaluated on the fuzzy discs set.

6.2.5 A fuzzy generalisation of the Hausdorff distance, using distance profiles

One particularly interesting application of the *distance profile* is that it offers a useful interpretation for the *Hausdorff distance* [110], which allows it to be further generalised to fuzzy objects in a straightforward manner, bypassing the need to define a 'border contour' for the objects involved — a concept which would be hard to define for fuzzy objects.

Given a *binary* segmentation mask and a *binary* gold standard, their Hausdorff distance is defined as the 'most distant point of failure', i.e. the largest distance from any point of the contour corresponding to the border of the gold standard, to that of the segmentation. The Hausdorff distance is therefore often used as a (non overlap-based) validation metric, where a Hausdorff distance of 'zero' indicates a 'perfect' segmentation, in that the contours of the two binary objects (and by extension the two objects themselves) coincide; as the Hausdorff distance increases, the quality of the segmentation is therefore deemed to decrease with respect to this criterion. In particular, two segmentation objects achieving the same Hausdorff distance with respect to a gold standard, are deemed to be of comparable quality in terms of this criterion, regardless of the respective areas / volumes covered by each object, or how much of each segmentation's contour can be found at, or close to, that maximal distance (we have mentioned in section 4.1.2 examples of when it may be preferable to use this validation strategy as opposed to an overlap-based metric).

If this 'most distant point of failure' is located *outside* the binary gold standard, then with respect to our formulation of the distance profile, the Hausdorff distance essentially becomes the largest such ℓ in the distance profile, for which m is *non-zero*. Conversely, if it is located *inside* the gold-standard, then the largest distance between the two contours can be captured in the same manner, by performing the same operation with the role of the two objects reversed (i.e. obtain the distance profile

of the gold standard with respect to the binary segmentation instead), to obtain a second, 'inward' candidate for the Hausdorff distance: the final, true Hausdorff distance is the largest of the two candidates, thus also ensuring the symmetry of the operation. For binary objects, the smallest such m possible (assuming it is obtained numerically via the sum of individual pixel values) is equal to '1' (i.e. in the case that only a single non-zero pixel is present at that distance).

We can easily generalise this to fuzzy objects, as the largest ℓ for which m is non-zero[6]. Alternatively, we could even define a *parameterised* version of a *fuzzy-generalised Hausdorff distance*, by introducing a degree of tolerance; that is by obtaining the largest ℓ value for which m is still larger than or equal to the tolerance threshold m_0 (e.g. '1'). Note, also, that the same information can be obtained using the 'step function'-based approach instead of a distance profile (e.g. compare the end-points between panels 'b' and 'c' in fig. 6.11).

Other generalisations of the Hausdorff distance exist, with various underlying semantics (an analysis of their strengths and weaknesses is beyond the scope of this section; for an overview and criticisms of such algorithms see [120]); however, we would argue that the distance profile approach is useful in being simple in its formulation and intuitive from a semantic point of view. Furthermore, it has the advantages that:

- it can be applied equally well to local-performance measures, as opposed to just the segmentation objects themselves
- it retains the specific semantics that one chooses to interpret these with (in terms of an appropriate choice of t-norms, etc).

In other words, as well as being able to answer the specific question "what is the farthest 'distance' between the (fuzzy) gold standard and segmentation candidate" —

[6] In practice, when implementing such an algorithm, for numerical reasons relating to floating-point arithmetic, it is safer to obtain the largest ℓ for which m is larger than a very small number close to zero.

essentially allowing the use of the Hausdorff distance as a validation metric for *fuzzy* objects — this interpretation enables one to also answer questions like: "what is the farthest away from this fuzzy gold standard where overlap / classification accuracy is below 0.5", or "the farthest away where false positives are present specifically", etc.

6.2.6 A note on the notion of distance from a fuzzy object

The astute reader will have noticed that the 'base' distance transforms used in the examples demonstrated so far, were obtained from either binary objects or point-sources rather than 'true' fuzzy objects. We deferred this discussion of what constitutes a meaningful distance transform for fuzzy masks until now, to more clearly demonstrate the concept of a fuzzy *distance mask* in more general terms first, and show how they could be used to evaluate segmentations. However, the transition from obtaining a base distance transform with respect to a binary object, to obtaining one from a fuzzy object is not a trivial or straightforward matter, since for fuzzy objects the notion of distance with respect to a crisply-defined object border is ill-defined.

Saha *et al.* defined a *Fuzzy Distance Transform* (FDT) [121] relying on the concepts of a 'fuzzy path length' with respect to a fuzzy object, and the related concept of a '*fuzzy distance*': for any given path π between two points A and B over the image domain $\Big($parameterised as $\pi(t)$ such that $\pi(0) = A$ and $\pi(1) = B\Big)$ the corresponding 'fuzzy path length' is defined (at least in the continuous sense) as the integral of an infinite number of infinitely small path segments of (canonical) length $\mathrm{d}\pi$, multiplied (i.e. weighted) by the corresponding fuzzy value corresponding to that segment, with respect to an underlying fuzzy object also defined over the image domain. Formally:

$$\text{fuzzy path length} = \int_0^1 \mu\Big(\pi(t)\Big) \left| \frac{\mathrm{d}\pi(t)}{\mathrm{d}t} \right| \mathrm{d}t$$

Fuzzy distance is then defined as the 'length' corresponding to the 'shortest' possible path from A to B (which may not necessarily be a straight line).

The *Fuzzy Distance Transform* of a *bounded* fuzzy object can now be defined as a mapping from each point *inside* the fuzzy object of interest, to a value equal to the fuzzy distance from that point to the *nearest* point located *outside* the object (i.e. the nearest point with a fuzzy value of exactly zero). In this sense, 'fuzzy distance' is essentially interpreted as a sort of optimal 'total traversal energy' (or 'total traversal time') required to "escape" the fuzzy object from any particular starting point position within it, given that the traversal between any two pixels in the set requires a certain amount of 'traversal energy' (or 'traversal time'), which is contingent on their fuzzy values as well as the physical distance between them (i.e. as if pixels with higher fuzzy membership values were generally 'denser' and harder to traverse, contributing more to the 'energy' or 'time' taken to traverse them).

If desired, the FDT can be parameterised further using a *tolerance* parameter τ, such that the FDT then maps each point to the fuzzy distance from that point to the nearest point with a fuzzy value less or equal to τ instead — the rationale here being that, with respect to the distance transform, and to the extent that they are of much lower value compared to the remaining 'core' of the object, points in the object with fuzzy membership values below such a tolerance threshold can be safely discounted as not representative of the object as a whole, and therefore not worthy of being included in such a distance calculation (or, alternatively, interpreted as though their contribution to the 'total traversal energy / time required for escape' is negligible and can therefore be trivially discounted).

We immediately face a few problems with interpreting the FDT by Saha *et al.* as a 'base' distance transform for use with the fuzzy objects we have been dealing with so far. The first is that, this FDT concerns itself with 'distances' from *inside* the object, as opposed to a notion of measuring the extent to which a point *outside* the object is found at a certain distance from it. Naturally, we could mitigate this by computing an FDT for the *complement* of the fuzzy object, such that the 'distance' value given to each point outside the object, reflects the 'energy / time' required to "escape the *background*" by sufficiently 'penetrating' the object's innermost 'crisp'

core. But this then poses a second interpretation problem: is it appropriate to treat the innermost 'crisp' core of the fuzzy object, that is, the set of pixels with a value of '1', as the 'escape / penetration boundary' for the FDT with respect to the background? Remember that for some of the fuzzy objects we're dealing with (e.g. see the retinal vessel sets), there may not even be such a 'crisp inner core' in which case such an 'inverse FDT' is undefined. Of course, this could be mitigated by applying a tolerance τ to artificially create slightly more permissive 'penetration boundaries', compatible with fuzzy objects whose maximal membership value is at most $1 - \tau$. But this seems unintuitive, and contrary to the function of such a *tolerance* threshold; does it make intuitive sense to talk about a 'distance' from an 'innermost core', expressed as 'energy', as opposed to a *physical* distance from an *outermost* core, perceived as being representative of the object's boundary?

Therefore, a more intuitive and straightforward way to define a notion of an external distance from a fuzzy object, in a manner that is compatible with the particular spatial semantics of fuzziness we have been discussing so far, may instead be to obtain a standard distance transform with respect to the 'outermost core' of the fuzzy object, as defined by crisply retaining pixels that *exceed* a certain low fuzzy-value *tolerance* threshold. In other words, pixels with fuzzy values falling below this tolerance threshold, are deemed to be too 'weak' and unrepresentative of the object as a whole (in the specific context of establishing a meaningful measure of distance from it), and therefore their contribution with respect to evaluating such a distance from the object as a whole can be discounted as trivial or as 'effectively noise', and the remainder of the fuzzy object is interpreted as a binary object with respect to obtaining a standard distance transform from its boundary. While this approach might strike one as not being 'purely fuzzy', in contrast to the themes presented in this thesis thus far, and the particular choice of tolerance threshold *can* generally affect the resulting distance transform, in this particular case it is a more meaningful and intuitive choice with regard to the spatial semantics in question, and the act of selecting a tolerance threshold is fairly intuitive within such a context of retaining 'only non-trivial volume', making the semantics of the resulting distance transform

easier to interpret. Naturally, if a more 'fuzzy' insight is required, then a *distribution* of base distance transforms can be obtained for a suitable range of tolerance values.

Conversely, it may be desirable to define a 'distance' (in the more general sense) involving 'attraction' semantics from the object as a whole — rather than in terms of a representative border or 'penetration boundary' — in a 'purely fuzzy' manner that does not rely on arbitrary tolerance thresholds at all. To this end, we also propose an alternative approach, which treats each point inside the fuzzy object as a point-source exerting a certain 'attraction force' over the domain, such that this force decays towards zero (according to a decay curve) as one moves further away from it, and whose strength, or 'influence' at the source is proportional to its fuzzy value. Therefore, for any point outside the object, we can calculate the influence it receives from each point belonging to the fuzzy object; the actual final influence received is the strongest influence received from all points inside the object. The 'distance' then becomes a (normalised) measure of the extent to which such a point is 'outside the object's sphere of influence', i.e. the complement of the resulting fuzzy 'influence' mask. If the decay curve chosen is linear (with '0' influence set to e.g. the maximal distance possible in the image domain, such that the normalised distance from 0 to 1 corresponds to a particular number of pixels in the image domain) then the resulting 'base' distance transform is mostly linearly-increasing over this domain. Note that in the presence of a binary object, this transform also converges to the standard distance transform. Also, similar to the previous methods, such a distance in the $[0, 1]$ range can be meaningfully transformed back to a physical distance in pixels, by applying the normalisation factor in reverse.

Fig. 6.12 demonstrates the three transforms in action, using the gold standard from the fuzzy discs set. Given the symmetry of the object, we only show the fuzzy value profile along the row that goes through the disc's centre. Note that, technically, as described above, the 'influence-based' approach, unlike the other two approaches, does not take extra parameters such as a 'tolerance threshold'. However, one can be artificially introduced if required, by artificially 'pushing' all

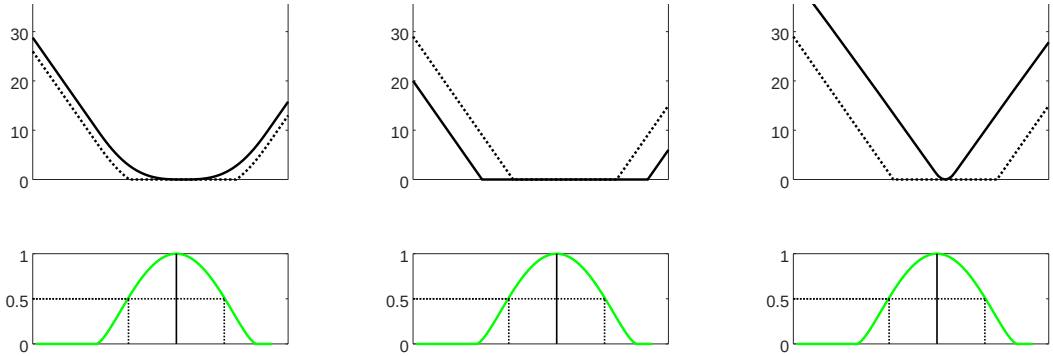


Figure 6.12: Base distance transform approaches for fuzzy objects. In all three cases, the bottom row shows the fuzzy membership values of the 'centre' row of the gold standard from the *fuzzy discs* set. The top row shows the Saha *et al.* Fuzzy Distance Transform for tolerances of 0 (solid line) and 0.5 (dotted line), the standard distance transform obtained after application of a tolerance threshold of 0 (solid line) and 0.5 (dotted line), and the 'distance as degree of immunity from object influence' approach respectively over the same domain (where the dotted line corresponds to the introduction of an 'artificial' 0.5 threshold — see text for details). Distance is shown 'de-normalised' here, i.e. reported in terms of 'pixel-lengths' resulting from each method. Note that when converting these base transforms to a fuzzy mask, the values would become normalised in the [0,1] range.

values above a certain threshold in the fuzzy object up to '1', so as to increase their influence (which is similar to how we apply the tolerance to the FDT). We have done so here for the 0.5 tolerance threshold, for a better visual comparison of its effect with respect to the other two methods.

However, the main point we are making with regard to obtaining a 'base' distance transform from a fuzzy object, is that in the end, as with most fuzzy problems, there are several ways by which to do so, and the particular choice of approach depends on the particular interpretation of 'distance' the experimenter might wish to ascribe with regard to the particular problem at hand.

6.2.7 A note on pre-applying spatial masks on segmentation and gold-standard masks directly

In section 6.1.2 we described how a *region*-mask could be applied at each pixel position of the segmentation and the gold-standard *separately*, such that the resulting mask pairs at each position could be validated against each other, resulting in a performance map providing a 'regional Tanimoto coefficient' measurement at

each pixel position.

It is possible to follow a similar logic when applying *spatial* masks. So far we have applied spatial masks directly onto *pre-derived* measures. However, in theory, the reverse strategy is also possible, i.e. we can apply a particular spatial map on the segmentation and gold-standard masks separately, and then derive the Tanimoto coefficient for that resulting pair; by extension, this approach could be used to obtain alternative 'directional profiles' or 'distance profiles', analogous to the ones discussed in sections 6.2.2 and 6.2.4. Figs 6.13 and 6.14 show the equivalent such plots to figs 6.9 (p. 192) and 6.11 (p. 198) respectively.

However, note that, while the two approaches seem similar, they answer two very different questions — as is implied from their resulting 'profile' plots, which we only demonstrate here for the sake of comparison and completeness. The approaches described previously in sections 6.2.2 and 6.2.4 relate to local measures which are distributed in a certain manner over the image domain, such that the application of any spatial map attempts to relate the distribution of the region concerned with that of the measure over the whole image domain. Whereas the latter approach shown here, isolates a particular spatially-connected subset of the image space over both sets, and then evaluates the similarity of the two subsets in isolation. In other words, the former approach asks questions such as "How many pixels are inaccurate / misclassified left of the gold standard (and also, how is that distributed within that region)?", or "How much of the total inaccuracy / misclassification that exists in the image domain, occurs within that specific region?" whereas the latter approach asks the different question "To what extent are these sub-regions (dis)similar in the two masks, when evaluated in isolation (and expressed as a single measure of (in)accuracy)?".

In particular, note that while such a question may be sensible when evaluating direction or distance with respect to a gold standard as represented by a *centroid*, it does not make sense when evaluating the direction or distance from a gold-standard

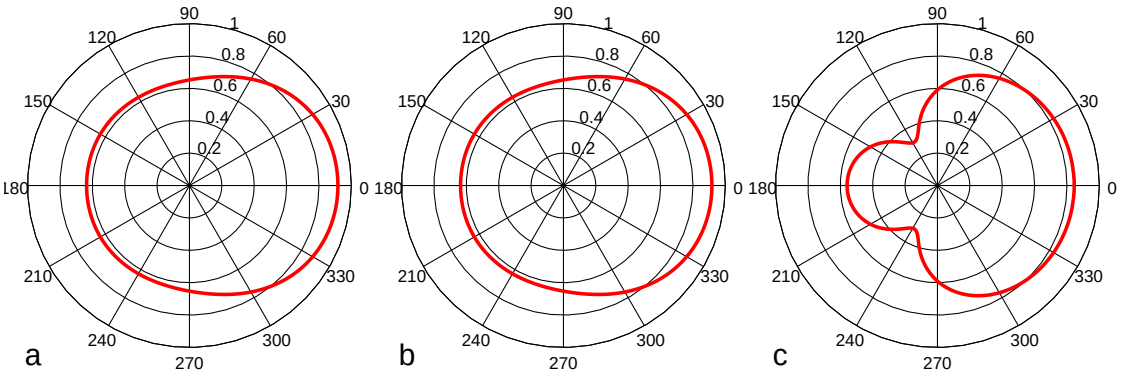


Figure 6.13: Evaluating directional inaccuracy as isolated regions. These plots are obtained in an equivalent manner to those of fig. 6.9 (p. 192), except the respective directional mask of each step (i.e. angle) is applied to the segmentation and gold standard masks *separately*, producing two sets representing the corresponding isolated regions from each mask. A Tanimoto coefficient is then calculated for each such pair; the values reported here as 'inaccuracy' (for the sake of more meaningful comparison to fig. 6.9) are simply the complement of the resulting Tanimoto coefficient (i.e. $1 - T_c$) at each step. Note that, contrary to fig. 6.9, plot 'c' is not 'area'-normalised in this case.

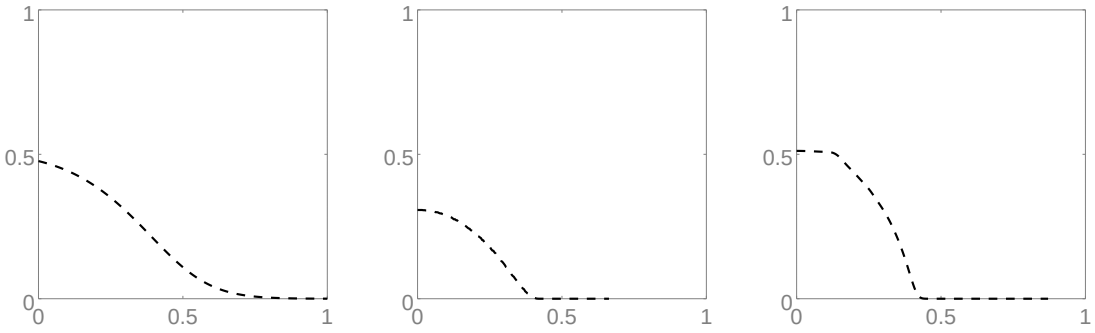


Figure 6.14: Evaluating accuracy for increasingly 'distant' isolated regions. These plots are obtained in an equivalent manner to those of fig. 6.11 (p. 198), except the respective distance mask of each step (i.e. distance increment) is applied to the segmentation and gold standard masks *separately*, as above. The plots show the resulting Tanimoto coefficient at each distance step (i.e. as a function of the distance ℓ under evaluation). Note that, as above, contrary to fig. 6.11, plot 'c' is not 'area'-normalised in this case.

as a *whole* object, since one would expect that the independent evaluation of a segmentation's particular region against the equivalent region in the gold-standard, where that region is by definition *outside* of the gold standard, to always lead to a Tanimoto coefficient of zero. This may be a useful question to ask, however, if one is evaluating the independent accuracy of sub-regions with respect to clinical landmarks, which may be unrelated to the gold-standard.

6.2.8 A practical demonstration

We demonstrate in this section an example of how the information provided by the directional and positional extent of misclassification can complement traditional single-valued validation when comparing the quality of two segmentation outputs against a gold standard.

Fig. 6.15 shows the output from two segmentation algorithms, which have been designed to have *exactly* the same number of true positive, false positive and false negative pixels, and therefore the same Tanimoto coefficient on validation, but which exhibit very different behaviour. Clearly the two outputs are very different, and one might be justified in remarking that, at a glance, the output on the left seems to be of '*better*' quality by comparison. While to say that the output on the left is '*better*' is subjective to an extent (since, after all, it has exactly the same misclassification rate as the output on the right by definition), it may well be 'better' in the sense that it is more useful to a researcher with respect to a particular criterion, such as compactness, or a more easily detectable / correctable defect at post-processing.

In this case, the output on the left can be seen to exhibit oversegmentation and undersegmentation respectively strongly in very specific directions, and at a very particular distance from the gold-standard. Conversely, the output on the right seems to be inaccurate in a more general since, spanning all directions and distances much more uniformly by comparison. Such information can be of great use to the researcher; an ensemble algorithm could make use of segmentation failures of the left kind, for instance, as they would be much more likely to result in a compact object

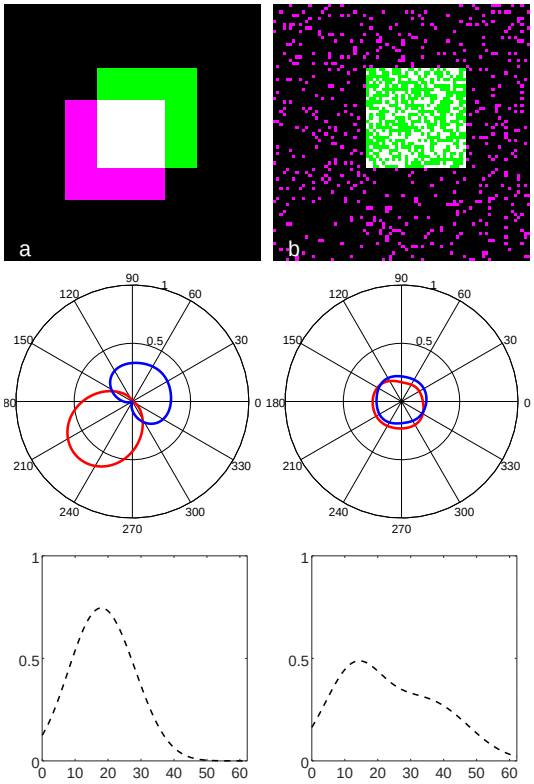


Figure 6.15: Synthetic set demonstrating an example application of directional and distance profiles on two algorithms (left versus right column) with identical Tanimoto coefficients but different failure mechanisms. **Top:** Fusion images of the two segmentation outputs against the same gold standard; True positives result in white colour, false negatives in green, and false positives in purple. **Middle:** Directional evaluation of oversegmentation (red) and undersegmentation (blue) with respect to the gold-standard centroid for each segmentation output, evaluated using a cosine directional mask, and reported as a percentage of the overall oversegmentation / undersegmentation respectively. **Bottom:** Evaluation of misclassification as a function of distance ℓ (in pixels) from the gold standard centroid; misclassification is reported as a percentage of overall misclassification in each case.

than the noisy version on the right. Therefore, parameterisations that are found to be more likely to result in such modes of failure can be selected preferentially in an automated manner, such as by choosing parameter sets, either *offline* (i.e. during a training phase) or *online* (i.e. if assessed with respect to some other known landmark rather than requiring a gold standard) that result in directional / distance profiles exceeding a certain threshold.

6.3 Evaluating failure caused by the presence of particular features

Sometimes, segmentation algorithms fail in predictable patterns, but the circumstances under which these patterns occur are not straightforward to predict and correct for in advance. For instance, we previously mentioned in sections 3.3.3 (p. 67) and 3.3.7 (p. 79) how in the Cocosco *et al.* algorithm, the presence of bright

fatty regions near the Right Ventricle, can sometimes cause them to be mistakenly identified as part of the ventricle; we have seen that the accuracy of atlas-based algorithms can suffer under conditions leading to sub-optimal registration (section 3.3.3, p. 71); and that model-based methods may fail to adequately match the shape of more elaborate borders depending on the model's tendency to favour smoother shapes, as a result of 'elastic' forces acting on the model by design (section 3.4.4 – fig. 3.9, p. 95).

So far we have discussed how failures relating to a particular direction or distance from an object or point of interest, may be quantified in an automated or semi-automated manner through the use of fuzzy spatial / anatomical relationship masks. Occasionally, however, segmentations fail through the process of exhibiting particular fuzzy 'features' in the segmentation, i.e. the presence of a particular shape, texture, or attribute (e.g. tissue density, compactness, smoothness, etc), which is present in parts of the image, and which are an artefact of the particular algorithm used, and not present in the gold standard. The automated retinal segmentation algorithm used in [23] (see section 5.2.1), is a good example of this, as shown in fig. 6.16, where on certain occasions, artifacts of a specific nature can be detected in the segmentation output.

We introduce here the concept of a '*validation sweep*' (or '*validation map*', in terms of the *output* of such a 'sweep' operation), and show how this can be used to detect segmentation failure due to the presence of such unwanted fuzzy features, given a segmentation / gold-standard pair.

We note that this concept is still somewhat under development and presented here mostly for completion in the context of the previous concepts presented in this chapter for evaluating modes of failure.

6.3.1 Validation sweeps

There is a certain family of 'sweeping' operators / functions $\mathcal{S}(y; f, k, g)$, where:

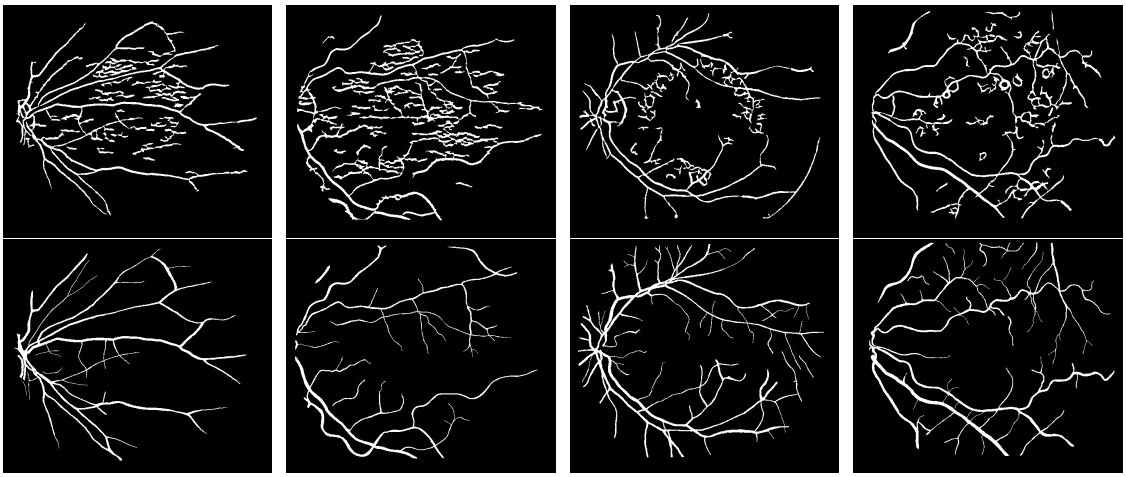


Figure 6.16: Examples of segmentations failing due to the presence of specific features, taken from the STARE dataset of retinal segmentations [23] (see section 5.2.1). The top row shows results obtained using the automated algorithm described in [23], and the bottom row corresponds to a gold standard. The first two segmentations can be seen to display artifacts in the form of horizontal lines, and the last two display can be seen to display artifacts in the form of circular loops

- f is any function $f : x \mapsto \mathbb{R}^N$
- k is a 'kernel' function $k : x \mapsto \mathbb{R}^N$
- y represents an 'offset' (independent) variable
- g is a higher-order 'evaluation' function $g : f, k; \tilde{f}, \tilde{k}, y \mapsto \mathbb{R}^N$, where $\tilde{f}[f]$ is a higher order function modifying $f(x)$, and $\tilde{k}[k, y]$ is a higher order function modifying $k(x)$, in a manner that depends on 'offset' y – i.e., $\tilde{k}$ results in a 'shifted' (or more generally, 'adjusted') version of k by y (plus any other modifications, e.g. reflection).

In other words, by varying the offset y, the kernel k is made to 'sweep' over the domain shared with f, such that for each position y in the sweep, a particular function g evaluates the application of k w.r.t. f, the result of which becomes the output of $\mathcal{S}$ at that position y.

A prominent example of such an operator is the *convolution* of a function $f(x)$ with a kernel $k(x)$, where function g takes the specific form of an *integral transform*:

- $\tilde{f}[f(x)] = f(x)$

- $\tilde{k}[k(x), y] = k(y - x)$ (i.e. k is reflected and offset by y)
- $g(f, k; \tilde{f}, \tilde{k}, y) = \int_{x \in \mathbb{R}^N} \tilde{f}[f(x)] \;\; \tilde{k}[k(x), y] \;\; \mathrm{d}x$

By contrast, an example which follows the same form, but where the kernel 'adjustment' does *not* represent an 'offset' in the 'space sweeping' sense but is a 'multiplier' instead, is the *Fourier* transform (or *Fourier series* to be more exact), where:

- $\tilde{f}[f(x)] = f(x)$
- $k(x) = e^{-2\pi i x}$
- $\tilde{k}[k(x), y] = k(xy)$
- $g(f, k; \tilde{f}, \tilde{k}, y) = \int_{x \in \mathbb{R}^N} \tilde{f}[f(x)] \;\; \tilde{k}[k(x), y] \;\; \mathrm{d}x$ (i.e. same as in convolution)

Examples of other 'sweeping' functions are morphological operations like dilation and erosion, where g takes a different form involving set operations (i.e. it's not in the form of an integral transform); yet another is in image registration, where g is of the form of a similarity function plus a regulariser, etc.

We might refer to such functions taking the form $\mathcal{S}(y; f, k, g)$ more generally as *kernel-based transforms*. However, we are specifically interested here in the first type of such transforms, where the 'offset' variable y represents a *physical* offset (as opposed to a multiplier or other type of 'adjustment'), such that the effect of calculating $\mathcal{S}$ at each offset y in the context of images, is that of a kernel undergoing a '*sweep*' over the entire image domain, and resulting in a particular output for each position in the 'sweep'.

We define a *validation sweep*, as a special transform $\mathcal{S}(y; f, k, g)$ of this kind, whose output is a valid fuzzy 'map', where:

- f represents a suitable fuzzy mask $\mathbf{f}$ (or more generally a fuzzy function $f \mapsto [0, 1]$).
- k represents a fuzzy mask $\mathbf{k}$, or more generally a fuzzy function with local support, representing a particular fuzzy attribute or criterion, which can be

evaluated locally w.r.t. f to judge local similarity or fuzzy membership w.r.t. to that particular criterion.

- y represents an 'offset' variable taking *discrete* values, allowing k to be evaluated at all possible overlap positions with respect to $\mathbf{f}$
- g is an 'evaluation' function outputting a fuzzy value, corresponding to the evaluation of k w.r.t. f, localised at point y; e.g. where both k and f are simple fuzzy masks, then any suitable fuzzy validation operator is a valid g function.

When dealing with fuzzy masks explicitly, we refer to the output of such an operation as 'the *validation map* of mask $\mathbf{f}$ with respect to a kernel-mask $\mathbf{k}$'. In the context of a particular validation operator, we denote this operation more succinctly as $\mathcal{S}_{\{\mathbf{f}|\mathbf{k}\}}$. Note that this operation is *not* symmetric $\left(\text{i.e. } \mathcal{S}_{\{\mathbf{f}|\mathbf{k}\}} \not\equiv \mathcal{S}_{\{\mathbf{k}|\mathbf{f}\}}\right)$. In the special case that the mask $\mathbf{f}$ and kernel $\mathbf{k}$ represent the *same* object, we refer to this operation as the *autovalidation sweep* of $\mathbf{f}$ (and the corresponding output as its *autovalidation map*), denoted simply as $\mathcal{S}_{\{\mathbf{f}\}}$. Fig. 6.17 demonstrates how a validation map may be calculated in practice for two fuzzy masks (defined over a one-dimensional domain).

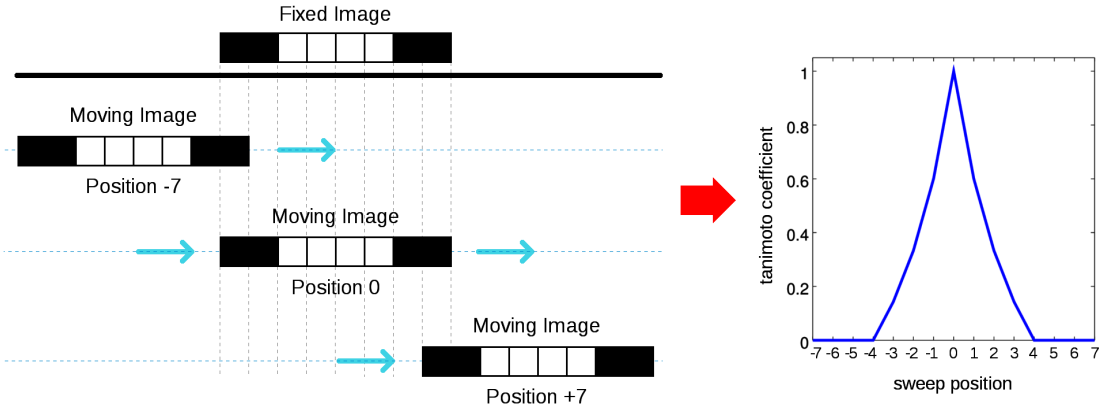


Figure 6.17: Obtaining a one-dimensional validation map: A suitable mask $\mathbf{f}$ (such as the segmentation mask or gold-standard mask we intend to assess) acts as the 'fixed' image, and a separate 'kernel' mask $\mathbf{k}$ (or in the case of *autovalidation* as is shown here, the mask $\mathbf{f}$ itself) acts as a 'moving' image. The 'moving' mask then undergoes a *sweep*, i.e. it is superimposed at each available position, such that all overlap positions are explored. A validation operator (such as the Tanimoto coefficient) is calculated for each position in the sweep; 'padding' pixels in this scenario are assumed to have a mask value of zero.

The important point here, is that a validation operator results in values in the $[0, 1]$ range, and therefore the resulting *validation map* is generally expected to be a valid fuzzy mask[7]. This means we can combine validation maps further with other 'fuzzy relationship' maps (such as the ones explored in the previous sections), or apply them onto segmentations or local performance maps, using standard fuzzy operations. Note that, in theory, when dealing with masks explicitly, any valid (summary) performance metric, such as the mean pT_c, or mean misclassification etc, can also be used to produce such a map; for the sake of generality and consistency, we use the term 'validation sweeps' and 'validation maps' here in the more general sense, to refer to any valid evaluation metric in general, that can evaluate a kernel k w.r.t. a function f for local similarity or membership. However, an exploration of the properties for a range of particular metrics with respect to validation sweeping is beyond the scope of this section, and we will be focusing more in the context of the overlap measures discussed above (and membership functions more generally).

6.3.2 Validation sweeps for quantifying failure caused by the presence of particular features

For appropriate kernels representing 'fuzzy features' or 'attributes', validation sweeps can essentially be used as 'fuzzy feature / attribute quantifiers', in the sense that they quantify the extent to which the 'accuracy' or 'presence' of a 'feature' or 'attribute', as detected by a validation operator (or other evaluation function), is high in the different areas of a fuzzy mask.

The process can be compared to the process of extraction of a 'feature map' from an image using *convolution*, where a suitable 'feature detector' (i.e. a kernel) undergoes a "sweep" over each position in the image, thereby quantifying the presence of that 'feature' for the area around each pixel. While the underlying operations

[7] At least in the case of overlap-based operators; in theory a validation map based on a generalised Hausdorff distance is also possible; however, if the resulting map is required to be a valid fuzzy mask, so as to be evaluated further using fuzzy operations, then the resulting distances need to be presented in normalised form (e.g. normalised by the largest distance possible within the image domain).

and intent of the two processes is different, a comparison between the two is useful for better understanding the nature of validation sweeps, therefore we make the comparison between these two kernel-based transforms explicit below. We distinguish between validation sweeps and convolution, in terms of the underlying operations involved, in terms of the nature of the kernels involved, in terms of the nature of the outputs, and lastly in terms of intent:

Nature of underlying operations In convolution, the 'evaluation' operation g involved at each position in the sweep is $g(y) = \int_{-\infty}^{+\infty} f(x)k(y-x)\mathrm{d}x$, i.e. "multiply k by f, then integrate the result". This is effectively a measure of *correlation* between the kernel k and underlying function f at that particular position in the sweep, and therefore the output of such an operation is a map, expressed in the same domain as f which highlights areas of high and low correlation with respect to the particular 'feature' k chosen. The 'multiply then integrate' operation is not a 'fuzzy' one, since its output is not restricted to the $[0, 1]$ interval, although a sigmoid function can be used further to achieve such a transformation, if required, which would then turn this into a validation operator of sorts. By contrast, the operation performed on k and f in a validation sweep is not restricted to the above, but can be *any* compatible validation operator, or fuzzy membership function in general, evaluating k w.r.t. f according to the particular semantics we're interested in. Naturally, the 'multiply $\rightarrow$ integrate $\rightarrow$ transform via sigmoid' operation *could* still be used, if one is really interested in a fuzzy 'correlation' map, but one could also use, e.g. a Gödel-based operator, expressing an 'optimistic fit', or a Łukasiewicz-based validation operator expressing a 'pessimistic fit', or more generally a fuzzy function evaluating the 'membership' of a criterion k w.r.t. f at each locality.

Nature of outputs In a similar manner to how *numeric* outputs from a number of convolution operations could participate in further *arithmetic* computations (such as averaging all the components), in the case of validation sweeps, because the outcomes are *fuzzy*, these can be used as components in the

context of larger (fuzzy) *logic formulas*; that is, a feature may be defined as the fuzzy conjunction, disjunction, or positive symmetric difference, etc of two or more sweep components (e.g. we could express a compound feature as something that demonstrates subfeature 'A' `and` subfeature 'B', but `not` 'C', `and` overall has higher fuzzy membership than 'D', etc).

Nature of kernel functions In convolution, there is no requirement for k to be 'fuzzy' (if anything, the more useful kernels also contain negative values to capture inverse correlation as well). However, a kernel k is generally required to be a function independent of f, and definable as a function, in the same domain as f (albeit typically with local support over that domain). In validation sweeps, this is an unnecessary restriction. While k can be a fuzzy mask, in which case the 'evaluation' function g may be as simple as a validation operation w.r.t. f in the context of a particular locality, it could also represent a more 'fuzzy' criterion, such as the vague statement "there exists tissue that is 'between 2-3mm' `and` is 'compact"', etc. In the latter case, an evaluation function g will simply be a mathematically meaningful membership function, evaluating the membership of f at that locality, with respect to the fuzzy criterion in k.

Note that this kind of effect could potentially also be simulated in traditional convolution, by using a 'dictionary', of predefined kernels corresponding to compatible 'image features' (i.e. a large collection of shapes or morphologies, which were chosen or generated explicitly as being representative of that criterion), and then combining all the outputs into a single map representative of the presence of any of the image features in the dictionary, However, a more general approach as described above is likely to be more intuitive, and potentially computationally simpler, by allowing one to specify a simple mathematical representation of the criterion represented by k (as we did for clinical variables in chapter 3, and with anatomical relationships in the earlier sections of this chapter), as well as by allowing any 'logical formula' steps to

be performed at the level of the evaluation function g, rather than having to be performed at the end from a collection of sweep outputs.

Intent of operation Convolution, particularly in the context of detecting 'image features' from 'real' images (i.e. as opposed to 'label' images), for the purposes of classification or regression, typically tries to identify instances, or a weighting among a collection of features, that minimize classification or regression error. Often this also involves transforming the resulting feature maps further, to isolate / localise interesting features more clearly. In the context of classifiers like Convolutional Neural Networks, the intent may even be to 'learn' such a collection of suitable features via backpropagation.

With validation sweeps, the intent is not one of detection and localisation per se, but one of quantification throughout the label image domain. We desire to quantify the extent to which a segmentation fails due to a particular mode of failure, relating to the presence of a fuzzy 'feature' or attribute in the segmentation map itself, or in a local performance / misclassification map w.r.t. a gold-standard. We are less interested in 'learning' features; indeed one's intent would be the reverse: to explicitly express such features as mathematical membership functions deemed to be representative of particular attributes (which may or may not take the form of a fuzzy mask combined with a suitable fuzzy validation operator), and then evaluate their membership throughout the domain.

Fig. 6.18 shows an example of such an application in action, using the first pair illustrated in fig. 6.16. The segmentation candidate $\mathbf{s}$ in question shows multiple false positives w.r.t. gold-standard $\mathbf{g}$, resulting from the presence of multiple 'roughly horizontal lines'. A suitable kernel (denoted here as $\mathbf{f}$) is created to reflect such a 'roughly horizontal lines' feature — in this case, via the union of a 'left' and a 'right' cosine-based directional fuzzy mask, preferentially capturing features to the extent they conform to a more horizontal orientation, and ignoring features to the extent they conform to a more vertical orientation.

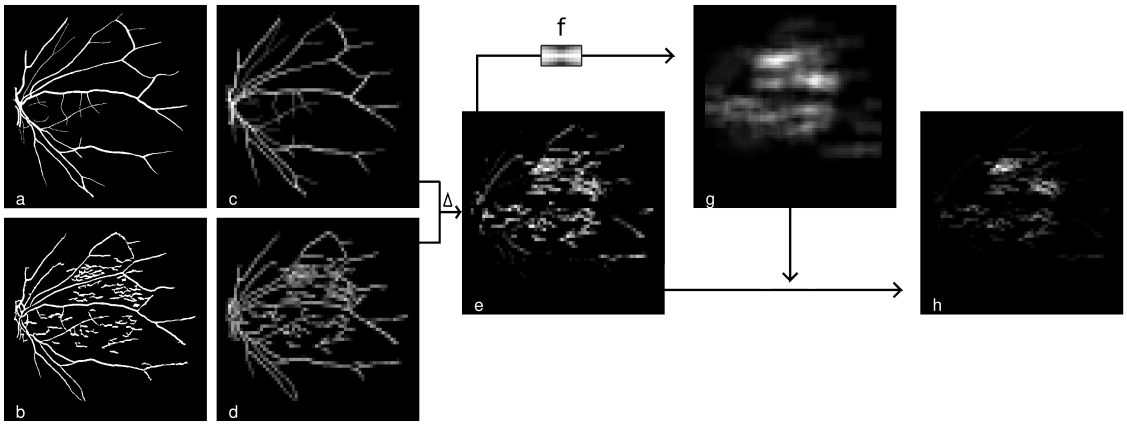


Figure 6.18: Quantifying failures consistent with the presence of a particular feature.
a. Gold standard retinal mask $\mathbf{g}_0$ at 700×605 resolution.
b. Automated segmentation mask $\mathbf{s}_0$. Note areas of false positive 'horizontal line' features.
c-d. Fuzzy masks $\mathbf{g}$ and $\mathbf{s}$ (at 70×60 resolution) derived from $\mathbf{g}_0$ and $\mathbf{s}_0$ respectively via 10×10 block-averaging.
e. The subset of the standard Gödel symmetric difference $\mathbf{g} \veebar_G \mathbf{s}$, for which $\mathbf{s} > \mathbf{g}$. (Note: intensities scaled to [0,1] range for better visual contrast)
g. Validation map of the mask in 'e' produced using the kernel $\mathbf{f}$ (shown here between the arrows) representing the fuzzy attribute 'roughly horizontal lines', and using generalised d-Norm validation. (scaled for better contrast)
h. The result of applying the 'heat map' from 'g' onto 'e', performed using product semantics, giving a fuzzy map of the extent of misclassification affected by that feature (scaled for better contrast). Based on this result (before contrast scaling), we can estimate that about 14% of the misclassification seen in 'e' occurs directly as a result of the presence of fuzzy feature 'f'.

As we saw previously, the positive symmetric difference of $\mathbf{s}$ and $\mathbf{g}$, corresponds to misclassification due to oversegmentation. Performing a validation sweep over this 'oversegmentation' mask results in a fuzzy 'heat' map, relating the extent of the presence of that fuzzy feature / attribute in the oversegmentation. In the previous sections, we applied 'fuzzy anatomical relationship' maps to local performance masks, thereby isolating the fuzzy portion of the latter affected by the relationship represented by the former. In a similar fashion, now that we have a 'fuzzy attribute presence' map, we can apply this to the original oversegmentation mask, to gauge the extent to which this is affected by that attribute. This gives a mask containing about 14% of the original oversegmentation mask in terms of fuzzy 'mass'. In other words, we have quantified the proportion of oversegmentation occurring specifically due to the presence of 'roughly horizontal lines' to be 14%.

Note that the above kernel captures features that are 'roughly horizontal' in their orientation, but it does not guarantee "straight" features (i.e. a 'mostly horizontal' but also rather curved feature would still be detected). As per our previous discussion, an alternative approach to ensure 'straightness' while still allowing some degree of freedom in the actual orientation relative to the horizontal, would be to either use a number of more crisply defined straight-line feature-detectors / kernel masks (e.g. $\{\mathbf{f}_1, \ldots, \mathbf{f}_K\}$) to obtain several 'heat map' subcomponents as above, and obtain the final 'heat map' as the fuzzy union of all the subcomponents (i.e. $\bigcup_{k=1}^{K} \left[\mathcal{S}_{\{\mathbf{s} \trianglelefteq \mathbf{g} | \mathbf{f}_k\}} \right]$), *or* to use a bespoke evaluation function directly, itself designed to assign a membership value in the range $[0, 1]$ at each pixel, reflecting the extent to which there exists a straight line in that locality, effectively corresponding to the scenario that a kernel is the fuzzy predicate 'is straight line', obviating the need to define an explicit kernel *mask* for validation as such.

Furthermore, as always, such a heat map is still a valid fuzzy mask, and could therefore by analysed further still, using any of the operations described earlier, e.g. combining with a *directional profile* or a *distance profile* to establish the direction or distance in which such failures tend to occur most with respect to, say, the optic disc.

6.4 Conclusion

In this chapter, we have demonstrated how one might complement traditional validation with further information characterising a segmentation's quality, by investigating the ways in which it fails; this can be done by 'asking' appropriate questions relating to particular modes of failure, where such questions can be mathematically represented using corresponding spatially-relevant fuzzy masks. We have shown how various useful metrics can be applied for this purpose to act as local-performance maps, and subsequently 'queried' by applying spatial masks relating to direction or distance from a useful point of reference, such as the gold

standard itself or a known landmark, or used to detect segmentation failures due to the presence of particular fuzzy features.

We do not intend to cover all the possible combinations of local-performance measures for all possible fuzzy relationships here, as we are more interested in showing the general *approach* behind the assessment of such "explicit" modes of failure, by asking appropriate questions, represented as suitable fuzzy relationship masks. For instance, the above concepts could easily be adapted to create "rotation profiles" by evaluating performance metrics for particular rotations of the gold standard mask from a suitable centre of rotation.

However, it goes without saying that *any* of the aforementioned local-performance measures could be used, with a variety of anatomical relations suitably represented as fuzzy membership masks, combined or in isolation, to provide answers to explicit questions like "in which direction does the algorithm oversegment", or "to what extent does misclassification occur far from the object", or any question that can be represented as a fuzzy relationship in general.

It is worth restating, however, that what we have provided here is not simply "a collection of tools and methods" from which one might randomly select combinations to generate a 'pool' of validation metrics, but rather a methodology for asking relevant, *explicit* questions represented as appropriate fuzzy masks, that might help the researcher to characterise particular aspects of the segmentation, or to query and discover particular modes of segmentation failure with respect to a gold standard. Therefore — with the exception perhaps of the directional and distance profiles, which may serve as a more general, 'exploratory' overview of how a segmentation or performance measure varies over the image domain — there is no single 'recommendation' for a combination of performance measure, spatial mask, and t-norm, that is a 'more appropriate' combination in the general sense. Rather, any particular choice reflects a particular question asked; at best if a 'general recommendation' is needed, this only makes sense in the sense of suggesting

a set of standard 'questions' (like the ones above), which might then serve as a useful general starting point for the researcher.

Summary

- Validation typically answers the question "to what *extent* is a segmentation accurate?", but fails to address the *modes* in which it succeeds or fails. This can be investigated with the use of appropriate fuzzy maps / masks.
- Local-performance maps can be used to visualise how segmentation performance is *distributed* over the image domain.
- We propose a number of such fuzzy maps: a Pixelwise Tanimoto coefficient map (as well as a 'regional' counterpart), Symmetric Difference (representing *misclassification*), Over- / Under-segmentation, and False Positive / Negative maps.
- Fuzzy masks can be used to represent a direction or distance from an object or point of interest. These can then be applied to local-performance maps allowing one to investigate *how*, or *to what extent* performance varies in a particular direction or at a particular distance.
- We have proposed and demonstrated specific examples of *directional masks* using a *cosine*, a *linear separation boundary*, and a *line segment* approach, and used them to evaluate the distribution of an object's total 'mass' *around* a point or object of interest.
- We have proposed and demonstrated specific examples of *distance masks* defined via the application of a *Gaussian membership function*, a *step membership function*, and a *delta membership function* over a 'base' distance transform, and used them to evaluate the distribution of an object's total 'mass' *as a function of distance*.
- A 'fuzzy' generalisation of the Hausdorff distance can be straightforwardly defined for fuzzy objects using distance masks and appropriately defined 'fuzzy' distance transforms.
- The suitability of the 'Fuzzy Distance Transform' algorithm by Saha *et al.* is discussed in the context of obtaining distance transforms from the *outside* of fuzzy objects, and two alternative approaches are proposed, each addressing different fuzzy 'semantics'.
- We propose a new operation, called a *validation sweep*, and show it can be used to determine the extent to which segmentation failure is due to the presence of particular, unwanted features.

7

Conclusion and future outlook

This chapter reviews and summarizes the main discussion points, claims, and contributions presented in this thesis, followed by a discussion of open questions, both flagged and generated by this work, and potential future work and further areas of application of the ideas presented herein.

Contents

7.1 Summary of contributions

The present thesis focused on the particular properties and applications of fuzzy and probabilistic segmentations, and in particular the semantics of such fuzziness in the context of medical imaging (with a particular focus on cardiac magnetic resonance imaging as the illustrative case); the use of fuzziness as measures of evidence, whose uncertainty can be quantified further; and the appropriate validation of

such segmentations.

7.1.1 Introduction, motivation and background theory

In chapters 1 and 2, we presented the motivation for this thesis and a summary of its claims (section 1.2), an overview of medical image segmentation with a particular focus on cardiac magnetic resonance imaging as the illustrative case (sections 2.1 and 2.2), and provided a comprehensive mathematical background (section 2.4) on the applicability of fuzzy set theory, fuzzy logic and fuzzy measures (including a short introduction to probability theory), as a suitable framework for investigating the properties of non-deterministic segmentations, and the semantics that they represent, in terms of ambiguity and clinical uncertainty.

7.1.2 Fuzziness, probability, and uncertainty over soft segmentations

Chapter 3 discussed the semantics of fuzziness in the context of medical image segmentation, and explored the concept of treating soft segmentations not as simple 'uncertainty maps over labels', but rather more usefully as numerical measures demonstrating quantifiable uncertainty of various forms themselves, that can be put to further use.

We explored this concept further in section 3.3 were we demonstrated fusion of soft segmentations based on pixelwise measures of uncertainty: a measure of algorithmic inconsistency / soundness, and a measure based on entropy, as a fuzzy generalization of Saha and Udupa's 'intensity-based class-uncertainty' [20]. This approach seems to work better than naive fusion of the soft segmentations by themselves, but did not necessarily exceed the accuracy of one of the original soft components. We theorized that clinical measures of uncertainty might work better in this situation, and therefore went on to define one such measure in section 3.4, and show its ability to create better soft segmentations, both in terms of converting a deterministic algorithm to a representative soft one such that it reliably retains its clinical estimates, but also in terms of allowing a clinician to guide

a segmentation algorithm towards optimal outputs for their particular clinical setup, without needing knowledge of the intrinsic algorithmic parameters for fine-tuning, but by providing simple clinical information to be used as physiological constraints instead (published as [21]).

7.1.3 Appropriate validation for soft segmentations

In chapter 4 we discussed the nature and particular challenges inherent to validation in the context of medical imaging (section 4.1), and noted that with respect to the expansive field of medical image segmentation, validation is a much under-researched field by comparison, ironically lacking the degree of rigorous assessment for validation algorithms that tends to be exhibited when 'validating' segmentation algorithms.

We identified as the main conventional approach used in the literature for the validation of soft segmentations to be that of thresholding (section 4.2.2), and relayed the historical reasons behind this, and the inherent dangers and unreliability of this approach.

Building on ideas from fuzzy literature, we produced a theoretical framework which places core concepts of fuzzy set theory and fuzzy logic at the heart of the semantic significance of the non-deterministic nature of segmentation mask pixels (section 4.3), particularly in the context of the interpretation of such pixels as boundary pixels exhibiting Partial Volume Effect (section 4.5). We showed that an understanding of this framework reveals particular properties and constraints for fuzzy validation (section 4.4), which are generally not respected, let alone exploited by existing methods for more accurate and precise validation.

In chapter 5, we proposed two alternative methods for validation (published as [22]), based on the novel concept of directional t-norms, as an illustrative case of the many direct and practical potential applications stemming from the aforementioned theoretical framework.

We confirmed these two novel implementations to be of higher accuracy, precision, and therefore reliability compared to extant state-of-the-art validation methods, and in particular the established but unreliable conventional approach of thresholding.

We concluded chapter 5 with a plea to the medical imaging community, to reconsider the use of thresholding as an academic gold-standard for the evaluation of fuzzy and probabilistic segmentations and the algorithms producing them, in favour of more appropriate validation methods designed specifically with the assessment of soft segmentations in mind.

7.1.4 Characterisation of segmentation failure modes for more informative validation

In chapter 6 we sought to provide methods to qualify and quantify the modes in which segmentations succeed or fail with respect to the gold standard. To this end, we proposed the use of *local-performance maps* (section 6.1), which demonstrate the distribution of suitably localised evaluation metrics over the entire image domain, and provided examples of such maps, each dealing with a particular question to be answered.

We then demonstrated several examples of fuzzy masks acting as representations of particular spatial / anatomical relationships with respect to the gold standard or other object of interest (section 6.2), and how these could be used to query segmentation performance with respect to explicit spatial constraints corresponding to clinical questions, or more generally in providing a *directional* or *distance* profile for the segmentation. While the concept of a fuzzy mask acting as a representation of a spatial relationship over an image domain isn't new, its use in the context of characterising validation, particularly combined with performance maps, is a novel approach.

Furthermore we examined how the concept of distance from an object applies to fuzzy objects, and proposed ways to assess this, comparing with current standards in the literature as applied to this particular context (section 6.2.6). Using such

a distance and the concept of distance profiles, we also proposed a 'fuzzy' version of the Hausdorff distance, which is a measure of the maximum 'spatial' distance between two fuzzy objects (section 6.2.5).

Finally we introduced the concept of a *validation sweep* (section 6.3.1), and demonstrated how this can be used to identify segmentation failures stemming from the unwanted presence of specific features in a segmentation (section 6.3.2).

7.2 Open questions and future outlook

Much of the work and contributions in this thesis were of a largely theoretical nature, with practical applications demonstrated mostly as 'proof of concept' work, rather than via extensive testing on large datasets. This was both as a result of the theoretical focus and nature of the work, as well as due to time constraints and a large number of ideas and avenues followed that have not made it into the final thesis. Therefore to some extent, the open questions still remaining relate to more extensive testing and identification of edge cases, more specific applications relating to the concepts already presented here, and linking together some complementary, but at the moment isolated contributions, which have remained somewhat patchy due to the existing time constraints, to form a more integrated, well-knit overall framework of uncertainty in soft segmentation and validation, such that the work would be more suitable for further publication. Beyond that, the future outlook for this thesis involves expanding application of the concepts discussed within, to other fields, or the reverse, i.e. improve the applications currently proposed, using concepts from other fields.

Combining segmentations using uncertainty

The proof-of-concept experiments presented in this thesis are promising, demonstrating the potential for improvement from using this method over simple consensus methods, in that the use of external information in the form of pixelwise uncertainty,

is shown to lead to better segmentation outcomes, for quality measures of uncertainty. However, they have also demonstrated some of the weaknesses, i.e. such measures need to be chosen carefully, and considered in terms of what they represent, e.g. whether they address systematic biases in the algorithms adequately. Therefore, a more extensive exploration of different types of uncertainty would be desirable for a journal publication, as well as testing these in larger datasets; since this method is not heart-segmentation specific, these could also include datasets from other disciplines.

We mentioned that the most useful type of uncertainty to have over an algorithm, would be a measure demonstrating uncertainty in the clinical sense. We made use of one such measure to create fuzzy variants, and to fuse segmentations with a bias towards more clinically relevant fused results, and demonstrated that this works in another proof-of-concept experiment. While this is probably far from being an ideal measure of clinical uncertainty that could be applicable to all scenarios, time constraints prevented us from returning to confirm its value in the context of combining segmentations as above. Therefore further work would test this measure, as well as the 'spatial-anatomical uncertainty' mentioned in this thesis, to confirm and demonstrate this hypothesis more conclusively on further datasets.

Secondly, while the proof of concept experiments within have shown the superiority of fusion using external uncertainty information, over fusion that relies only on consensus between the soft segmentations themselves in the absence of such external information, the consensus methods used were simple. For any further serious publication on the matter, it would be required to compare this against 'more state of the art' consensus methods, such as the various variants of STAPLE mentioned during the literature review; unfortunately, although considered, this was not performed in this thesis, partly due to technical problems with the implementation of the STAPLE algorithm in our particular setup, and partly due to existing time constraints rendering their effective resolution too costly in this context, so the decision was made to defer such comparison as future work.

Guiding segmentations using clinical knowledge expressed as physiological constraints

As mentioned at the conclusion of section 3.4, further work could investigate the use of further clinical constraints in the construction of a measure of clinical uncertainty, particularly relating to constraints of an anatomical nature, similar to what a radiologist might describe, and expressed in terms of fuzzy spatial / anatomical uncertainty maps, as per section 3.2.4 (figs. 3.1 and 3.2, p. 64).

Also, as above, the experiments were useful to demonstrate the value of this approach as a proof of concept, but larger, more variable datasets (from across disciplines), and testing on a larger variety of algorithmic substrates would be desirable, in order to pursue a journal publication.

Furthermore, as already hinted to in the introduction to this section, in many ways this application and the previous one of combining soft segmentations via pixelwise uncertainties, are complementary. The previous method used uncertainty on soft masks, this one has not only demonstrated a way to generate one useful such measure of a more 'clinically' relevant measure of uncertainty, but has also further provided a way to combine multiple 'weaker' uncertainties into a single 'stronger' one (reminiscent of boosting methods). The work on combining segmentations only used simple, single measures of uncertainty as proof of concept; combining the two approaches would allow us to use a collection of 'weaker' uncertainties in the same way, building a 'stronger' uncertainty map that combines uncertainty information coming from many different sources, and representing different types of uncertainty, to be used within the context of combining segmentations.

Finally, another useful application for this framework would be to use such an approach as a method of obtaining a consensus, relying on "clinical" criteria rather than a consensus relying purely on overlap. For example, in STAPLE, the consensus is provided in effect by weighing a single rater's output against all others, resulting in measures of sensitivity and specificity for each rater. This concept could be adapted to physical constraints, by weighing a component in terms of its clinical

estimates and anatomical relations, against the predictions provided from the rest of the candidates, acting as the baseline constraints. As a trivial example, an algorithm which results in smaller Ejection Fractions than its peers, or is "out of position" with respect to an anatomical landmark as compared to its peers, would be weighed less in the fusion outcome; and this could either be done at the segmentation level, or at the pixel level. So, in effect, this framework would still be a fusion using uncertainty, with the difference that the uncertainty here does not represent external information[1], but is calculated for each segmentation based on its deviation from the clinical attributes exhibited by its colleagues.

Fuzzy validation and directional norms

The theoretical ideas that led to the creation and use of directional norms, have been shown to be robust under experimental conditions, both through a synthetic and a clinical test. However, a lot of simplifying assumptions were made during the creation of this framework. It would be useful to put those assumptions to the test, and explore on a wider set of algorithms and datasets (including cardiac, but also from other disciplines, as reiterated above), not necessarily consistent with these assumptions, i.e. for more complicated models of PVE, for segmentations where softness does not necessarily straightforwardly denote tissue distribution, or for more models of boundary pixels rather than just the 'linearly separated' form.

Furthermore, the notions of 'boundariness' and 'neighbour-compatibility' mentioned in section 4.5 could be used to correct for weaknesses in obtaining boundary pixel orientations from the mask gradient, or to create smooth, higher-resolution masks from low-resolution fuzzy masks (as long as these denote or can be meaningfully transformed to reflect degree of quantization) by ensuring optimality for these criteria. Note that these criteria (neighbour-compatibility in particular) can also be used with relevant models of boundary pixels, such as the 'linearly separated' model used in this thesis, or more elaborate models, such as a smooth bezier-curve

[1] although, naturally, external uncertainty could be incorporated on top and weighed in, if necessary, in the manner described above

based model (e.g. the '2D vector' representation of fuzziness we proposed could be extended to higher dimensions, such as a 3rd dimension denoting 'curvature', for appropriately defined higher-dimensional t-norms and other fuzzy operators).

Finally, the notion of propagating the uncertainty over soft masks to their validation outcomes has been mentioned, but not adequately explored. A promising framework that was considered and that would be very suitable for this (but was not explored further in this thesis due to time constraints), is the concept of *subjective logic*, pioneered by Audun Jøsang [122, 123], which augments standard fuzzy logic terms with an uncertainty dimension (as well as a plausibility dimension, if required) in an intuitive manner, such that standard fuzzy logic operations can be performed in a way that results in the uncertainty over the terms to be straightforwardly propagated in the outcome. This means that such *subjective* operators could be used directly in place of fuzzy ones in the context of fuzzy validation operators examined herein, yielding validation outputs with an associated uncertainty that has been propagated in a natural way. This framework has the added benefit that a subjective term also has a one-to-one mapping with an equivalent beta distribution, meaning the validation output could be fed straight into a probabilistic framework for further statistical analysis, if required.

Characterisation of modes of segmentation failure

Future work in this area would expand on the theoretical and proof-of-concept examples presented in this thesis and test them on larger datasets and with specific segmentation algorithms, allowing a better understanding of when and how these algorithms fail in general.

Further work would also include the use of such qualitative and quantitative characterisation for guiding segmentation algorithm optimisation more effectively using *clinical* criteria, and away from specific *types* of failure; this could be tested both in the context of designing stronger classifiers from weaker ones, or

combining segmentation outputs using such clinical / anatomical localised measures of uncertainty, as per our methods described in chapter 3.

Explore application areas outside medical image analysis

The above concepts were presented in the context of medical image analysis. However, there's no reason why any of the above techniques couldn't be used in more traditional image analysis and computer vision techniques.

As an example, many computer vision techniques, including Convolutional Neural Networks (CNN) which is current state of the art for feature extraction [124, 125], rely on the detection of 'features' using 'feature detectors', or 'kernels'. In a CNN, the feature detector is convolved with an image to create a 'feature map', i.e. an image (typically of the same size as the original), yielding a measure at each pixel position, of how likely the area around that pixel is to contain that feature, where this is measured as the correlation of the intensity values in that area with those of the kernel; the objective of the CNN then is to figure out the collection of features with the highest predictive value.

The convolution's "base" operation (i.e. 'multiply then sum') at each position could easily be replaced with a validation operation, such that the feature map is essentially conceptually equivalent to a validation map, obtained with a particular fuzzy operator. In other words, we can introduce "optimality" or "pessimality" in our detectors by choosing the Gödel or Łukasiewicz operators as appropriate. Since a CNN usually relies on a sequence of convolution iterations, this could become part of a strategy, e.g. start with optimal detection, but progressively move to more and more selective detectors by making use of an appropriately parameterised Gödel - Łukasiewicz family of t-norms.

Alternatively, a d-norm operator could be used to ensure optimal feature detection for features that are expected to exhibit a particular orientation as well as texture; the fuzzy orientations of the kernel elements could be gradient-based, or preferably predefined for optimal feature detection.

In the same way a d-norm operator operates on pixels that are essentially 2D vector metrics (i.e. exhibiting intensity and directionality), we could extend this principle further to images of multidimensional pixels, such as colour (i.e. RGB) images, or depth images (i.e. containing RGBD pixels). Feature detection could consist of a suitable RGBD kernel, and validation sweeps performed in each dimension, resulting in a fuzzy feature map per dimension, which can then be collapsed into a single feature map through fuzzy intersection or fuzzy union (weighted or otherwise).

Summary

- We have presented a thesis on soft segmentations defined using fuzzy and probabilistic measures, and measures of uncertainty over them. We examined the semantics of fuzziness in such segmentations, and proposed methods to make use of measures of uncertainty defined over such fuzzy and probabilistic segmentations; a method to combine segmentations using uncertainty and a method to make use of a measure of clinical uncertainty to improve segmentation outcomes and clinical usability.
- Furthermore we provided a theoretical framework and predictions on the semantics of fuzziness with respect to the validation of soft segmentations, and used it to propose a novel validation method suitable for soft segmentations, as well as prove that the conventional approach currently used in the literature is unreliable and should be abandoned.
- Finally we demonstrated how to characterise modes of segmentation failure by using local-performance maps and fuzzy spatial / anatomical relationship masks to express performance as a function of the direction or distance from the gold standard or other point of interest, and validation sweeps for identifying failure due to the unwanted presence of specific features in the segmentation.
- Future work will focus on further experimentation with larger datasets and more algorithmic substrates, aiming for publication; and further exploration of the concepts discussed within, aiming to create a more integrated framework of uncertainty in soft segmentation and validation.
- While the applications presented in this dissertation have been demonstrated in the context of medical image segmentation, the methods and concepts presented in this thesis can be generalised to other areas of application, such as image analysis and computer vision in general.

Appendices

A

Code implementations

Contents

A.1 Context-specific directed t-norm for square 2D pixels

```
%%% in file 'exactDNorm.m' %%%

function Tn = exactDNorm (FVs,FVg,Gsx,Gsy,Ggx,Ggy,PlotFlag,CanonicalGradient)
% This function returns a 'TNorm' from two fuzzy values representing pixels
% originating from a segmentation image and a gold-standard image
% respectively, for which the local gradients are known (i.e.  can be
% calculated).
%
% To elaborate:  in the same way Goedel is subpixel overlap when both coming
% from the 'left', product is overlap when coming from 'left' and 'below',
% and lukasiewicz is overlap when coming from 'left' and 'right', this
% TNorm would be the overlap of the two components which when coming from
% the direction specified by the local orientation / direction give that
% fuzzy value.
%
% Inputs:
% FVs -> The Fuzzy Value of the segmentation pixel
% FVg -> The Fuzzy Value of the gold-standard pixel
% Gsx -> The horizontal ('x-axis') Gradient in the segmentation pixel.
% Gsy -> The vertical ('y-axis') Gradient in the segmentation pixel.
% Ggx -> The horizontal ('x-axis') Gradient in the gold-standard pixel.
```

```
% Ggy -> The vertical ('y-axis') Gradient in the gold-standard pixel.
% Plotflag -> If true, show a graphical representation of the intersection
% CanonicalGradient -> If false, the gradients are inverted, so as to be
% co-directional to the segmentation and GS front propagation.  If true, the
% gradients naturally increase from 0 to 1, and are therefore
% anti-directional.
% In other words, the actual situation is that the front evolves in a
% direction "opposite" to the gradient, but this is less intuitive to
% visualise, so the option to invert the gradients is provided; given the
% symmetry of the situation, the end-value of the Norm does not change.
%
% Output:
% Tn -> The resulting T-norm value.
%
% Please note that some edge-cases do not produce a visual representation even
% if Plotflag is set to true.  (e.g.  0-valued Fuzzy values or Gradients).

% Copyright Oct 2014 Tasos Papastylianou
% UNVECTORISED VERSION - SCALAR INPUT / OUTPUTS.

  %% Sanitisze inputs
  assert(FVs >=0 && FVs<=1, ...
    'Fuzzy Value for Segmentation cannot be outside of the 0-1 range');
  assert(FVg >=0 && FVg<=1, ...
    'Fuzzy Value for Segmentation cannot be outside of the 0-1 range');

  if nargin < 7; PlotFlag = false; end
  if nargin < 8; CanonicalGradient = false; end

  %% Uncomment lines below if interested in debugging iterations
  % (e.g.  for debugging arrayfun calls)
  %persistent debug_iters;
  %if isempty(debug_iters); debug_iters=0; else debug_iters=debug_iters+1; end;

  %% Preprocess Gradients
  % The algorithm treats the variables Gsx, Gsy, Ggx and Ggy as the directions
  % of expansion of the growing fronts, rather than the true gradients which are
  % anti-directional.  This is done because it is conceptually easier to
  % visualise and work with the former.  The correction below only serves for
  % visualisation purposes; the end-value is unaffected because of symmetry.
  if CanonicalGradient; Gsx = -Gsx; Gsy = -Gsy; Ggx = -Ggx; Ggy = -Ggy; end

  %(gradients of zero - within tolerance) - default to goedel norm
  if Gsx < 0.001 && Gsx > -0.001; Gsx = 0; end
  if Gsy < 0.001 && Gsy > -0.001; Gsy = 0; end
  if Ggx < 0.001 && Ggx > -0.001; Ggx = 0; end
  if Ggy < 0.001 && Ggy > -0.001; Ggy = 0; end

  %% Take care of edgecases
  % Full pixels
  if (FVs == 1 || FVg == 1); Tn = min(FVs,FVg); return; end
  if (FVs == 0 || FVg == 0); Tn = 0; return; end
  if (Gsx == 0 && Gsy == 0) || (Ggx == 0 && Ggy == 0)
    Tn = goedel(FVs,FVg); return;
```

```
  end

  % -> Past this point, nonzero gradients can be assumed :)

  % Get 2 points where the segmentation 'front' (line) intersects the
  % pixel square's sides.  Also get the square's corner from which the
  % gradient is advancing.
  [sx1, sy1, sx2, sy2, sx0, sy0] = getTnormComponent(FVs,Gsx,Gsy);
  [gx1, gy1, gx2, gy2, gx0, gy0] = getTnormComponent(FVg,Ggx,Ggy);
  Tn = calculateTnormFromComponents( ...
         sx1,sy1,sx2,sy2,sx0,sy0,gx1,gy1,gx2,gy2,gx0,gy0,Gsx,Gsy,Ggx,Ggy);

  % plot points
  if PlotFlag == true
    plotSquare; hold on;
    axis equal; axis([-0.1 1.1 -0.1 1.1]);
    plotSquareTangents(Gsx,Gsy,'r:');
    plot([sx1 sx2]',[sy1 sy2]','r','linewidth',2);
    plotSquareTangents(Ggx,Ggy,'g:');
    plot([gx1 gx2]',[gy1 gy2]','g','linewidth',2);
    scatter(sx0,sy0,10,'r','filled');
    scatter(gx0,gy0,5,'g','filled');
    hold off;
    title(['Tnorm Result = ' num2str(Tn)]);
  end

end

function plotSquare
  plot([0 1]',[0 0]','k', 'linewidth',2, ...
       [1 1]',[0 1]','k', 'linewidth',2, ...
       [1 0]',[1 1]','k', 'linewidth',2, ...
       [0 0]',[1 0]','k', 'linewidth',2);
end

function plotSquareTangents(Gx,Gy,C)
  x = 0:1;
  m = -Gx * (Gy ^ -1); % written like this to avoid div by zero warnings
  x1 = 0; y1 = 0;
  x2 = 1; y2 = 0;
  x3 = 1; y3 = 1;
  x4 = 0; y4 = 1;
  b1 = y1 - m*x1;
  b2 = y2 - m*x2;
  b3 = y3 - m*x3;
  b4 = y4 - m*x4;
  plot(x,m*x+b1,C,'linewidth',2, ...
       x,m*x+b2,C,'linewidth',2, ...
       x,m*x+b3,C,'linewidth',2, ...
       x,m*x+b4,C);
end

function I = goedel(A,B)
  I = min(A,B);
```

```
end

function [x1, y1, x2, y2, x0, y0] = getTnormComponent(FV,Gx,Gy)
  % Create "Triangle Crawling" vectors, such that they have the same direction
  % as the initial gradient components, but with magnitudes reversed, and
  % normalised such that the largest component has magnitude 1.
  Tmin = min(abs(Gx),abs(Gy)); Tmax = max(abs(Gx),abs(Gy));
  Tx = sign(Gx) * abs(Gy) / Tmax; if sign(Gx) == 0; Tx = 1; end
  Ty = sign(Gy) * abs(Gx) / Tmax; if sign(Gy) == 0; Ty = 1; end

  %Determine corner of square pixel to serve as the origin.
  if Gx >= 0; x0 = 0; else x0 = 1; end
  if Gy >= 0; y0 = 0; else y0 = 1; end

  %% Find the points where the segmentation line intersects the pixel boundary
  %
  % V represents the area of the triangle in the
  % following pic, where one of the triangle sides is equal to the square side
  % (i.e. equal to '1'), and the other side has a slope defined by the gradient
  % of the segmentation pixel. The following scheme is made by four lines of
  % that slope, each going through one of the square's corners.
  %    _____
  %   |\ \  |
  %   | \ \ |
  %   |__\_\|
  %
  % Also note that the 'short' side corresponds to the 'long' gradient component
  % and vice versa, as per the definition of the "Triangle Crawling" vectors
  % above.
  V = abs(Tx * Ty / 2);
  if FV <= V
    % If the expanding front represented by the Fuzzy Value has not exceeded V
    % then it can be represented as a 'scaled-down' version of V.
    % Therefore, for a multiplier 'a', you can get a (FV) triangle such that
    % (aTx * aTy) / 2 = FV
    a = sqrt((FV * 2)/abs(Tx .* Ty));
    x1 = (Tx * a) + x0; y1 = y0;
    x2 = x0; y2 = (Ty * a) + y0;
  elseif FV > V && FV <= (1-V)
    % In this case, if FV has expanded into the central 'parallelogram' portion
    % its value represents the triangle V, plus a scaled-down version of the
    % central parallelogram in the diagram above, which has area 1-2V.
    % The amount of scaling required is b = (FV-V)/(1-2V).
    b = (FV-V)/(1 - V - V);
    if abs(Tx) <= abs(Ty);
      x1 = x0 + Tx + (1-abs(Tx)) * sign(Tx) * b; y1 = y0;
        if sign(Tx) == 0; x1 = x0 + sign(Gx) * b; end
      x2 = x0 + (1-abs(Tx)) * sign(Tx) * b; y2 = y0 + Ty;
        if sign(Tx) == 0; x2 = x0 + sign(Gx) * b; end
    else % Tx > Ty
      x1 = x0; y1 = y0 + Ty + (1 - abs(Ty)) * sign(Ty) * b;
        if sign(Ty) == 0; y1 = y0 + sign(Gy) * b; end
      x2 = x0 + Tx; y2 = y0 + (1-abs(Ty)) * sign(Ty) * b;
        if sign(Ty) == 0; y2 = y0 + sign(Gy) * b; end
```

```
    end
  else
    % FV has expanded past the parallelogram portion and into the final
    % triangle.  This case is equivalent to flipping the starting points and
    % gradient vectors, and then calculating in the same way as the first case.
    x0t = double( x0); y0t = double( y0); Txt = -Tx; Tyt = -Ty;
    a = sqrt((2 * (1-FV)) / abs (Txt .* Tyt));
    x1 = (Txt * a) + x0t; y1 = y0t;
    x2 = x0t; y2 = (Tyt * a) + y0t;
  end
end

function Tn = calculateTnormFromComponents( ...
                sx1,sy1,sx2,sy2,sx0,sy0,gx1,gy1,gx2,gy2,gx0,gy0,Gsx,Gsy,Ggx,Ggy)

  s0 = [sx0 sy0]'; s1 = [sx1 sy1]'; s2 = [sx2 sy2]';
  g0 = [gx0 gy0]'; g1 = [gx1 gy1]'; g2 = [gx2 gy2]';
  [Ix, Iy] = getIntersectionPoint(sx1,sy1,sx2,sy2,gx1,gy1,gx2,gy2);

  if isnan(Ix) || isnan(Iy) % lines coincide
    % Special Cases
    if (s0  = g0); Tn = 0; return; end
    if (sx1 == sx2) && (sx0  = gx0); Tn = 0; return; end
    if (sy1 == sy2) && (sy0  = gy0); Tn = 0; return; end
    if (sx1 == sx2) && (sx0 == 0); Tn = sx1; return; end
    if (sx1 == sx2) && (sx0 == 1); Tn = 1 - sx1; return; end

    % General Case
    OriginSign = getSideOfCurve(sx0,sy0,sx1,sy1,sx2,sy2);
    if OriginSign == 0
      OriginSign = -getSideOfCurve(0.5,0.5,sx1,sy1,sx2,sy2);
    end

    ContributingPoints = [s1,s2];
    if (getSideOfCurve(0,0,sx1,sy1,sx2,sy2) == OriginSign)
      ContributingPoints = [ContributingPoints, [0,0]'];
    end
    if (getSideOfCurve(1,0,sx1,sy1,sx2,sy2) == OriginSign)
      ContributingPoints = [ContributingPoints, [1,0]'];
    end
    if (getSideOfCurve(1,1,sx1,sy1,sx2,sy2) == OriginSign)
      ContributingPoints = [ContributingPoints, [1,1]'];
    end
    if (getSideOfCurve(0,1,sx1,sy1,sx2,sy2) == OriginSign)
      ContributingPoints = [ContributingPoints, [0,1]'];
    end

    Centroid = mean(ContributingPoints,2);
    Tn = getTnArea(Centroid,ContributingPoints);

  elseif (Ix < 0 || Ix > 1) || (Iy < 0 || Iy > 1) % lines do not intersect
                                                  % within pixel

    % General Case
```

```
  ContributingPoints = [];
  Centroid = mean([s1,s2,g1,g2],2); % centroid of Seg and Gs line endpoints

  % If the Segmentation front has expanded beyond the centroid, then the
  % GS points contribute to the TnArea regardless of the GS origin
  % (and vice-versa)
  [ix iy] = getIntersectionPoint ( ...
              sx0, sy0, Centroid(1), Centroid(2), sx1, sy1, sx2, sy2);
  if norm([ix iy]'-s0) > norm(Centroid - s0)
    ContributingPoints = [ContributingPoints, g1, g2];
  end

  [ix iy] = getIntersectionPoint( ...
              gx0, gy0, Centroid(1), Centroid(2), gx1, gy1, gx2, gy2);
  if norm([ix iy]'-g0) > norm(Centroid - g0)
    ContributingPoints = [ContributingPoints, s1, s2];
  end

  if length(ContributingPoints) == 0; Tn = 0; return; end

  SegOriginSign = getSideOfCurve(sx0,sy0,sx1,sy1,sx2,sy2);
  if SegOriginSign == 0
    SegOriginSign = -getSideOfCurve(0.5,0.5,sx1,sy1,sx2,sy2);
  end

  GsOriginSign = getSideOfCurve(gx0,gy0,gx1,gy1,gx2,gy2);
  if GsOriginSign == 0
    GsOriginSign = -getSideOfCurve(0.5,0.5,gx1,gy1,gx2,gy2);
  end

  if (getSideOfCurve(0,0,sx1,sy1,sx2,sy2) == SegOriginSign) ...
     && (getSideOfCurve(0,0,gx1,gy1,gx2,gy2) == GsOriginSign)
    ContributingPoints = [ContributingPoints, [0,0]'];
  end
  if (getSideOfCurve(1,0,sx1,sy1,sx2,sy2) == SegOriginSign) ...
     && (getSideOfCurve(1,0,gx1,gy1,gx2,gy2) == GsOriginSign)
    ContributingPoints = [ContributingPoints, [1,0]'];
  end
  if (getSideOfCurve(1,1,sx1,sy1,sx2,sy2) == SegOriginSign) ...
     && (getSideOfCurve(1,1,gx1,gy1,gx2,gy2) == GsOriginSign)
    ContributingPoints = [ContributingPoints, [1,1]'];
  end
  if (getSideOfCurve(0,1,sx1,sy1,sx2,sy2) == SegOriginSign) ...
     && (getSideOfCurve(0,1,gx1,gy1,gx2,gy2) == GsOriginSign)
    ContributingPoints = [ContributingPoints, [0,1]'];
  end

  % Find centroid of the Contributing Points
  Centroid = mean(ContributingPoints,2);
  Tn = getTnArea(Centroid,ContributingPoints);

else % lines intersect within pixel

  % For vectors a and b, defined as having origin point (Ix,Iy), i.e.  the
  % intersection point, and endpoints (gx1,gy1) and (gx2,gy2) respectively,
```

```
    % choose the one whose projection onto Gs has the opposite direction as Gs
    % (since it has already been passed by the segmentation front)
    % Note, the dot product won't be zero, because that case has been eliminated
    % in the first branch of this if block.
    if dot([Gsx,Gsy]',[gx1-Ix, gy1-Iy]') > 0; gopt = g2; else gopt = g1; end
    if dot([Ggx,Ggy]',[sx1-Ix, sy1-Iy]') > 0; sopt = s2; else sopt = s1; end

    % Find the square corners that contribute to the desired area; these are
    % the corners that are on the same side of
    % both their respective expanding curves as their origins of expansion.
    SegOriginSign = getSideOfCurve(sx0,sy0,sx1,sy1,sx2,sy2);
    if SegOriginSign == 0
      SegOriginSign = -getSideOfCurve(0.5,0.5,sx1,sy1,sx2,sy2);
    end

    GsOriginSign = getSideOfCurve(gx0,gy0,gx1,gy1,gx2,gy2);
    if GsOriginSign == 0
      GsOriginSign = -getSideOfCurve(0.5,0.5,gx1,gy1,gx2,gy2);
    end

    % Collect all points contributing to the desired area
    ContributingPoints = [[Ix,Iy]', sopt, gopt];
    if (getSideOfCurve(0,0,sx1,sy1,sx2,sy2) == SegOriginSign) ...
       && (getSideOfCurve(0,0,gx1,gy1,gx2,gy2) == GsOriginSign)
      ContributingPoints = [ContributingPoints, [0,0]'];
    end
    if (getSideOfCurve(1,0,sx1,sy1,sx2,sy2) == SegOriginSign) ...
       && (getSideOfCurve(1,0,gx1,gy1,gx2,gy2) == GsOriginSign)
      ContributingPoints = [ContributingPoints, [1,0]'];
    end
    if (getSideOfCurve(1,1,sx1,sy1,sx2,sy2) == SegOriginSign) ...
       && (getSideOfCurve(1,1,gx1,gy1,gx2,gy2) == GsOriginSign)
      ContributingPoints = [ContributingPoints, [1,1]'];
    end
    if (getSideOfCurve(0,1,sx1,sy1,sx2,sy2) == SegOriginSign) ...
       && (getSideOfCurve(0,1,gx1,gy1,gx2,gy2) == GsOriginSign)
      ContributingPoints = [ContributingPoints, [0,1]'];
    end

    % Find centroid of the Contributing Points
    Centroid = mean(ContributingPoints,2);
    Tn = getTnArea(Centroid,ContributingPoints);

  end
end

function [Ix Iy] = getIntersectionPoint(sx1, sy1, sx2, sy2, gx1, gy1, gx2, gy2)

%% Explanatory formulas derivation in latex format:
% y = \frac{sy_2 - sy_1}{sx_2 - sx_1}(x - sx_1) + sy_1
%   = \frac{gy_2 - gy_1}{gx_2 - gx_1}(x - gx_1) + gy_1 \Rightarrow
%     \frac{sy_2 - sy_1}{sx_2 - sx_1}x
%   - \frac{sy_2 - sy_1}{sx_2 - sx_1}sx_1
%   + sy_1 = \frac{gy_2 - gy_1}{gx_2 - gx_1}x
%   - \frac{gy_2 - gy_1}{gx_2 - gx_1}gx_1 + gy_1 \Rightarrow
```

```
% x \left ( \frac{sy_2 - sy_1}{sx_2 - sx_1}
%   - \frac{gy_2 - gy_1}{gx_2 - gx_1} \right )
%   = gy_1 - sy_1 + \frac{sy_2 - sy_1}{sx_2 - sx_1}sx_1
%   - \frac{gy_2 - gy_1}{gx_2 - gx_1}gx_1 \Rightarrow
% x = \frac{gy_1 - sy_1 + \frac{sy_2 - sy_1}{sx_2 - sx_1}sx_1
%   - \frac{gy_2 - gy_1}{gx_2 - gx_1}gx_1}{\frac{sy_2 - sy_1}{sx_2 - sx_1}
%   - \frac{gy_2 - gy_1}{gx_2 - gx_1}},
% y = \frac{sy_2 - sy_1}{sx_2 - sx_1}(x - sx_1) + sy_1


  % Special Cases
  if isequal (sx1, sx2, gx1, gx2); Ix = nan; Iy = nan; return; end
  if isequal (sy1, sy2, gy1, gy2); Ix = nan; Iy = nan; return; end

  if ((sx1 == sx2) && (gx1 == gx2)); Ix = inf; Iy = inf; return; end
  if ((sy1 == sy2) && (gy1 == gy2)); Ix = inf; Iy = inf; return; end;

  if (sx1 == sx2); Ix = sx1; Iy = (Ix-gx1)*(gy2-gy1)/(gx2-gx1)+gy1; return; end
  if (gx1 == gx2); Ix = gx1; Iy = (Ix-sx1)*(sy2-sy1)/(sx2-sx1)+sy1; return; end

  % General case
  if ((sy2-sy1)/(sx2-sx1) - (gy2-gy1)/(gx2-gx1))  = 0
    denom = 1 / ((sy2-sy1)/(sx2-sx1) - (gy2-gy1)/(gx2-gx1));
  else
    denom = inf; % multiply by inf instead of divide by zero to avoid warnings
  end
  Ix = (gy1-sy1+sx1*(sy2-sy1)/(sx2-sx1)-gx1*(gy2-gy1)/(gx2-gx1)) * denom;
% i.e.  Ix = (gy1-sy1+sx1*(sy2-sy1)/(sx2-sx1)-gx1*(gy2-gy1)/(gx2-gx1))
% / ((sy2-sy1)/(sx2-sx1) - (gy2-gy1)/(gx2-gx1));
  Iy = (Ix-sx1)*(sy2-sy1)/(sx2-sx1)+sy1;
end

function SideOfCurve = getSideOfCurve(X,Y,x1,y1,x2,y2)
  % Returns -1 or +1 depending on which side of the curve point X,Y is, below
  % or above (or 0 if point is on the curve)
  if (x2-x1)  = 0
    denom = 1/(x2-x1);
  else
    denom = inf; % multiply by inf instead of divide by zero to avoid warnings
  end
  SideOfCurve = sign(Y - ((X-x1)*(y2-y1)*(denom)+y1));
% i.e.  SideOfCurve = sign(Y - ((X-x1)*(y2-y1)/(x2-x1)+y1));
end

function TnArea = getTnArea(Centroid,ContributingPoints)
  PointVectors = (ContributingPoints - ...
                  repmat(Centroid,1,size(ContributingPoints,2)));
  PointAngles = atan2d(PointVectors(2,:),PointVectors(1,:));
  [ , I] = sort(PointAngles);
  P1 = PointVectors(:,I);
  P2 = P1(:,[end,1:end-1]);
  P1 = [P1;zeros(1,size(P1,2))];
  P2 = [P2;zeros(1,size(P2,2))];
  Areas = abs(cross(P1,P2))/2; % Works without need for norms because
                               % plane hence 0 elements
```

```
  TnArea = totalsum(Areas);
end

function Out = totalsum(A)
  Out = sum(A(:));
end
```

References

[1] Caroline Petitjean and Jean-Nicolas Dacher. "A review of segmentation methods in short axis cardiac MR images." In: *Med Image Anal* 15.2 (Apr. 2011), pp. 169–184. URL: http://dx.doi.org/10.1016/j.media.2010.12.004.

[2] Nelly Gordillo, Eduard Montseny, and Pilar Sobrevilla. "State of the art survey on MRI brain tumor segmentation". In: *Magnetic Resonance Imaging* 31.8 (2013), pp. 1426–1438. URL: http://www.sciencedirect.com/science/article/pii/S0730725X13001872.

[3] Zhen Ma, João Manuel R.S. Tavares, Renato Natal Jorge, and T. Mascarenhas. "A review of algorithms for medical image segmentation and their applications to the female pelvic cavity". In: *Computer Methods in Biomechanics and Biomedical Engineering* 13.2 (2010). PMID: 19657801, pp. 235–246. eprint: http://dx.doi.org/10.1080/10255840903131878. URL: http://dx.doi.org/10.1080/10255840903131878.

[4] Alireza Norouzi, Mohd Shafry Mohd Rahim, Ayman Altameem, Tanzila Saba, Abdolvahab Ehsani Rad, Amjad Rehman, and Mueen Uddin. "Medical Image Segmentation Methods, Algorithms, and Applications". In: *IETE Technical Review* 31.3 (2014), pp. 199–213. eprint: http://dx.doi.org/10.1080/02564602.2014.906861. URL: http://dx.doi.org/10.1080/02564602.2014.906861.

[5] Sepideh Yazdani, Rubiyah Yusof, Alireza Karimian, Mohsen Pashna, and Amirshahram Hematian. "Image Segmentation Methods and Applications in MRI Brain Images". In: *IETE Technical Review* 32.6 (2015), pp. 413–427. eprint: http://dx.doi.org/10.1080/02564602.2015.1027307. URL: http://dx.doi.org/10.1080/02564602.2015.1027307.

[6] Koen L Vincken, Andrés E Koster, and Max A Viergever. "Probabilistic segmentation of partial volume voxels". In: *Pattern Recognition Letters* 15.5 (1994), pp. 477–484.

[7] Einar Heiberg, Martin Ugander, Henrik Engblom, Matthias Götberg, Goran K Olivecrona, David Erlinge, and Hakan Arheden. "Automated quantification of myocardial infarction from MR images by accounting for partial volume effects: animal, phantom, and human study 1". In: *Radiology* 246.2 (2008), pp. 581–588.

[8] Miguel Ángel González Ballester, Andrew P. Zisserman, and Michael Brady. "Estimation of the partial volume effect in MRI". In: *Medical Image Analysis* 6.4 (2002), pp. 389–405. URL: http://www.sciencedirect.com/science/article/pii/S1361841502000610.

[9] Petronella Anbeek, Koen L Vincken, Matthias JP van Osch, Robertus HC Bisschops, and Jeroen van der Grond. "Automatic segmentation of different-sized white matter lesions by voxel probability estimation". In: *Medical image analysis* 8.3 (2004), pp. 205–215.

[10] J. Ashburner and K.J. Friston. "Unified segmentation". In: *NeuroImage* 26 (2005), pp. 839–851.

[11] M. Jorge Cardoso, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C. Fox, and Sebastien Ourselin. "STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation". In: *Medical Image Analysis* 17.6 (2013), pp. 671–684. URL: http://www.sciencedirect.com/science/article/pii/S1361841513000200.

[12] I. Isgum, M. Staring, A. Rutten, M. Prokop, M.A. Viergever, and B. Van Ginneken. "Multi-Atlas-Based Segmentation With Local Decision Fusion - Application to Cardiac and Aortic Segmentation in CT Scans". In: *Medical Imaging, IEEE Transactions on* 28.7 (2009), pp. 1000–1010.

[13] Rashed Karim, Piet Claus, Zhong Chen, R.James Housden, Samantha Obom, Harminder Gill, YingLiang Ma, Prince Acheampong, Mark O'Neill, Reza Razavi, Tobias Schaeffter, and KawalS. Rhode. "Infarct Segmentation Challenge on Delayed Enhancement MRI of the Left Ventricle". In: *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges*. Ed. by Oscar Camara, Tommaso Mansi, Mihaela Pop, Kawal Rhode, Maxime Sermesant, and Alistair Young. Vol. 7746. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 97–104. URL: http://dx.doi.org/10.1007/978-3-642-36961-2%5C_12.

[14] Victor Lempitsky, Michael Verhoek, J Alison Noble, and Andrew Blake. "Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography". In: *Functional Imaging and Modeling of the Heart*. Springer, 2009, pp. 447–456.

[15] Ján Margeta, Ezequiel Geremia, Antonio Criminisi, and Nicholas Ayache. "Layered spatio-temporal forests for left ventricle segmentation from 4D cardiac MRI data". In: *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges*. Springer, 2012, pp. 109–119.

[16] J. A. Noble and D. Boukerroui. "Ultrasound image segmentation: a survey". In: *IEEE Transactions on Medical Imaging* 25.8 (Aug. 2006), pp. 987–1010.

[17] Amol Pednekar, Uday Kurkure, Raja Muthupillai, Scott Flamm, and Ioannis A Kakadiaris. "Automated left ventricular segmentation in cardiac MRI". In: *Biomedical Engineering, IEEE Transactions on* 53.7 (2006), pp. 1425–1428.

[18] Zhao Yi, Antonio Criminisi, Jamie Shotton, and Andrew Blake. "Discriminative, Semantic Segmentation of Brain Tissue in MR Images". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*. Ed. by Guang-Zhong Yang, David Hawkes, Daniel Rueckert, Alison Noble, and Chris Taylor. Vol. 5762. Lecture Notes in Computer Science. 10.1007/978-3-642-04271-3_68. Springer Berlin / Heidelberg, 2009, pp. 558–565. URL: http://dx.doi.org/10.1007/978-3-642-04271-3%5C_68.

[19] Kelly H Zou, William M Wells III, Michael R Kaus, Ron Kikinis, Ferenc A Jolesz, and Simon K Warfield. "Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumors". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2002*. Springer, 2002, pp. 315–322.

[20] P.K. Saha and J.K. Udupa. "Optimum image thresholding via class uncertainty and region homogeneity". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.7 (July 2001), pp. 689–706.

[21] Tasos Papastylianou, Christopher Kelly, Benjamin Villard, Erica Dall' Armellina, and Vicente Grau. "Fuzzy Segmentation of the Left Ventricle in Cardiac MRI Using Physiological Constraints". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2015, pp. 231–239.

[22] Tasos Papastylianou, Erica Dall' Armellina, and Vicente Grau. "Orientation-Sensitive Overlap Measures for the Validation of Medical Image Segmentations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 361–369.

[23] Adam Hoover, Valentina Kouznetsova, and Michael Goldbaum. "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response". In: *Medical Imaging, IEEE Transactions on* 19.3 (2000), pp. 203–210.

[24] *Cardiac MR Left Ventricle Segmentation Challenge*. Kitware Inc. Midas Journal, 2009. URL: http://www.midasjournal.org/browse/journal/49.

[25] Caroline Petitjean, Maria A Zuluaga, Wenjia Bai, Jean-Nicolas Dacher, Damien Grosgeorge, Jérôme Caudron, Su Ruan, Ismail Ben Ayed, M Jorge Cardoso, Hsiang-Chou Chen, et al. "Right ventricle segmentation from cardiac MRI: a collation study". In: *Medical image analysis* 19.1 (2015), pp. 187–202.

[26] Phi Vu Tran. "A fully convolutional neural network for cardiac segmentation in short-axis MRI". In: *arXiv preprint arXiv:1604.00494* (2016).

[27] Torsten Rohlfing and Calvin R Maurer. "Shape-based averaging". In: *IEEE Transactions on Image Processing* 16.1 (2007), pp. 153–161.

[28] Richard Francis Mould. "The early history of x-ray diagnosis with emphasis on the contributions of physics 1895-1915". In: *Physics in medicine and biology* 40.11 (1995), p. 1741.

[29] William R Hendee. "Cross sectional medical imaging: a history." In: *Radiographics* 9.6 (1989), pp. 1155–1180.

[30] Joseph J Schreiber, Paul A Anderson, Humberto G Rosas, Avery L Buchholz, and Anthony G Au. "Hounsfield units for assessing bone mineral density and strength: a tool for osteoporosis management". In: *The Journal of Bone & Joint Surgery* 93.11 (2011), pp. 1057–1063.

[31] James S Duncan and Nicholas Ayache. "Medical image analysis: Progress over two decades and the challenges ahead". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.1 (2000), pp. 85–106.

[32] Simon K Warfield, Kelly H Zou, and William M Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation". In: *Medical Imaging, IEEE Transactions on* 23.7 (2004), pp. 903–921.

[33] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* smc-9.1 (Jan. 1979), pp. 62–66.

[34] Evangelia Micheli-Tzanakou, Elsa Angelini, Yinpeng Jin, and Andrew Laine. "State of the Art of Level Set Methods in Segmentation and Registration of Medical Imaging Modalities". In: *Handbook of Biomedical Image Analysis*. Ed. by Jasjit S. Suri, David L. Wilson, and Swamy Laxminarayan. Topics in Biomedical Engineering. International Book Series. 10.1007/0-306-48608-3_2. Springer US, 2005, pp. 47–101. URL: http://dx.doi.org/10.1007/0-306-48608-3%5C_2.

[35] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. "Snakes: Active contour models". In: *International journal of computer vision* 1.4 (1988), pp. 321–331.

[36] Chenyang Xu and Jerry L Prince. "Snakes, shapes, and gradient vector flow". In: *Image Processing, IEEE Transactions on* 7.3 (1998), pp. 359–369.

[37] Jianbo Shi and J. Malik. "Normalized cuts and image segmentation". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.8 (Aug. 2000), pp. 888–905.

[38] Anil K Attili, Andreas Schuster, Eike Nagel, Johan H C Reiber, and Rob J van der Geest. "Quantification in cardiac MRI: advances in image acquisition and processing." In: *Int J Cardiovasc Imaging* 26 Suppl 1 (Feb. 2010), pp. 27–40. URL: http://dx.doi.org/10.1007/s10554-009-9571-x.

[39] Timothy N. Bloomer, Sven Plein, Aleksandra Radjenovic, David M. Higgins, Timothy R. Jones, John P. Ridgway, and Mohan U. Sivananthan. "Cine MRI using steady state free precession in the radial long axis orientation is a fast accurate method for obtaining volumetric data of the left ventricle". In: *Journal of Magnetic Resonance Imaging* 14.6 (2001), pp. 685–692. URL: http://dx.doi.org/10.1002/jmri.10019.

[40] Andrew C. Larson, Richard D. White, Gerhard Laub, Elliot R. McVeigh, Debiao Li, and Orlando P. Simonetti. "Self-gated cardiac cine MRI". In: *Magnetic Resonance in Medicine* 51.1 (2004), pp. 93–102. URL: http://dx.doi.org/10.1002/mrm.10664.

[41] Manuel D. Cerqueira, Neil J. Weissman, Vasken Dilsizian, Alice K. Jacobs, Sanjiv Kaul, Warren K. Laskey, Dudley J. Pennell, John A. Rumberger, Thomas Ryan, and Mario S. Verani. "Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart". In: *Circulation* 105.4 (2002), pp. 539–542. eprint: http://circ.ahajournals.org/content/105/4/539.full.pdf+html. URL: http://circ.ahajournals.org/content/105/4/539.short.

[42] Niek H Prakken, Birgitta K Velthuis, EJ Vonken, Willem P Mali, and MJ Cramer. "Cardiac MRI: standardized right and left ventricular quantification by briefly coaching inexperienced personnel". In: *Open Magn Reson J* 1 (2008), pp. 104–111.

[43] Simon F Eskildsen and Lasse R Østergaard. "Active surface approach for extraction of the human cerebral cortex from MRI". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*. Springer, 2006, pp. 823–830.

[44] Tim McInerney and Demetri Terzopoulos. "Deformable models in medical image analysis: a survey". In: *Medical image analysis* 1.2 (1996), pp. 91–108.

[45] Wenzhe Shi, Xiahai Zhuang, Haiyan Wang, Simon Duckett, Declan Oregan, Philip Edwards, Sebastien Ourselin, and Daniel Rueckert. "Automatic Segmentation of Different Pathologies from Cardiac Cine MRI Using Registration and Multiple Component EM Estimation". In: *Functional Imaging and Modeling of the Heart*. Ed. by Dimitris Metaxas and Leon Axel. Vol. 6666. Lecture Notes in Computer Science. 10.1007/978-3-642-21028-0_21. Springer Berlin / Heidelberg, 2011, pp. 163–170. URL: http://dx.doi.org/10.1007/978-3-642-21028-0%5C_21.

[46] Tim Cootes. "Model-Based Methods in Analysis of Biomedical Images: An Introduction to Active Shape Models". In: *Image Processing And Analysis*. Ed. by R Baldock and J Graham. Oxford University Press, 2000. Chap. 7, pp. 223–248.

[47] Hans C van Assen, Mikhail G Danilouchkine, Alejandro F Frangi, Sebastián Ordás, Jos J M Westenberg, Johan H C Reiber, and Boudewijn P F Lelieveldt. "SPASM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data." In: *Med Image Anal* 10.2 (Apr. 2006), pp. 286–303. URL: http://dx.doi.org/10.1016/j.media.2005.12.001.

[48] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. "Active Appearance Models". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.6 (2001), pp. 681–685.

[49] Lisa Gottesfeld Brown. "A survey of image registration techniques". In: *ACM computing surveys (CSUR)* 24.4 (1992), pp. 325–376.

[50] Dan V Iosifescu, Martha E Shenton, Simon K Warfield, Ron Kikinis, Joachim Dengler, Ferenc A Jolesz, and Robert W McCarley. "An automated registration algorithm for measuring MRI subcortical brain structures". In: *Neuroimage* 6.1 (1997), pp. 13–25.

[51] Xiahai Zhuang, K.S. Rhode, R.S. Razavi, D.J. Hawkes, and S. Ourselin. "A Registration-Based Propagation Framework for Automatic Whole Heart Segmentation of Cardiac MRI". In: *Medical Imaging, IEEE Transactions on* 29.9 (Sept. 2010), pp. 1612–1625. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5444972.

[52] Pierre-François D-Haese, Valerie Duay, Thomas E Merchant, Benoit Macq, and Benoit M Dawant. "Atlas-based segmentation of the brain for 3-dimensional treatment planning in children with infratentorial ependymoma". In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*. Springer, 2003, pp. 627–634.

[53] Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Joseph V Hajnal, and Daniel Rueckert. "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy". In: *Neuroimage* 46.3 (2009), pp. 726–738.

[54] Yongfu Hao, Tianzi Jiang, and Yong Fan. "Iterative multi-atlas based segmentation with multi-channel image registration and Jackknife Context Model". In: *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*. IEEE. 2012, pp. 900–903.

[55] Chris A Cocosco, Wiro J Niessen, Thomas Netsch, Evert-Jan P A Vonken, Gunnar Lund, Alexander Stork, and Max A Viergever. "Automatic image-driven segmentation of the ventricles in cardiac cine MRI." In: *J Magn Reson Imaging* 28.2 (Aug. 2008), pp. 366–374. URL: http://dx.doi.org/10.1002/jmri.21451.

[56] J. Surowiecki. *The Wisdom of Crowds*. Knopf Doubleday Publishing Group, 2005. URL: http://books.google.co.uk/books?id=hHUsHOHqVzEC.

[57] Lior Rokach. "Ensemble-based classifiers". In: *Artificial Intelligence Review* 33.1-2 (2010), pp. 1–39.

[58] Robert E Schapire. "The boosting approach to machine learning: An overview". In: *Lecture Notes In Statistics - New York - Springer Verlag* (2003), pp. 149–172.

[59] Neil I Weisenfeld and Simon K Warfield. "SoftSTAPLE: Truth and performance-level estimation from probabilistic segmentations". In: *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE. 2011, pp. 441–446.

[60] Olivier Commowick and Simon K. Warfield. "A Continuous STAPLE for Scalar, Vector and Tensor Images: An Application to DTI Analysis". In: *IEEE Transactions on Medical Imaging* 28.6 (June 2009), pp. 838–846.

[61] Olivier Commowick and Simon K Warfield. "Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. Springer, 2010, pp. 25–32.

[62] Juan Eugenio Iglesias and Mert R Sabuncu. "Multi-atlas segmentation of biomedical images: a survey". In: *Medical image analysis* 24.1 (2015), pp. 205–219.

[63] Wenjia Bai, Wenzhe Shi, Declan P O'Regan, Tong Tong, Haiyan Wang, Shahnaz Jamil-Copley, Nicholas S Peters, and Daniel Rueckert. "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images". In: *IEEE transactions on medical imaging* 32.7 (2013), pp. 1302–1315.

[64] Y.J. Zhang. "A survey on evaluation methods for image segmentation". In: *Pattern Recognition* 29.8 (1996), pp. 1335–1346. URL: http://www.sciencedirect.com/science/article/pii/0031320395001697.

[65] Jayaram K Udupa, Vicki R Leblanc, Ying Zhuge, Celina Imielinska, Hilary Schmidt, Leanne M Currie, Bruce E Hirsch, and James Woodburn. "A framework for evaluating image segmentation algorithms". In: *Computerized Medical Imaging and Graphics* 30.2 (2006), pp. 75–87.

[66] George Klir and Bo Yuan. *Fuzzy sets and fuzzy logic*. Vol. 4. Prentice Hall New Jersey, 1995.

[67] Taffee T Tanimoto. "Elementary mathematical theory of classification and prediction". In: (1958).

[68] Lee R Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

[69] Lotfi A Zadeh. "Fuzzy sets". In: *Information and control* 8.3 (1965), pp. 338–353.

[70] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M Stadlan. "Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease". In: *Neurology* 34.7 (1984), pp. 939–939.

[71] George J Klir and Tina A Folger. "Fuzzy sets, uncertainty, and information". In: (1988).

[72] Isabelle Bloch and Henri Maître. "Fuzzy mathematical morphologies: a comparative study". In: *Pattern Recognition* 28.9 (1995), pp. 1341–1387.

[73] Richard T Cox. "Probability, frequency and reasonable expectation". In: *American journal of physics* 14.1 (1946), pp. 1–13.

[74] Rafael C Gonzalez, Richard E Woods, et al. *Digital image processing*. 2002.

[75] Peter Santago and Howard D Gage. "Statistical models of partial volume effect". In: *Image Processing, IEEE Transactions on* 4.11 (1995), pp. 1531–1540.

[76] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[77] C.M. Bishop. *Pattern Recognition And Machine Learning*. Information Science and Statistics. Springer, 2006. URL: http://books.google.co.uk/books?id=kTNoQgAACAAJ.

[78] Zhen Qian, Dimitris N Metaxas, and Leon Axel. "A learning framework for the automatic and accurate segmentation of cardiac tagged MRI images". In: *Computer Vision for Biomedical Image Applications*. Springer, 2005, pp. 93–102.

[79] E.Harvey Estes, Frank M. Dalton, Mark L. Entman, Henry B. Dixon, and Donald B. Hackel. "The anatomy and blood supply of the papillary muscles of the left ventricle". In: *American Heart Journal* 71.3 (1966), pp. 356–362. URL: http://www.sciencedirect.com/science/article/pii/0002870366904753.

[80] PA Bromiley, NA Thacker, and E Bouhova-Thacker. "Shannon entropy, Renyi entropy, and information". In: *Statistics and Inf. Series (2004-004)* (2004).

[81] Bart Kosko. "Fuzzy entropy and conditioning". In: *Information sciences* 40.2 (1986), pp. 165–174.

[82] T. Downarowicz. "Entropy". In: *Scholarpedia* 2.11 (2007). revision #126991, p. 3901.

[83] P. E. Latham and Y. Roudi. "Mutual information". In: *Scholarpedia* 4.1 (2009). revision #122173, p. 1658.

[84] Massimo De Santo, Consolatina Liguori, and Antonio Pietrosanto. "Uncertainty characterization in image-based measurements: a preliminary discussion". In: *Instrumentation and Measurement, IEEE Transactions on* 49.5 (2000), pp. 1101–1107.

[85] A Moreno, C Takemura, O Colliot, O Camara, and I Bloch. "Heart segmentation in medical images using the fuzzy spatial relation "Between"". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU*. 2006, pp. 2052–2059.

[86] A Moreno, Celina Maki Takemura, Olivier Colliot, Oscar Camara, and Isabelle Bloch. "Using anatomical knowledge expressed as fuzzy constraints to segment the heart in CT images". In: *Pattern Recognition* 41.8 (2008), pp. 2525–2540.

[87] Laszlo Balkay. *DICOMDIR reader*. [software]. University of Debrecen. 2011. URL: http://www.mathworks.co.uk/matlabcentral/fileexchange/7926-dicomdir-reader.

[88] *MATLAB, v8.2 (R2013b)*. The MathWorks Inc. 2012. URL: http://www.mathworks.com.

[89] John W. Eaton, David Bateman, Søren Hauberg, and Rik Wehbring. *GNU Octave version 4.2.0 manual: a high-level interactive language for numerical computations*. 2016. URL: http://www.gnu.org/software/octave/doc/interpreter.

[90] Nikos Komodakis, Georgios Tziritas, and Nikos Paragios. "Fast, approximately optimal solutions for single and dynamic MRFs". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE. 2007, pp. 1–8.

[91] Ben Glocker, Aristeidis Sotiras, Nikos Komodakis, and Nikos Paragios. "Deformable Medical Image Registration: Setting the State of the Art with Discrete Methods*". In: *Annual review of biomedical engineering* 13 (2011), pp. 219–244.

[92] U Sechtem, PW Pflugfelder, RG Gould, MM Cassidy, and Ch B Higgins. "Measurement of right and left ventricular volumes in healthy individuals with cine MR imaging." In: *Radiology* 163.3 (1987), pp. 697–702.

[93] M. Hemmendorff, M.T. Andersson, T. Kronander, and H. Knutsson. "Phase-based multidimensional volume registration". In: *Medical Imaging, IEEE Transactions on* 21.12 (Dec. 2002), pp. 1536–1543.

[94] Gabor T Herman, Jingsheng Zheng, and C Bucholtz. "Shape-based interpolation". In: *IEEE Computer Graphics and Applications* 12.3 (1992), pp. 69–79.

[95] Fa-Yueh Wu. "The potts model". In: *Reviews of modern physics* 54.1 (1982), p. 235.

[96] Torsten Rohlfing and Calvin R Maurer. "Multi-classifier framework for atlas-based image segmentation". In: *Pattern Recognition Letters* 26.13 (2005), pp. 2070–2079.

[97] Punam Kumar Saha, Bipul Das, and Felix W. Wehrli. "An object class-uncertainty induced adaptive force and its application to a new hybrid snake". In: *Pattern Recognition* 40.10 (2007), pp. 2656–2671. URL: http://www.sciencedirect.com/science/article/pii/S0031320307000222.

[98] J Petrovičová. "On the entropy of partitions in product MV algebras". In: *Soft Computing* 4.1 (2000), pp. 41–44.

[99] Dymitr Ruta and Bogdan Gabrys. "An overview of classifier fusion methods". In: *Computing and Information systems* 7.1 (2000), pp. 1–10.

[100] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. "Learning from crowds". In: *Journal of Machine Learning Research* 11.Apr (2010), pp. 1297–1322.

[101] Lance M Kaplan, Supriyo Chakraborty, and Chatschik Bisdikian. "Fusion of classifiers: A subjective logic perspective". In: *Aerospace Conference, 2012 IEEE*. IEEE. 2012, pp. 1–13.

[102] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion". In: *NeuroImage* 33.1 (2006), pp. 115–126.

[103] David MJ Tax, Martijn Van Breukelen, Robert PW Duin, and Josef Kittler. "Combining multiple classifiers by averaging or by multiplying?" In: *Pattern recognition* 33.9 (2000), pp. 1475–1485.

[104] A. A. Othman, H. R. Tizhoosh, and F. Khalvati. "EFIS x2014;Evolving Fuzzy Image Segmentation". In: *IEEE Transactions on Fuzzy Systems* 22.1 (Feb. 2014), pp. 72–82.

[105] A Othman, HR Tizhoosh, and F Khalvati. "Self-Configuring and Evolving Fuzzy Image Thresholding". In: *arXiv preprint arXiv:1509.04664* (2015).

[106] Einar Heiberg, L Wigstrom, Marcus Carlsson, AF Bolger, and M Karlsson. "Time resolved three-dimensional automated segmentation of the left ventricle". In: *Computers in Cardiology, 2005*. IEEE. 2005, pp. 599–602.

[107] *CMR42*. [software]. Circle Cardiovascular Imaging Inc. URL: https://www.circlecvi.com/.

[108] M Justin S Zaman, Julie Sanders, Angela Crook, Gene Feder, Martin Shipley, Adam Timmis, and Harry J Hemingway. "Cardiothoracic ratio within the'normal'range independently predicts mortality in patients undergoing coronary angiography". In: *Heart* (2006).

[109] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[110] Daniel P Huttenlocher, Gregory Klanderman, William J Rucklidge, et al. "Comparing images using the Hausdorff distance". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15.9 (1993), pp. 850–863.

[111] Martin J Bland and Douglas G Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The lancet* 327.8476 (1986), pp. 307–310.

[112] Barry J Maron. "Cardiology patient pages. Hypertrophic cardiomyopathy." In: *Circulation* 106.19 (2002), pp. 2419–2421.

[113] Christiaan F Mooij, Cornelis J de Wit, Dionne A Graham, Andrew J Powell, and Tal Geva. "Reproducibility of MRI measurements of right ventricular size and function in patients with normal and dilated ventricles". In: *Journal of Magnetic Resonance Imaging* 28.1 (2008), pp. 67–73.

[114] Michel Marie Deza and Elena Deza. *Encyclopedia of distances, 2nd ed.* Springer, 2013.

[115] Herng-Hua Chang, Audrey H. Zhuang, Daniel J. Valentino, and Woei-Chyn Chu. "Performance measure characterization for evaluating neuroimage segmentation algorithms". In: *NeuroImage* 47.1 (2009), pp. 122–135. URL: http://www.sciencedirect.com/science/article/pii/S1053811909003279.

[116] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC Press, 2000.

[117] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[118] William R Crum, Oscar Camara, and Derek LG Hill. "Generalized overlap measures for evaluation and validation in medical image analysis". In: *Medical Imaging, IEEE Transactions on* 25.11 (2006), pp. 1451–1461.

[119] Didier Dubois and Henri Prade. *Fundamentals of fuzzy sets*. Vol. 7. Springer Science & Business Media, 2012.

[120] Peter Brass. "On the nonexistence of Hausdorff-like metrics for fuzzy sets". In: *Pattern Recognition Letters* 23.1-3 (2002), pp. 39–43. URL: http://www.sciencedirect.com/science/article/pii/S0167865501001179.

[121] Punam K. Saha, Felix W. Wehrli, and Bryon R. Gomberg. "Fuzzy Distance Transform: Theory, Algorithms, and Applications". In: *Computer Vision and Image Understanding* 86.3 (2002), pp. 171–190. URL: http://www.sciencedirect.com/science/article/pii/S1077314202909744.

[122] Audun Jøsang. "A logic for uncertain probabilities". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9.03 (2001), pp. 279–311.

[123] Audun Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016.

[124] Ujjwal Karn. *An Intuitive Explanation of Convolutional Neural Networks*. 2016. URL: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/.

[125] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.