Learning Sonographic Views using a Multiple Proposal Approach



Khair Razlan Othman

Wolfson College

University of Oxford

Supervised by

Professor J. Alison Noble and Dr. Mohammad Yaqub

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019



"Oo My Lord, increase me in knowledge"

To my beloved

Othman Ahmad & Rahmah Masbor
Tengku Zunida Tengku Yahya Shah
Safiyyah Khairah
Uswah Khairah
Tasneem Khairah
Faqih al-Khair

Acknowledgements

Alhamdulillah, all praise to the Lord Almighty, the Most Merciful, the Most Beneficent, Who has taught a man how to read and write, and Who has given me Guidance all the way through this DPhil project.

A tribute should be dedicated to a group of wonderful people who had thrown in massive effort and who had strived hard to ensure this project was memorable. This thesis would not have been what it is without their contribution.

First and foremost, I would like to express my utmost gratitude to my supervisor Prof. J. Alison Noble. Thank you Alison for being there through thick and thin. All this while, you have been like a mother figure to me. You are the epitome of a limitless source of enthusiasm, encouragement and perseverance. Throughout many periods of self-doubt and difficulties, I could not recall any single event where I failed to smile upon leaving your office. "Khair, at this stage, you did pretty well!" That's only one classic example out of many. I could not have imagined having a better advisor and mentor for my DPhil study.

Special thanks to Dr Mohammad Yaqub for his relentless support and immense knowledge, and for being a guiding force behind the research, which is paramount to the completion of this project. I would also like to thank my friends and colleagues in the BioMedIA laboratory here at IBME, Oxford, Dr Carlos Arteta, Dr Ana Namburete, Dr Ali Maraci, Hasmila Omar and Joao Domingos. Their kindness and assistance have carried me through these past few years. I am also hugely indebted to Prof Stephen Kennedy, Dr Aris Papageorghiou and the medical staffs of the INTERGROWTH-21st research centre for their support and the help in medical data collection.

My biggest gratitude also goes to various friends across this vibrant academic city with whom I have had the good fortune of working and socialising with: uncle Moznu, uncle Rauf, brother Kamruz Zaman, the Bangladeshi Mosque community as a whole, and special mention to all members of Ikatan Muslimin Malaysia of European chapter. I am also grateful to the funding received towards my DPhil through Yayasan Telekom Malaysia (YTM), Wolfson College and Yousef Jameel-Oxford Centre for Islamic Studies (OXCIS) scholarship.

Finally, as customarily done, I dedicate the last few lines of the acknowledgement page to my family members: my parents, my wife and my children. Saying thank you is surely an understatement to repay their endless support and love throughout my DPhil journey. Special appreciation to my dad, Othman and my mother, Rahmah who had taught me the value of constantly seeking knowledge. My heartfelt gratitude is dedicated to my beloved wife, Tengku Zunida and my lovely children, Safiyyah, Uswah, Tasneem and Faqih, for their unconditional love and endless support during the challenging times that I faced of going through graduate school and life.

For those who had directly or indirectly contributed to make this thesis a dream come true, a big thank you to you all as well.

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Khair Razlan Othman, Wolfson College.

Copyright © 2019
Khair Razlan Othman
All Rights Reserved

Abstract

Accurate fetal biometric measurement and diagnosis rely on the quality of acquisition of 2D ultrasound (US) diagnostic planes. Optimal imaging output mostly depends on the ability of the sonographer to identify and localise a set of key anatomical landmarks which normally exhibit certain geometrical relationship patterns. Localisation of structures in US can be challenging because of the variability of image appearance and pose with respect to the transducer. In this doctoral thesis, a multiple proposal approach is introduced to facilitate an automatic localisation of multiple anatomical structures. Inspired by the image acquisition process, the knowledge of the sonographer is modelled implicitly in a multiple-stage hypothesis learning framework based on a database of 2D fetal US images acquired from multiple arbitrary orientations. Information learned from each stage allows the model to incorporate geometric constraints in a structured manner which further helps to reduce the number of possible valid hypotheses due to the variety of viewpoints and articulations. In return, the multiple proposal framework ensures good recall and precise localisation of key anatomical structures on unconstrained images of arbitrary viewing angles.

There are a couple of key research contributions brought forward by this doctoral thesis. Firstly, a comprehensive dataset with annotated ground truth for respective anatomical structures in each image is explicitly described. This dataset includes images from multiple machines and sonographers' expertise, thereby guaranteeing a collection of challenging images of various shapes, textures and layout configurations. An efficient coarse classifier based on integral channel features is then designed to provide automatic abdomen localisation. This first-stage proposal investigates a simple yet elegant implementation of a soft-cascade rejection threshold architecture with score recalibration and shows its impending practicality for US image object detection at potentially real-time speed. This improves object recall at a much-reduced time complexity. Next, an automated segmentation-driven detection technique based on probabilistic superpixels-based Bag-of-Words (BOW) is presented. It is shown that this model, which is largely inspired by perceptual grouping of low level and mid-level image descriptors, is a good alternative to a conventional sliding-window mechanism in localising multiple key anatomical structures, such as the stomach and the umbilical vein (UV). This is achieved by reducing searching time and complexity compared to multi-scale sliding mechanism, and by jointly learning and modelling the spatial relationship of multi-class objects at training and testing time. Using this approach, the model transforms a pixel-wise probabilistic output into meaningful objects based on the knowledge of neighbouring superpixel regions, thus refining the location of these key structures. This method improves the average precision of the object detection (specifically for structures of deformable nature like the UV) by localising the exact object region as compared to the bounding boxes generated from the sliding-window approach. Overall, this doctoral thesis has sought to investigate computer vision and machine learning techniques augmented by the knowledge of sonography for fetal US image analysis interpretation.

Contents

List of Figures **v**

List of Tables **viii**

1 Introduction **1**

1.1 Motivation 1

1.2 Aims and Scope 3

1.3 Key Challenges 3

1.4 Overview of Our Detection Framework 6

1.5 Outline of the Dissertation 9

2 Literature Review **11**

2.1 Clinical Background 12

2.1.1 Fetal Biometric Measurements 12

2.1.2 Objective Image-Scoring System for Automated Quality Assessment 14

2.2 Three R's of Computer Vision 16

2.2.1 Recognition algorithm 17

2.2.2 Re-organisation algorithm 22

2.2.3 Reconstruction algorithm 24

2.3 Automated Fetal Ultrasound Image Analysis 26

2.3.1 Pattern Recognition and Image Processing Methods 26

2.3.2 Learning-Based Methods for Fetal Biometrics Analysis 28

2.3.3 Summary 32

3 Datasets and Evaluation **37**

3.1 INTERGROWTH-21st Study 38

3.1.1 Abdominal Localisation Dataset 42
3.1.1.1 Annotation Methodology 42
3.1.1.2 Testing Dataset 46
3.1.2 Segmentation Dataset 47
3.1.2.1 Annotation Methodology 47
3.2 Evaluation Methodology 48
3.2.1 Classification versus Localisation 49
3.2.1.1 Evaluation Measure of the Classification Performance 50
3.2.1.2 Evaluation Measure of the Localisation Performance . 52
3.2.2 Evaluating Segmentation 54
3.3 Summary 55

4 Ultrasound Object Localisation using Aggregate Channel Features 56
4.1 Introduction 58
4.2 Overall Architecture 61
4.2.1 Phase 1 – Data-driven Learning Process 61
4.2.1.1 Overview of Feature Sets 63
4.2.1.2 Boosting Decision Tree 67
4.2.1.3 Soft-Cascade Threshold 69
4.2.2 Phase 2 – Localisation Process 72
4.3 Experimental Setup 73
4.3.1 *K*-Fold Experiments 74
4.3.2 Model Parameterisation 74
4.3.3 Overall System Architecture 76
4.3.3.1 Key Statistics Training Dataset 76
4.3.3.2 Full Image Localisation 77
4.4 Results and Discussion 81
4.4.1 Design Parameters 81
4.4.1.1 Windows and Padding Size 81
4.4.1.2 Feature Parameterisation 83
4.4.1.3 Classifier Settings 88
4.4.2 Calibration and Soft-Cascades Architecture 91
4.4.3 Scaling and Striding Effects 92
4.4.4 Evaluation of Abdominal Localisation Framework 94
4.4.4.1 Quantitative Analysis 94

4.4.4.2 Qualitative Analysis 97
4.5 Summary 99

5 Segmentation-Driven Detection of Multiple Key Anatomical Structures 101
5.1 Introduction 102
5.2 Proposed Algorithm 106
5.2.1 Multiclass Pixel-Based Detector 106
5.2.2 Learning to Combine Mid-Level Cues for Object Localisation . . 110
5.2.2.1 Superpixelisation 112
5.2.2.2 Probabilistic Superpixels 115
5.2.2.3 Visual Tokens - Deriving a Texton Map 116
5.2.2.4 Object Detection Framework via Refined Superpixels . 121
5.2.3 Post-processing 125
5.3 Experimental Setup 128
5.3.1 Datasets 129
5.3.2 Evaluation 129
5.4 Results and Discussion 129
5.4.1 Model parameterisation 130
5.4.1.1 Superpixelisation parameters 130
5.4.1.2 Choice of Region Grouping Cues 130
5.4.1.3 Choice of Seeding Cues 133
5.4.2 Quantitative Analysis 139
5.4.3 Qualitative Analysis 141
5.4.3.1 Post-processing vs None 143
5.4.3.2 Abdomen Detection (Module B) Employment 143
5.5 Summary 147

6 Conclusions and Future Work 149
6.1 Summary of Contributions 150
6.1.1 Clinical Dataset 150
6.1.2 Coarse Recognition Module 150
6.1.3 A Fine Bottom-Up Re-Organisation Framework 151
6.1.4 Multiple Proposal Framework 152
6.2 Supplementary Future Work 153
6.2.1 Joint Detection and Pose Estimation Framework 153

6.2.2 Deep Learning Algorithm 158
6.2.3 Generalisation to Other Imaging Plane 159
6.2.4 Clinical Evaluation on 3D Ultrasound Dataset 159

Bibliography **161**

List of Figures

1.1 Standard Plane Detection and Verification. 4
1.2 Key challenges to automatic detection algorithm in fetal ultrasound. . 5
1.3 Overview of the proposed method based on the possible scenario in a multiple structural localisation problem. 7

2.1 Standard fetal biometrics measurements used in clinical practice in the second trimester. 12
2.2 Overview of the main research area relevant to this doctoral thesis. . . 16

3.1 Dataset preparation from the Intergrowth-21st data repository. 39
3.2 Statistics of the datasets according to the gestational age. 41
3.3 Summary of the main abdominal dataset according to gestational age. 41
3.4 Sample preparation for the training-validation dataset. 43
3.5 Sample patches from the training-validation dataset. 46
3.6 Sample of full-frame images from the testing dataset. 47
3.7 Segmentation dataset. 48
3.8 Evaluation framework for classification and localisation tasks. 49
3.9 ROC vs. DET: A comparison of different performance illustration for similar experiment. 51
3.10 Evaluation methodology for measuring localisation performance 53

4.1 Problem formulation for coarse object localisation framework. 59
4.2 Overall object detection architecture. 62
4.3 Visualisation of channel features samples. 64
4.4 Feature extraction from local-phase information. 66
4.5 Sample of boosting decision tree with soft-threshold rejection. 70
4.6 Illustration of Window Size Parameters. 75

4.7 Histogram distributions of size ratios for various pair of structures in the training-validation dataset. 77

4.8 Fast feature pyramids scheme consists of feature channel scaling and approximation. 78

4.9 Evaluation of several principal designed parameters. 82

4.10 Feature heat map of the best 1024-weak classifier model consisting of grayscale, HOG and magnitude gradient features. 85

4.11 Evaluation of channel features properties. 87

4.12 Evaluation of pre and post-smoothing effect on the performance of channel features. 88

4.13 Comparison of different number of weak classifiers in the boosting decision tree. 89

4.14 Evaluation of the impact of various calibration parameters ϕ and soft-threshold rejection values, θ^* for different number of weak classifies over computational time. 90

4.15 Evaluation of execution time (sec/frame) for various calibration parameters, ϕ and soft-threshold rejection values, θ^* on a single-core machine. 92

4.16 Evaluation of spatial and striding scale effect on full frame image. . . . 93

4.17 Illustration of Top-Scoring Detection versus Top-Overlapped Detection. 95

4.18 Quantitative assessment of abdominal localisation framework on the test set. 96

4.19 Qualitative analysis of test set. 98

5.1 Problem formulation for fine localisation of multi-class structures via segmentation-driven framework 103

5.2 General overview of the processes involved in fine localisation of the stomach and the UV. 105

5.3 Training module for pixel-wise classification using the Random Forests (RF) framework 107

5.4 Illustration of feature sets used for Random Forest training session . . 107

5.5 Node training in the Random Forests framework 109

5.6 Visual comparison of superpixels produced by various combinations of sampling size and compactness. 113

5.7 The effect of sampling size and compactness parameters in the initialisation of key structures localisation. 114

5.8 Sample illustration of mean probabilistic superpixel map for the respective classes. 115

5.9 Flow diagram illustrating an example process by which a visual codebook may be constructed. 116

5.10 Graphical illustration of our image textonisation approach and sample patches of cluster centres. 119
5.11 Illustration of texton maps. 120
5.12 Superpixel Optimisation Framework. 121
5.13 Post-processing formulation. 127
5.14 Evaluation of the effect of various sampling size S and compactness parameters m, in the key structures localisation. 131
5.15 Evaluation of several parameters of various features use for generating object region. 132
5.16 Comparison of various perceptual cues combination method. 134
5.17 Qualitative comparison of various perceptual cues combination method. 135
5.18 Comparison of object localisation performance given background vs foreground probabilities as seed grouping criteria. 137
5.19 Qualitative analysis on background vs foreground probabilities as seed grouping criteria. 138
5.20 Quantitative analysis of multiple key structure localisation on a test set. 140
5.21 Qualitative Analysis I - Illustration of rectification of false positive cases into true positive detections using the proposed post-processing approach142
5.22 Qualitative Analysis II - Illustration of improvements on true positive detections using the proposed post-processing approach. 144
5.23 Qualitative Analysis III - Illustration of true positive detections by combining abdomen detection (Module B). 145
5.24 Qualitative Analysis IV - Illustration of false detections upon the implementation of the abdomen detection (Module B). 146

6.1 Coarse-to-fine multiple proposal framework. 155
6.2 Joint detection and pose estimation module. 156
6.3 A comprehensive unifying multiple proposal framework. 157

List of Tables

2.1 Criteria for score-based objective evaluation 13
2.2 Historical listings of the work done in fetal ultrasound detection (1) . . 33
2.3 Historical listings of the works done in fetal ultrasound detection (2) . 34
2.4 Historical listings of the work done in fetal ultrasound detection (3) . . 35
2.5 Historical listings of the work done in fetal ultrasound detection (2) . . 36

3.1 Summary of the abdominal dataset and its extended sets. 44
3.2 Summary of the chapter, showing key objectives and where the dataset is explained and used in the thesis. 55

4.1 Summary of the 5-fold cross-validation dataset 74
4.2 Time taken (sec) and average *MR* performance comparison for various parameter setting. 92
4.3 Average time taken (sec/frame) and Average Precision (AP) performance comparison of various scaling and striding options for cross-validation set 93
4.4 Selected parameters of our 'Proposal B' coarse localisation framework to detect an abdomen 100

1

Introduction

1.1 Motivation

Efficient detection of anatomical structures is a fundamental task in medical image analysis and often serves as a prerequisite for the higher level analysis and understanding of medical images, *e.g.* segmentation, detection, and registration. As an example, and considered here in this dissertation, reliable landmark detection (for instance the stomach and the umbilical vein) in a fetal abdominal ultrasound (US) scan provides the essential groundwork for successful and accurate biometry measurement.

2D US biometry plays a critical role in the practice of modern obstetrics. It has been extensively used as a tool to predict the gestational age of the fetus, estimate fetal weight, and to identify growth patterns and abnormalities (Dudley & Chapman 2002, Hadlock et al. 1985, Neilson 2008). In the United Kingdom, the National Health Service (NHS) recommends two routine antenatal screening for pregnancy (NHS Fetal Anomaly Screening Programme (NHS FASP) 2015). The first scan, also known as a 'dating scan', usually takes place between 11^{+2} weeks to 14^{+1} weeks in which the GA of the fetus is estimated by measuring the crown rump length (CRL) of the fetus.

Following that, a subsequent fetal anomaly scan is recommended to be undertaken between 18^{+0} weeks to 20^{+6} weeks of pregnancy to identify any abnormalities.

In current clinical practice, fetal size is estimated by using 2D US measurements of the head, the abdomen, and the femur, at around $18-20$ weeks gestational age (Neilson 2008, The American Institute of Ultrasound in Medicine (AIUM) 2010). Fetal growth at any stage during pregnancy can be assessed by using population-based growth charts (Hadlock et al. 1985, Papageorghiou et al. 2014). Given how important fetal growth indices are, Dudley & Chapman (2002) argued that inaccurate measurements at any stage may lead to inappropriate or untimely intervention and perinatal compromise. In view of that, Salomon et al. (2006) proposed an objective image-scoring method to assess the quality of the scan plane which suggested among others that the presence or absence of certain anatomical landmarks is a major indicator for defining a good scan plane.

This score-based method has demonstrated some measure of objectivity into image-quality evaluation and has inspired good inter- and intra-reviewer reproducibility. Objective evaluation is also likely to improve clinical training and practice, as image-scoring based on whether quality criteria are being met or not, could provide vital feedback mechanism for better coaching and skill development.

However, the current manual visual-scoring method has some limitations: 1) the quality of the evaluation is user-dependent and, 2) clinical scoring can be time-consuming and is an exhaustive process. Hence, the principal motivation of our work is to study and explore the criteria used in visual-scoring which could assist in the design of automatic plane detection and view description. Based on the selected machine learning techniques, the design framework aims for an automatic analysis that produces an accurate and reproducible fetal ultrasound biometry scan.

1.2 Aims and Scope

By automatically learning the variability of structural appearances from a fetal abdominal scan, we aim to localise anatomical structures, and quantify its importance towards the objective assessment of an ultrasound scan. In doing this, the proposed techniques aim to satisfy the following requirements:

- **Accuracy:** The localisation accuracy of the proposed multiple proposal approach should be at least as good as state-of-the-art published methods, or ideally better. This will be validated qualitatively and quantitatively against manually annotated ultrasound images.
- **Repeatability:** The object detection should be operator and machine independent, and measured objectively.
- **Automatic:** No or very little user intervention should be required in localising the structures of interest.
- **Fast and efficient:** The framework should be realisable as a real-time implementation.

1.3 Key Challenges

We focus our doctoral research on automatic detection and characterisation of abdominal ultrasound scans. Figure 1.1 illustrates the standard quality assessment framework which highlights how the proposed objective criteria could stimulate a better practice which in turn provides an improvement in the quality of ultrasound scans. It shows how crucial the localisation of particular structures *e.g.* stomach and umbilical vein are in determining the correct standardised image plane from where the measurement will be taken. Clearly, any poorly localised view will result in an erroneous measurement which further risks potential failure to detect an abnormality in growth or creates unnecessary maternal anxiety and medical intervention due to

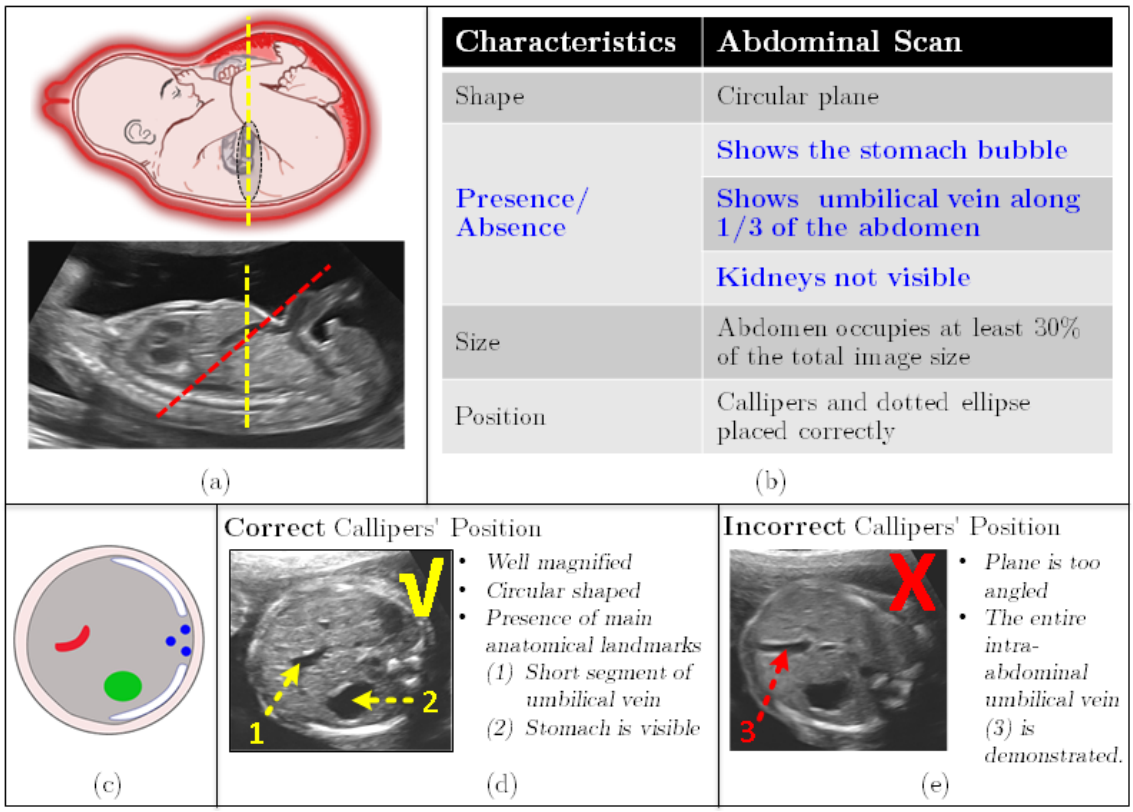


Figure 1.1: Standard Plane Detection and Verification. (a)- The intersection of the callipers is placed through the fetal abdomen for correct measurement as indicated by the yellow line in the ultrasound image. (b)- The objective scoring system criteria for abdominal circumference (AC) ultrasound scan as suggested by Salomon et al. (2006). Our research puts the emphasis on the anatomical localisation problem as highlighted in blue. (c)- Illustration of the ideal location of stomach, umbilical vein, spine and rib (represented with coloured shape in *green*, *red*, *blue* and *white* respectively) within the abdominal circumference. (d) and (e)- Illustration of the effect of callipers placement from (a). These demonstrates the interrelationships between the callipers' position (*i.e.* image plane acquisition) and the quality of image from where the biometry is taken.

erroneous detection. In this section, we provide some insights into the challenges of designing a learning algorithm or computing the image features as we aim to detect and localise the stomach and the umbilical vein. A visual description of the challenges is further illustrated in Figure 1.2.

Image quality: US images can have acoustic shadowing, reverberation artefacts, attenuation and speckle. Presence of these artefacts can make detecting and visualizing the anatomical structures difficult.

Measurement standardisation: The findings and information gained from the routine obstetric ultrasound scan will only be helpful if it is performed in a consistent and reproducible fashion. However, in practice, biometry relies heavily on the

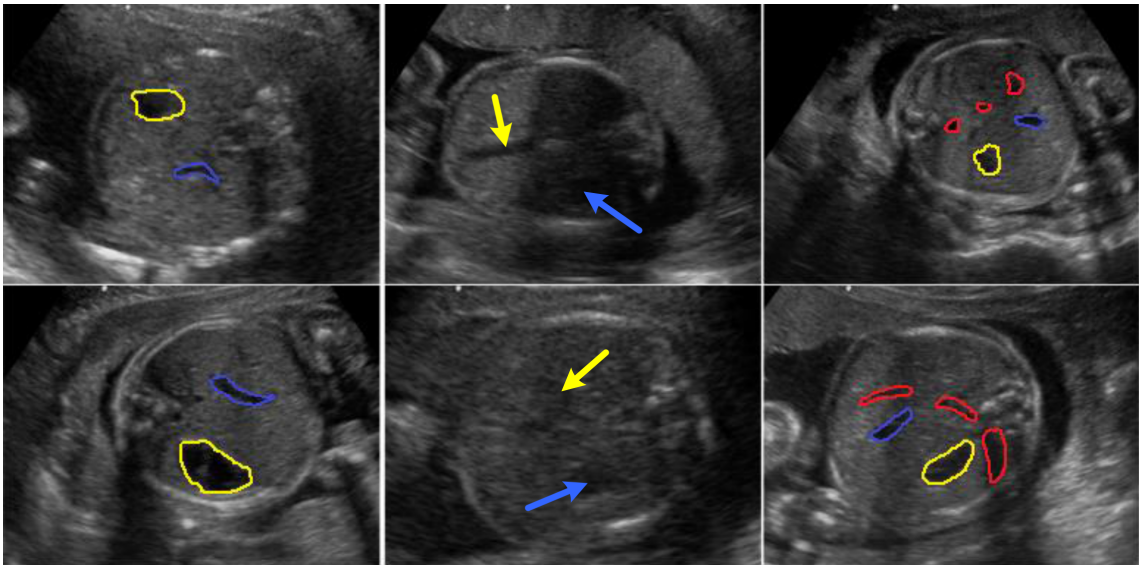


Figure 1.2: Key challenges to automatic detection algorithm in fetal ultrasound. Examples of the fetal abdominal ultrasound images which shows different key challenges to automatic detection algorithm. Fetal abdominal scans at 23 gestation weeks showing stomach (in yellow) and umbilical vein (in blue) of different size, shape and orientation (left), poor visibility due to image quality *i.e.* shadowing and contrast (middle) and similar looking structure (in red) as compared to both structures (right).

subjective judgement of the operator and is, therefore, prone to an undesirable level of intra- and inter-observer variation (Ville 2008). It is reported in (Sarris et al. 2012) that variability in fetal biometry performed by a trained sonographer is found to correlate with increasing gestational age. The level of sonographer expertise and experience in acquiring and analysing the scan also has significant impact on the data variability as has been reported in (Sarris et al. 2011). A constantly moving target, *i.e.* the fetus, also adds technical difficulty to the examination.

Data Variation: As a direct result of the two factors mentioned earlier, there can be a large variation in terms of shape, size, orientation, and position of the structures across gestational age. The fact that the measurements are taken with regard to arbitrary position of the fetus make this even more challenging. Added to that, the existence of multiple structure-alike objects which appear similar in appearance to the ones we attempt to localise imposes a greater demand on the learning process as well as to automate the recognition of untested images.

Biological property: The abdominal circumference contour has a weaker image contrast compared to both the head and the femur (*i.e.* the other two structures which are measured in a biometry exam) as it is formed by skin rather than a skull or bone, which has a higher acoustic reflection. It also lacks regularity due to deformation under the pressure from other maternal organs and the fetus itself (Yu et al. 2008b).

In summary, the US image quality assessment is important to ensure good biometry estimation. In this thesis, we look into some of the techniques and concepts from computer vision that could help us achieve our goal to produce an automated image quality assessment of obstetric ultrasound images, inspired by clinical US quality assessment.

1.4 Overview of Our Detection Framework

Figure 1.3 summarises the multiple-stage proposals approach that we have considered in this doctoral research. The unified multiple proposals framework which we have developed in this doctoral project aims to instantiate true hypotheses for the stomach (*green*) and the umbilical vein (*yellow*) which indirectly characterise the definition of standard plane detection. In other words, we seek an answer to "*How does a sonographer navigate over the fetus to locate and characterise the best abdominal image view given the potential whereabouts of the key landmarks structures?*"

In this thesis, we propose a unifying multiple proposals framework which describes an operator-independent algorithm for standard diagnostic plane identification and verification. Inspired by the acquisition process, we implicitly model the knowledge of the sonographer into a multiple-stage hypothesis framework by learning from an ultrasound image database that contain a vast collection of objects (anatomical structures) of varying appearance, shape and orientation.

This method performs a coarse-to-fine strategy interrogation on a 2D fetal ultrasound image in order to identify, extract and select respective key anatomical landmarks. Automated analysis performed on a set of selected variables contained

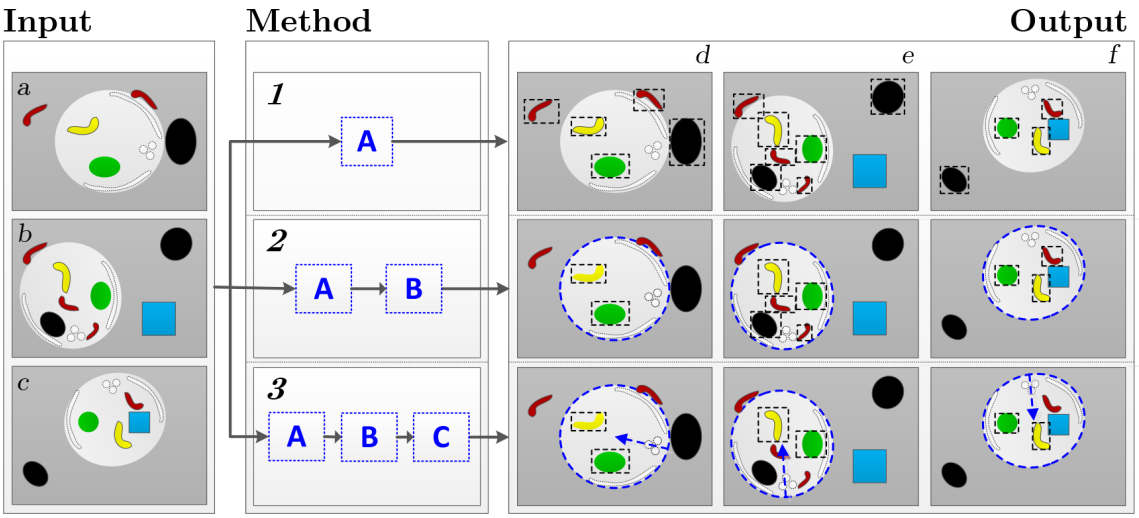


Figure 1.3: Overview of the proposed method based on the possible scenario in a multiple structural localisation problem. ***Input*** column shows three sample inputs denoted as *a*, *b* and *c*. ***Output*** column *d, e* and *f* are the respected output of image a, b and c using Method 1, 2 and 3 respectively. Meanwhile, several challenges exist. A non-object structure (in blue) which might possess notable dissimilarities in shape or pattern *etc.*can be considered as an easy true negative case. However, multiple structure-alike objects as exemplified by possible resemblance of the stomach (in black) and the UV (in red) that appear within or outside the abdomen may lead to false detections. Dotted box(es) shows possible detection(s).

within the 2D view help to establish geometrical context amongst structures of interest. This equips the system with an ability to eliminate multiple localisation problems (as exemplified by the output of *Method 1* and *2*). In particular, we investigate the importance of the interaction between two out of three major research themes popularly known as **recognition**, **re-organisation** and **reconstruction** [1] within the computer vision community - in providing a unifying solution to this medical image problem as illustrated in Figure 1.3.

Re-organisation: The dominant approach to object detection has been based on the sliding-window mechanism. It has been particularly made popular by the introduction of the integral image and cascaded architecture classifier (Viola & Jones 2001) as well as the histogram of oriented gradient (HOG)-based pedestrian detection (Dalal & Triggs 2005). This approach requires a system to independently model a dedicated classifier for each object. The learned classifiers are then used to evaluate a rectangular window of different sizes, locations and aspect ratios across the image to

[1]The categorisation of research works within the computer vision field into these three major themes is coined by Malik et al., in his recent works entitled 'The Three R's of Computer Vision: Recognition, Reconstruction and Reorganisation' (Malik et al. 2016).

search for possible candidate object(s). Given dedicated classifier for each structure, this search strategy can be considered as an exhaustive and time-consuming process. Besides, a fixed-size ratio classifier imposes certain limitations. It suits a rigid object better than an articulated object such as an umbilical vein. This technique also risks classifying the same patch as belonging to multiple objects.

Hence, we consider a pixel-wise object localisation technique within our first proposal (*Module A*) to localise the presence of the stomach and the umbilical vein (UV). This bottom-up *detection-by-segmentation* approach involves a classification process based on grouping and segmentation techniques which are strongly linked to 'perceptual organisation' in human vision or what is termed as ***re-organisation*** (Malik et al. 2016). Independent of the abdominal presence, we designed a multi-class pixel-based detector and superpixel-based Bag-of-Words framework to localise the key structures in a bottom-up fashion. Details of *Module A* proposal framework will be discussed in Chapter 5.

Recognition: Identification of standard diagnostic planes is itself an example of a recognition task. Recognition can be defined as a task to attach semantic category labels to objects or scenes (Malik et al. 2016). Specifically, locating the presence of an abdomen is crucial and our research suggests that finding it might be needed to be done prior to the process of extracting the position of the key landmarks. As highlighted in the first row of Figure 1.3, there is a possibility that the first proposal (*Module A*) ends up with some similar-looking structures outside the abdominal area. This scenario leads to the design of our second proposal (*Module B*) which coarsely and quickly scans for the presence and the location of an abdomen. Our recognition module is presented in Chapter 4 where we also explore various aspects of machine learning architecture and compare the performance of several image features.

Reconstruction: The original author, Malik et al. (2016) pointed out some three-dimensional (3D) works to define the reconstruction term. In the context of our work here, reconstruction would include estimation of pose and spatial layout of key anatomical landmarks. We can hypothesise that the pose of the abdomen plays a part

in pointing out the correct pair of the stomach and the UV. Although estimating the pose of the abdomen might improve the detection of the best diagnostic abdominal plane, it is out of the scope of this thesis. Method 3 which incorporates reconstruction process (*Module C*) has not been implemented in the thesis but progress towards this is described in Section 6.2.1 as part of our future work.

In summary, in this thesis, we present our 2-stage multiple proposal framework (Method 2) which demonstrate the interaction of the *recognition* and *re-organisation* processes operating in tandem with each other, either to provide input or to exploit output of each other. Method 2, which is a combination of *Module A* and *Module B* is expected to eliminate some falsely identified structures (mostly stomach) which sit outside the abdominal circumference.

1.5 Outline of the Dissertation

The aim of the doctoral thesis is to characterise the challenges faced in the localisation of key anatomical structure(s) in fetal abdominal US images and to propose solutions for suitable automated detection techniques. The following is the outline of the doctoral thesis.

- **Chapter 1** introduces the thesis and describes the rationale behind the study. Subsequent chapters are then summarised.
- **Chapter 2** describes the clinical background to this research, the role of objective quality metrics in defining planes, and provides insight on the challenges in designing the learning algorithm or computing the visual image features for US images. It also reviews the state of the art in computer vision and machine learning community, which are relevant to this thesis.
- **Chapter 3** summarises different datasets used in this thesis for various experiments. We also present different formulations of the evaluation criteria for the given datasets, which are used in later chapters.

- **Chapter 4** provides an experimental evaluation of a coarse patch-based detection method for localising the abdominal circumference. This includes a discussion on the integral channel features employed and the boosting decision tree framework that made up our second proposal module (*Module B*).

- **Chapter 5** considers a fine pixel-wise localisation technique based on Random Forests and probabilistic superpixel-based Bag-of-Words technique for localising the stomach and the umbilical vein. This includes details on the feature sets used for both techniques, including a post-processing framework using superpixel optimisation techniques.

- **Chapter 6** summarises our contributions and suggests several directions for future work.

2

Literature Review

In this chapter we review the existing relevant literatures that serve as background for the work presented in this doctoral thesis. First, we review relevant clinical background related to fetal biometry with an emphasis on objective image-scoring based evaluation for quality assessment. Secondly, we provide a general summary on the central problems in the machine learning and computer vision literature which are *recognition*, *reorganisation* and *reconstruction*. Here, we discuss the related techniques and feature sets used for each topic and most significantly, on the importance of the interaction between these different elements in producing a unifying framework for image analysis.

Finally, we focus and highlight recent published techniques in medical image analysis that were influential to our work, specifically in the area of automated fetal biometry measurement from ultrasound images.

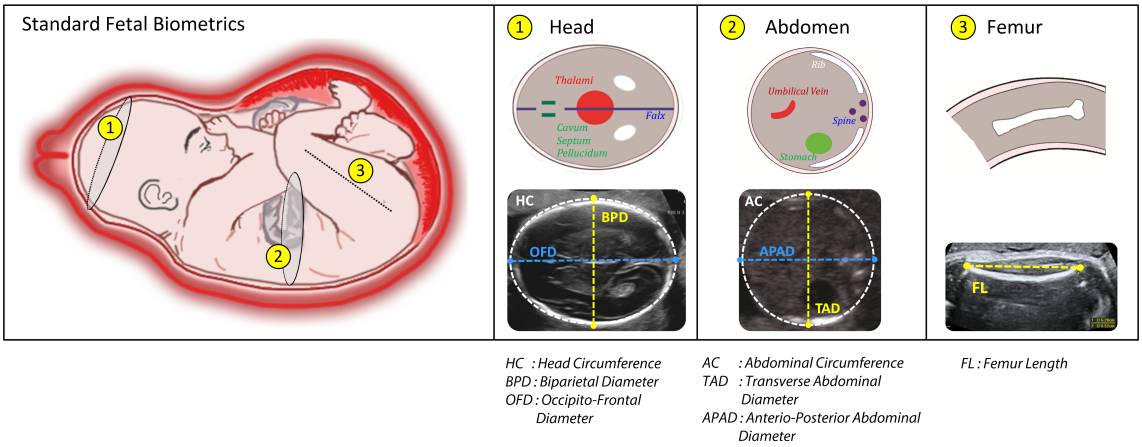


Figure 2.1: Standard fetal biometrics measurements used in clinical practice. Measurements were taken from two-dimensional (2D) US measurements of the head, abdomen, and femur were taken throughout pregnancy to estimate the fetal growth.

2.1 Clinical Background

2.1.1 Fetal Biometric Measurements

Two-dimensional (2D) US biometrics plays a critical role in the practice of modern obstetrics. It is used globally as a tool to predict the gestational age of the fetus, to estimate fetal weight, and to identify growth patterns and abnormalities (Blum & Kurtz 1990, Campbell & Wilkin 1975, Hadlock et al. 1982, The American Institute of Ultrasound in Medicine (AIUM) 2010). In current clinical practice, fetal size is estimated by using 2D US measurements of the head, the abdomen, and the femur, at 18—20 GA weeks (Dudley & Chapman 2002, Neilson 2008, The American Institute of Ultrasound in Medicine (AIUM) 2010). Henceforth, fetal growth at any stage during pregnancy can be assessed using population-based growth charts of head, abdomen and femur measurements (Chitty et al. 1994a;b;c, Hadlock et al. 1985, Melamed et al. 2009).

As depicted in Figure 2.1, fetal weight can be estimated based on a combination of several standardised parameters derived from a 2D ultrasound scan which include the *Head Circumference* (HC), the *Abdominal Circumference* (AC) and the *Femur Length* (FL). Campbell & Wilkin (1975) claimed that the AC provides the best correlation with fetal growth and it has been shown to be the most reliable indicator for birth weight

Table 2.1: Criteria for score-based objective evaluation according to Salomon et al. (2006). For every single structures, each fulfilled criterion scores one point.

Type of Image		
Cephalic	*Abdominal*	*Femoral*
Symmetrical plane	Symmetrical plane	Both ends of the bone clearly visible
Head plane occupying more than half of the total image size	Abdominal plane occupying more than half of the total image size	Femoral plane occupying more than half of the total image size
Plane showing the thalami	Plane showing the stomach bubble	45 angle to the horizontal
Plane showing the cavum septi pellucidi	Plane showing the portal sinus	-
Cerebellum not visible	Kidneys not visible	-
Calipers and dotted ellipse placed correctly	Calipers and dotted ellipse placed correctly	Calipers placed correctly

estimation. This has been strongly supported by the study of Nahum & Stanislaw (2003). Later models such as the one introduced by Hadlock et al. (1985) incorporate other fetal indices in an attempt to provide a more accurate estimation.

Chang et al. (1992) and Bernstein et al. (1997) have suggested that the AC is the most sensitive parameter for the detection of growth abnormalities. The fact that the AC shows a unique profile growth has been hypothesised. The degree of abdominal fatty accumulation is claimed to contribute more towards weight estimation as compared to the bone growth at later gestational age (Yu et al. 2008b). Melamed et al. (2009) also mentioned the difficulty in obtaining the diagnostic measurement plane for head measurements when the head engages in the pelvis during later gestation. These evidences suggest that the brain development and bone growth might have less correlation with the fetal weight increment compared to the AC. Thus, the AC measurement provides some interesting insights to be explored.

Dudley & Chapman (2002) raised the very basic yet most important question, which forms the premise of our research. They argued that inaccurate measurements at any stage may lead to inappropriate or untimely intervention and perinatal compromise. They postulated three main factors that contribute to the accuracy of measurements and their interpretations which are: (i) machine caliper accuracy; (ii)

validity of measurement charts for local methods and populations; and (iii) the quality of image plane acquisition and measurement. In this thesis, we aim to further examine the third factor with the AC as our parameter of interest.

2.1.2 Objective Image-Scoring System for Automated Quality Assessment

An essential aspect of any clinical examination of fetal biometry is the explicit definition of sonographic image landmarks and caliper placement. Modern obstetric practice requires the sonographer to continuously and actively engage with the real-time 2D sonography to find the correct imaging plane. During the entire examination, the sonographer must be involved in *simultaneous* acquisition, display and interpretation of images. These include image visualisation of the screen, manipulation of the transducer, adequate hand-eye coordination, in-depth knowledge and recognition of fetal anatomy, mental reconstruction of the 2D spatial relationship of several key landmarks as well as constant adaptation to dynamic changes such as fetal movement.

That is to say that the responsibility of acquiring the image in the correct plane rests entirely on the experience and subjective evaluation of the sonographer. In that sense, any inaccurate view will result in an erroneous measurement which further risks a failure to detect abnormality in growth or creates unnecessary maternal anxiety and medical intervention due to erroneous detection. Dudley & Chapman (2002) reported that in some cases, poor results are not due to inattention to detail; but apparently due to misinterpretation of the required landmark. Ville (2008) claimed that given a sonographic measurement are based on subjective judgement, it is prone to an undesirable level of interobserver and intraobserver variation, hence needed an automated quality assessment scheme. As 2D sonography is a highly operator-dependent process, the accuracy and reproducibility of the measurements between examiners is reported to achieve fair to moderate performance (Nakai & Oya 2002,

Sarris et al. 2012). Sonographic images of the same fetus acquired by two different operators may vary in quality, size, appearance, and orientation, all of which clearly can influence diagnostic interpretation.

Various image-scoring methods have been proposed to promote **objective quality assessment** of ultrasound images. These include the research work of Herman et al. (1998), Sairam et al. (2009) and Salomon et al. (2006; 2008). Herman et al. (1998) instigated an objective scoring tool for nuchal translucency measurement, which is used to detect fetal aneuploidy risk such as Down Syndrome in the first trimester screening. Meanwhile, Sairam et al. (2009) have made an attempt to assess the quality of fetal cardiac images at the time of screening and concluded that the scoring method could be further developed into a cardiac certification system.

In the field of fetal biometry, series of extensive work done by Salomon *et al.* (Salomon et al. 2006; 2009, Salomon & Ville 2009, Salomon et al. 2008) have demonstrated that manual quality assurance based on objective image-scoring method of certain metrics (refer Table 2.1) is feasible to obtain reproducible results. The standardisation process using the proposed metrics has been implemented in the INTERGROWTH-21st multicentre study (Papageorghiou et al. 2013, Sarris et al. 2013) to ensure an accurate and reproducible measurement of fetal biometrics. This standardised method has been reported to minimise systematic errors associated with pooling data from different study sites, thus leading to improvement in the accuracy and consistency of measurements (Sarris et al. 2011).

To date, subjective evaluation of predefined metrics is used in daily practice when reviewing an ultrasound examination performed by another operator. Nevertheless, the manual examination procedure may be lengthy even when reviewers are trained and experienced (Salomon et al. 2006). In contrast, an automated image measurement and assessment may introduce a much faster and a more systematic process. Meticulous standardisation and quality control of the measurement based on the objective image-scoring criteria listed in Table 2.1 ensures that all fetal biometric parameters are taken in an identical fashion. By doing this, it is promoting consistency

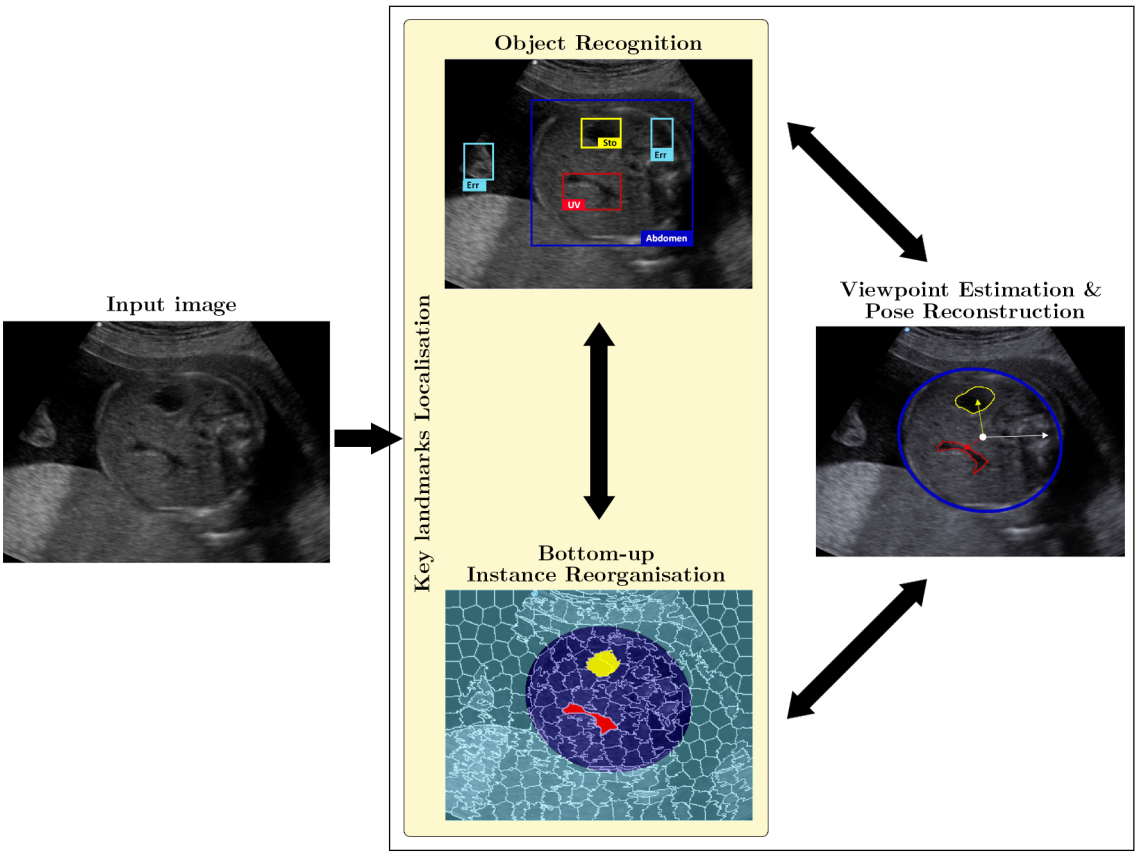


Figure 2.2: Overview of the main research area relevant to this doctoral thesis. Three main categories of research literature: *recognition*, *reorganisation* and *reconstruction* are identified as a basis for this doctoral project and its future extension work. First, we briefly review the evolution and the recent trend within a computer vision and machine learning community with regard to these topics. Later, we bring together related medical image literatures in particular one which focus on fetal biometrics for further discussion and comparison.

in the measurements taken by eliminating operator bias and reducing inter- and intra-operator variability.

2.2 Three R's of Computer Vision

In this section, we review selected work related to automated image processing and analysis using computer vision and machine learning frameworks. Basic terminologies, concepts and techniques that address three main challenges in creating a unifying framework for automated detection and characterisation of an abdominal ultrasound scan: ***recognition***, ***re-organisation*** and ***reconstruction*** will be reviewed. These closely relate to the research work presented in Chapter 4 - Chapter 5.

Even though the tasks mentioned here are quite distinct in the traditional computer vision literature, they tend to blend together while crafting the main unifying solution. Hence, we simplify and standardise the concepts and terminologies to distinctively group the cited literatures based on the dedicated objective of each task as exemplified in Figure 2.2.

As for key landmarks localisation, the dominant approach to object detection based on sliding windows could straightforwardly be applied to detect and recognise an abdomen (*recognition*). Alternatively, we opt for a bottom-up pixel-wise segmentation and grouping approach (*re-organisation*) in order to be able to jointly learn and localise the other two key landmarks; the stomach and the UV. Even so, these two approaches might be built via a common pipeline, which mostly will include a mixed combination of selected **image feature sets** and a **detection/classification algorithm**. The information discovered via these first two steps are later to be proposed to craft the third-stage classifier which tunes to estimate a pose viewing angle by recovering the geometrical relationship of key points implicitly (*reconstruction*). Each task is vital in creating a unifying framework solution aimed to guide and simplify the final verification of the biometric plane. This will be described in more depth next.

2.2.1 Recognition algorithm

Recognition can be defined as the task of attaching semantic category labels to objects or scenes (Malik et al. 2016). Recent years have seen great advances in the area of image-based object recognition using machine learning. The two major sub-tasks of object recognition are **image classification** and **object localisation**. Image classification deals with the task of defining the presence or absence of an object category without providing much information on the actual object location. *Object localisation* is more natural, in which an instance and potentially sub-parts are detected with its estimated location, in many cases up to a bounding box. The most common approach for object recognition is to perform an exhaustive search across all locations

and scales, and formulate object detection as a classification task, *e.g.* "*Does an image window at location i contains a given object*?"

Techniques based on this sliding-window mechanism have been popularised by the introduction of the integral image and cascaded architecture classifier (Viola & Jones 2001) as well as the histogram of oriented gradient (HOG)-based pedestrian detection (Dalal & Triggs 2005). It has been successfully used to detect rigid objects such as faces (Papageorgiou et al. 1998, Viola & Jones 2001; 2004), as well as articulated objects such as human pedestrian detection (Dalal & Triggs 2005, Papageorgiou et al. 1998). Alternatively, Lampert et al. (2009) used a branch-and-bound strategy to effectively search possible subwindows and hence reduce the computational cost of evaluating a large set of candidate subwindows.

Later on, Alexe et al. (2012) introduced the concept of '*objectness*' with the aim to reduce the number of possible windows being considered for a classification. The objectness measure which was derived from human attention theories (Desimone & Duncan 1995, Teuber 1955) suggested that the human vision system processes and analyses only certain parts of an image in detail, while leaving other part nearly unprocessed. In accordance to this theory, the approach suggests that some windows are more likely to represent an object, even without analysing its internal pattern. They argued that a well-defined closed boundaries and centres of an object are naturally the strongest distinctive characteristics that differentiate an object from its surrounding. Several works on this area (Cheng et al. 2014, Zitnick & Dollár 2014) have further confirmed this. In summary, by limiting the searching space to a small number of possible candidates, literatures based on the objectness approach has claimed to greatly increase the computational efficiency as well as to reduce the number of false-positive detections.

Regardless of the advancement, which contribute to create a more efficient and higher precision system, most approaches revolve around two main themes; 1) ***efficient image descriptors*** and 2) ***machine learning techniques***.

Image Features: Feature extraction generally captures intensity patterns, texture details, and/or shape and contour information. The introduction of Haar-like features (Papageorgiou et al. 1998, Viola & Jones 2001) has proven to be a big success . Haar features encode differences in intensity between different regions. It can be extended to multiple sizes, geometry and a combination thereof as well as to 3D (Cui et al. 2007). The Scale Invariant Feature Transform (SIFT) is another efficient feature descriptor that encodes gradient information in scale-space and possesses invariant properties to changes in scale, illumination and local affine distortion (Lowe 1999).

Dalal & Triggs (2005) proposed a novel feature called the Histogram of Oriented Gradient (HOG) descriptor. The HOG descriptor characterises the distribution of normalised local intensity gradients and edge directions within a small (overlapping) spatial region and has been experimentally shown to uphold invariance to geometric and photometric transformations. Emulating the success of bag of words in text retrieval, Sivic & Zisserman (2003) introduced bag-of-features which is essentially a quantised codebook vector that represents images by frequencies of 'visual words' *i.e.* features. Although demonstrating impressive classification performance, bag-of-features methods disregard information about the spatial layout of the features. The 'spatial pyramid' concept of Lazebnik et al. (2006) addresses this limitation. This approach works by repeatedly subdividing an image and computing histograms of image features over the resulting subregions.

Using a slightly different approach, Shotton et al. (2008) introduced the semantic texton forest (STF) which operates directly on the raw pixels value rather than a region or subregions. The STF can explicitly and/or implicitly model the classifier by building up a bag of semantic texton codebook within a learning model (*i.e.* Decision Forests). This bypasses the expensive steps of computing image descriptors while also incorporating the implicit hierarchical structure in a quick and efficient way.

Arguing that most computer vision literature shows increasing evidence in favour of features with richer and overcomplete representations, Dollár et al. (2009) proposed the 'Integral Channel Features' (ICF), with the idea of registering multiple

image channels (*i.e.* features such as intensity, gradient *etc.*) computed via linear and/or non-linear transformations of the input image. This idea was later extended to an 'Aggregated Channel Feature' (ACF) (Dollár et al. 2014), which was based on direct feature computation from a pixel value, rather than integral images as used in the earlier approach. It also introduced a feature approximation technique over a finely-sampled image pyramid which has been claimed to demonstrate significant speedups at reduced computational cost.

For other interesting local features, we refer readers to the work of Tuytelaars & Mikolajczyk (2008), for an extensive discussion and evaluation. We now proceed to briefly discuss some well-known classification techniques that are used to evaluate the extracted features.

Classification Techniques: The overall object detection architecture can be divided into two phases. The training phase creates a binary classifier that provides decisions (*i.e.* probability) on whether the image regions (or pixels) belong to the class or not *e.g.* background/foreground decision; while the testing phase uses the classifier to perform a dense scan on an unseen test image. Each image window will be encoded into a feature space and then fed into the classifier in which the likelihood score of each image windows corresponding to the respective object will be reported. There are various classifiers which can be used for this purpose, among them are Adaptive Boosting (AdaBoost), and Random Forests; both which belong to the class of ensemble methods.

AdaBoost (Freund & Schapire 1997) gained wide attention, thanks to the Viola and Jones detector, which offered a fast and practical solution to face detection (Viola & Jones 2004). AdaBoost demonstrates an effective feature selection where at each stage of the boosting process, it selects a new weak classifier that minimises the error rate and consequently '*tweaks*' in favour of those instances misclassified by previous classifiers. A single strong classifier is derived from the family of weighted weak classifiers which is then used for object classification. Nonetheless, despite its simplicity and high performance achievement as reported in various application, such

as face detection (Viola & Jones 2004) and pedestrian detection (Guo et al. 2014), the relatively slow training stage renders them impractical for applications with real-time requirements (Appel et al. 2013).

Compared to boosting, Random Forests technique has been demonstrated to achieve comparable accuracy, while being faster (Caruana & Niculescu-Mizil 2006, Lempitsky et al. 2009). The randomised decision forest (Breiman 2001), was originally conceived as a method of combining several decision trees (Breiman et al. 1984) using bagging technique (Breiman 1996). Its early development was influenced by the random subspace method (Tin Kam 1995; 1998), the feature selection exhibits a more robust method with respect to noise, addresses high dimensional data and also provides good generalisation. This is more likely due to its randomised decision criteria and feature selection process. In principal, a random decision tree selects a random feature that divides the training images at a node. Recursively, a complete tree and subsequently a collection of trees are built, which constitute a random forest. Final predictions are obtained by aggregating over the ensemble. Random Forests is now regularly used in medical imaging research (Donner et al. 2011, Lempitsky et al. 2009, Namburete & Noble 2013, Yaqub et al. 2011). On top of that, Random Forests is intrinsically suited for multi-class problems with each tree can be trained and tested independently of each other. In return, this had made it possible to parallelise the operation, and hence make it highly useful for real-time applications.

As reviewed, boosting (Schapire 1990) and its variant has been popularly known for its simplicity; in combining many weak learners to form a single strong one, while Random Forests (RF) or decision tree (Breiman 2001, Breiman et al. 1984) is typically acknowledged for its speed and efficiency in training and testing. Boosting decision trees (Grossmann 2004, Tu 2005) is another popular classification technique which reconciles the powerful combination of boosting and decision trees. It has shown a phenomenal achievement behind many state-of the-art methods across multiple domains such as text information retrieval (Peng et al. 2010), medical image analysis (Carneiro et al. 2008b) and object recognition (Dollár et al. 2010). The Boosting

Decision Tree approach of Appel et al. (2013) also has been reported to guarantee an equivalent performance to classical AdaBoost at a much reduced computational cost, both in training and testing.

Being fast and possessing a great noise suppression quality is great features for ultrasound imaging modality. Hence, we employed a shallow boosting decision tree with channel features given our motivation behind the work for recognition algorithm (*Module B*) is to produce a module closes to a real-time running speed. This will be presented in Chapter 4.

2.2.2 Re-organisation algorithm

As mentioned in the previous section, the dominant approach to object localisation has been based on a sliding-window classifier framework. However, this technique is somewhat unsatisfactory as its final result *i.e.* bounding-box output often simplifies the presence of the object as it does not capture the true shape of the object especially for deformable objects. In a multi-class object detection problem, the technique risks classifying the same pixels of the same image as belonging to multiple objects (Pantofaru et al. 2006). The fixed-ratio local patches are also reported to fail when local image information within the window is insufficient *e.g.* the target is occluded or too small (Murphy et al. 2006). An individual classifier is trained for each category object to be recognised. Given that the size, aspect ratio and location of the objects are initially unknown, the implementation of a sliding window mechanism especially for multi-class objects is highly computational and hence costly to evaluate.

In the spirit of overcoming this limitation, several works have been proposed on pixel-wise object class localisation. Such a pixel-labelling, or ***semantic segmentation*** approach aims at annotating each pixel in an image with a class label from a predefined set (*i.e.* grass, tree, road, and person). Objects and background classes are segmented into discrete non-overlapping regions and a label provided for each region, which indirectly produce an even more refined view of an object localisation.

Pantofaru et al. (2006) performed pixel-level object localisation by using a combined region and patches approach. This technique was found to do well on deformable objects with repetitive textures. Leibe et al. (2008) considered object categorisation and figure-ground segmentation as two interleaved processes, in which their Implicit Shape Model (ISM) yield a pixel-wise segmentation with confidence estimate as a result and extension of recognition. Marszatek & Schmid (2007) argued that the method produced by Leibe et al. (2008) suffers from parameterised hypotheses of the Hough space voting. They proposed online shape masks that are learned from sparse local features computation; enrich additional information of possible object locations and implicitly handle multiple viewpoints and articulations, resulting in precise localisations. In comparison to the work of Marszatek & Schmid (2007), Fulkerson et al. (2008) suggested that a better performance can be achieved by only performing local classification at the pixel level. The latter constructed compact dictionaries of dense features using integral images of spatial occurrences of features (Fulkerson et al. 2009).

Shotton et al. (2008) introduced an extremely fast Semantic Texton Forests (STF) which classifies pixels with reduced and simpler features. The STF method which learns a multi-class classifiers based on the RF framework has demonstrated a promising result for object localisation, in which one can label image pixels with the corresponding classes instead of rough localisation indicated by a bounding box. However, in the case of an ultrasound image as considered here, a decision made based on observation of individual pixels alone can be less meaningful and susceptible to noise. In short, there is a clear need to establish the relationship between pixels and the need to re-organise image information into meaningful assemblies *e.g.* region.

The use of bottom-up segmentation approach to provide pixel-wise class localisation has recently gained interest. Pantofaru et al. (2008) form superpixel-like objects by intersecting multiple segmentations and then obtain classification by averaging the results from all the member regions, while Fulkerson et al. (2009) has suggested a conditional random field (CRF) that operated directly on superpixels.

Superpixel algorithms group pixels into perceptually meaningful atomic region. The recently introduced simple linear iterative clustering (SLIC) by Achanta et al. (2012) has demonstrated a quick and memory-efficient superpixel algorithm, which also includes a more regular lattice representation. Schick et al. (2012) confirmed several properties of superpixels, such as aligns well with object boundaries, robust to noise than pixels and serves well to represent objects in the image. They also coined the term *probabilistic superpixel*, which is the proportion of foreground to background pixels within a superpixel. The work of Rantalankila et al. (2014) also demonstrate that object proposals generated via superpixels grouping has been able to produce fast and accurate object recognition.

In particular, we are interested to extend this approach, in which the probabilistic superpixels are formed from the posterior probabilities output of the RF multiclass classification output of the STF methods. Instead of deriving the binary segmentation output of foreground-background segmentation and producing a pixel-wise level solution, our segmentation-driven detection model proposes to reorganise pixels (or regions) based on mid-level information of a neighbouring superpixel regions to solve the anatomical localisation task.

2.2.3 Reconstruction algorithm

As humans, we can easily perceive and interpret the 2D scene or even the 3D shape of the shown object, even though it might be the first time we have seen that particular instance. For example, while seeking for the correct diagnostic plane, a sonographer would actively manipulate the US transducer and move it to certain positions based on the limited information displayed in the viewing field. Skill and expertise are needed to swiftly reconstruct 3D mental visualisation over successive 2D images obtained in real-time. At the 2D level, the operator will assess the presence of a particular fetal anatomy relative to a certain known configuration. Hence, identification of the pose of the object, which indirectly could be translated as a reconstruction of

geometrical relationship between multiple structures of interest, is mutually important as to localise the structures itself.

Some interesting projects has been conducted related to multi-view object prediction or pose estimation problem, such as in the field of vehicle detection (Juránek et al. 2015), face recognition (Yang et al. 2014a, Zhu & Ramanan 2012), and generic object detection (Savarese & Fei-Fei 2007). Generally, methods for pose estimation and multi-view object detection can be categorised into two major groups. The first group depend solely on just 2D information to solve the pose estimation task (Ghodrati et al. 2014, Ozuysal et al. 2009, Redondo Cabrera et al. 2014), while others require a complex 3D model (Liebelt & Schmid 2010, Savarese & Fei-Fei 2007).

Conceptual wise, there are two approaches in which a pose of the object can be described. From a more *global* standpoint, one may estimate the object pose using its **viewpoint** while a more *local* approach is via localisation and identification of a set of fixed **keypoints**. Tulsiani & Malik (2015) pointed out that viewpoint prediction is a general manifestation of a 2D image which can be represented by the spatial relationships of the different features of the object. Most of the works involve part based detectors with the Deformable Part-based Model (DPM) (Felzenszwalb et al. 2010) as its core. However they are computationally expensive, taking over several seconds per image. Other works include a separate DPM component representing a particular viewpoint (Xiang et al. 2014) or a fixed global descriptor to estimate viewpoint (Ghodrati et al. 2014). Meanwhile, keypoint prediction methods have mainly been used to recognise an object's keypoints given a bounding box as input. This includes model-based approach (Hogg 1983) to a pictorial-based approach (Yang & Ramanan 2011).

The work that interest us most are perhaps the one which jointly tackled the problem of object detection and pose estimation. While there are methods which trained several view-dependent detectors (Ozuysal et al. 2009, Savarese & Fei-Fei 2007) or treat the problem as a 1-vs-all classification detector (Ghodrati et al. 2014), we seek a simple approach, which can perform close to real-time speed. Juránek

et al. (2015) proposed a single compact model based on boosted decision trees, which claims to provide a pose estimation alongside a classification score at the speed of real-time operation. Hence, we consider this model to be part of the extension to our work, as a similar classification architecture could be used for both detection and pose estimation as well as feature extraction for a small increase in computational cost at test time. This is discussed further in Chapter 6.

2.3 Automated Fetal Ultrasound Image Analysis

In this section, we review and summarise some key relevant prior works done in the area of image-based analysis for anatomical structure recognition in ultrasound image. We observe and compare the roles of diverse image processing and pattern recognition techniques in facilitating the development of US image analysis. Then, we shift our attention to the domain of machine learning for the US image analysis. In particular, we focus and highlight recent published techniques in US medical image analysis that were influential to our work, specifically in the area of automated fetal biometry measurement.

2.3.1 Pattern Recognition and Image Processing Methods

Morphological Operator: Most of the very early works in fetal US image analysis have tackled the problem of automatic diagnostic measurement via a combination of thresholding approach and generic morphological operators, specifically to detect the edge and segment out the boundary of the femur (Thomas et al. 1991) and the head (Hanna & Youssef 1997, Lu et al. 2005). This is probably due to the low quality of digitised US signal at that time, and the fact that the bony structures such as skull and femur appear as bright objects, making them relatively easy to detect. Thomas et al. (1991) and Hanna & Youssef (1997) for example, employed a morphological operator such as opening and closing operation, edge linking *etc.*on a binary output of a threshold algorithm. Though, a hard decision based on an intensity thresholding

technique is sensitive to noise and might not always guarantee a good performance given a dynamic range of intensity levels across ultrasound images. Weak edges and large gaps that exist on the abdominal boundary often became a challenge to classical edge linking techniques such as described in morphological operations.

Active Contour Model: Instead of exploiting only low-level image information such as the previous edge-detection techniques do, some research apply an optimisation procedure using an information about the boundaries in a technique called active contours, or popularly known as snakes (Kass et al. 1988). Chalana et al. (1996) and Pathak et al. (1997) reported a segmentation of the fetal head using an active contour model. These semi-automatic approaches assumed that the head boundary is smooth and unbroken and further required the user to initialise the area of interest before iteratively fitting a contour around the boundary.

A novel method based on the integration of several segmentation techniques, which include the Gradient Vector Field (GVF) snakes and fuzzy C-means algorithm was developed to extract boundary information of the abdomen (Yu et al. 2008b). The implementation of GVF snakes is also reported in the work that study the detection of intrauterine growth retardation using fetal abdominal circumference (Nithya & Madheswaran 2009). Both works (Nithya & Madheswaran 2009, Yu et al. 2008b) similarly apply Hough transform technique to iteratively provide initial detection of the respective structures before refining the edges using the GVF. An active contour module however is sensitive to initial contour placement and might require intensive pre-processing to eliminate noise from irrelevant structures.

Other Methods: Ciurte et al. (2012) proposed an automated fetal abdomen segmentation using a semi-supervised graph-based patches. Limitation of the method includes problems related to the label initialisation. Other approaches include prior work on parametric deformable models (Jardim & Figueiredo 2005) as well as work that based on fuzzy connectedness framework (Rueda et al. 2015, Smith & Arabshahi 1996). We refer the reader to a recent published review on automatic techniques for the US imagery by Rawat et al. (2018).

2.3.2 Learning-Based Methods for Fetal Biometrics Analysis

Research on learning-based techniques or machine learning as it is popularly known have been encouraging in various domains and gained a lot of attention especially from the computer vision and pattern recognition communities. The application of this particular approach has only recently sparked much interest in medical image analysis, thanks to evolving technology trend, especially the availability of substantial medical data collection and high-performance computing.

In the field of US image analysis, research works have been dominated by several groups, in particular the work from the Siemens research group (Carneiro et al. 2008a; 2007; 2008b, Feng et al. 2009, Sofka et al. 2010, Zheng et al. 2009b), and the BioMedIA laboratory, Oxford University (Namburete et al. 2012, Noble 2010, Rahmatullah 2012, Rahmatullah et al. 2011a; 2012a;c; 2011b, Yaqub et al. 2012).

Perhaps, the first work which seriously attempts automatic detection and measurement of the abdominal structure is the work of Carneiro *et al.* (Carneiro et al. 2007; 2008b), which also includes other fetal biometry as well. They introduced the probabilistic boosting tree classifier technique that learned statistical data from manually annotated database. Their data-driven method was able to produce a robust detection and segmentation of the fetal head, abdomen and femur in US images with comparable accuracy to clinical evaluation. The detection algorithm ran from a coarse to fine stage searching the best classifier based on the appearance models. The notions of Haar features and constraining the tree growth by limiting the height and number of nodes enable their algorithm to execute in real time. However, this work did not emphasise on the fundamental challenge of finding a standardised abdominal plane.

As we aim to localise the stomach and the umbilical vein within an abdominal scan, it is also worth to highlight the recent work of Sofka et al. (2010) and Yaqub et al. (2012) which target multiple object detection in US images. Sofka et al. (2010) use sequential Monte Carlo based on Hierarchical Detection Network to model the selection order of multiple object detection, evaluated on the fetal faces and the

fetal brain structures of 3D US images. The model dynamically learns the spatial relationships between object poses in a sequential manner by recursively applying the prediction and updating steps to obtain the posterior distribution of the object. In the work of Yaqub *et al.* (Yaqub et al. 2012), they use the discriminative Random Forests classifier to localise fetal brain structures of 2D US images by learning from the 3D annotated volumes. The results provide a good insight on the application of Random Forests in US structure localisation, and the performance to relatively challenging abdominal structures is yet to be tested, which we intend to explore.

Target wise, the work of Rahmatullah *et al.* in (Rahmatullah 2012, Rahmatullah et al. 2011a; 2012a;b;c; 2011b) is the closest to our aim. This work shares a common goal, which is to localise the stomach and the umbilical vein in return to the optimal plane. It uses AdaBoost with the Haar and local phase features. In that approach, training for both structures is done separately. In the testing stage, the sliding window scans at multiple scales and positions giving out scores for both structures. The decision on the presence/absence of the structures is based on the maximum score. There is a major issue with this approach; the spatial relationship between both structures is ignored in which the detector might end up firing the maximum score on the same window, given that the stomach and the umbilical vein structures look almost similar in normalised patches. In a bid to overcome this issue, we aim to extend this work by studying the performance of Random Forests incorporated within the probabilistic superpixel framework in localising multiple key anatomical structures in the abdomen.

Following that, a series of work by Ni *et al.* (Ni et al. 2013; 2014, Yang et al. 2014b) explore the solution for localising the fetal abdominal standard plane based on selective search strategy. Ni et al. (2013) trained four individual classifier for the abdomen, stomach, spine and UV using an AdaBoost algorithm before sequentially detect the presence of each structure and decide on the standard plane. They first segment out the probable location of the stomach using an image processing and thresholding algorithm. Having a prior knowledge on the beam direction of

the US signal, they proposed an accumulative vessel probability map (AVPM) to subsequently search for the UV. They then modelled an anatomical relationship by defining the angles between the stomach, UV and the spine. This method is very much depending on the way the data were collected. Besides, as the local detectors were learned individually and separately detected, the geometrical relationship among the structures was established subsequently in sequential style.

Later, they expand their proposed solution to include a radial component model (RCM) in which they claimed to describe the geometric relationship of the standard plane (Ni et al. 2014, Yang et al. 2014b). They replaced the AdaBoost with the Random Forest classifier to train the local detector. The abdominal area was divided into 5 sub-regions to formulate the probable locations of the respective structures in the RCM. A second-level SVM classifier was then introduced to combine the individual score of each component detectors at all iterations and produced a newly formed combination score to identify the best abdominal plane. While the spirit of this method seems to be similar to our own by employing a coarse-to-fine detection mechanism, our unifying framework presents a joint object detection and pose estimation in a single compact model. We also jointly learned the probable locations of key anatomical structures using the Random Forests.

Research in medical image analysis over the last few years have witnessed a significant impact of the computer-assisted analysis in providing a better image interpretation, thanks to a revolution introduced by **deep learning** architecture. Shen et al. (2017) acknowledged that three main factors which are; (i) advancements of high-tech central processing units (CPUs) and graphics processing units (GPUs), (ii) availability of a huge amount of data (*i.e.* big data), and (iii) developments of learning algorithms are instrumental to the unprecedented success of deep learning framework.

Deep learning offers a different paradigm of image analysis than the techniques that we have explained earlier, by letting the data to discover and extract the information (*i.e.* features) in a self-taught manner. This approach takes on a different view of image analysis from ours, which still rely on hand-crafting the features based

on the knowledge of the target domain. As most of work related to deep learning algorithm appeared in the later stages of our research as reported here, we only provide a quick review on some literatures that is along this lines to put our research in context.

Recently, several works have proposed deep learning framework to standard plane detection in fetal US. Arguing that a deep learning algorithm works best given a large number of training data which is not always feasible in medical image analysis, a series of work by Chen *et al.* on multiple flavour of convolution neural network (CNN) (Chen et al. 2015a;b; 2014a;b) exploited powerful features representation learned from a large database of natural images towards the domain of fetal US plane identification.

Chen et al. (2014b) proposed a deep convolution neural network (DCNN) that learns the spatial feature representation via knowledge transfer. Experimental result shows huge improvement to their original work based on RCM (Ni et al. 2014) as reported earlier. They then expanded this model by exploring the temporal information on top of the spatial information (Chen et al. 2014a) to identify the fetal standard plane in the US video. Later, they introduced a multiple layered transferred CNN (TCNN) model (Chen et al. 2015b) which fuses the knowledge transferred from natural images with several layers of information learned from US imagery.

Likewise, Gao et al. (2016) reports preliminary work on describing and partitioning the US video has shown promising prowess of DCNN technique. CNN based framework also has demonstrated a real-time performance with a high classification accuracy in fetal standard plane detection using a freehand US videos (Baumgartner et al. 2017).

Elsewhere, it is also interesting to point out the work of Maryam *et al.* (Ahmed & Noble 2015; 2016a;b) which complement to our target although taken from a different research perspective. Her eye-tracking inspired framework provides several important insights, which could be the base guideline whilst designing the state of the art technique for automatic fetal standard plane identification. First, their experiments

demonstrate the importance of the spine besides the UV and the stomach in providing the correct plane view. Secondly, the eye movements extracted across multiple fixation points and their two-stage searching strategy suggest that the human vision employs a 'top-down' image perception in aiding the final 'reconstruction' view of what the standard plane is.

We provide a summary in Table 2.2 - Table 2.5, detailing the collection of techniques developed in the field of US image analysis, which is not only restricted to abdominal and detection problem but also to other fetal biometry as well.

2.3.3 Summary

This chapter has reviewed a collection of relevant literatures widely ranging from work on general computer vision to the application of computer vision technique in the area of medical image analysis.

We started our survey by looking at several popular approaches within a computer vision community and sorting them into three major research themes, namely recognition, re-organisation and reconstruction in order to see the current developments in this area and how the trend presses forward. Having reviewed literatures on prevalent techniques applied on various medical image modalities, we rounded up our review by comparing some closely related work on fetal abdominal ultrasound images.

Table 2.2: Historical listings of the work done in fetal ultrasound detection and/or segmentation. The table shows the following columns: reference, region of interest (ROI), algorithm *i.e.* techniques and features used, dataset and modality, the result(s), if the work is semi or fully automatic and some notes for references.

	Reference	ROI	Learning Algorithm		Dataset and Modality	Results	Semi/ Auto	Notes
			Techniques	Features				
Fetal Ultrasound	Yaqub et al. (2012)	Head (Brain)	Random Forest	Intensity (Cuboid, Unary, Binary, Haar)	Real 3D US Train - 10, Val - 20 Testing - 10	(*See reference)	Auto	2D detection from 3D classification
	Namburete et al. (2012)	Head (Chroid Plexus)	AdaBoost	Intensity, Nakagami, μ, ω	Real 2D US images Train - 570, Val - 61 Testing - 44	0.889 - Intensity 0.678 - Naka - μ 0.862 - Naka - ω	Auto	
	Rahmatullah et al. (2012c)	Abdominal	AdaBoost	Intensity (Haar) Local Phase (LP)	Real 2D US images (*See references)	0.83 -Stomach 0.63 -UV	Auto	Intensity vs. Intensity + LP
	Rahmatullah et al. (2011a)	Abdominal	AdaBoost	Intensity (Haar)	Real 3D US volumes Train/Val - 10/5 Testing - 30	91.29 -Recall 76.29 -Precision	Auto	Separate and combined training
	Yaqub et al. (2011)	Femur (Brain MRI)	Random Forest	Intensity (3D, Unary/Binary/ Rectangle/ Haar/Position/ Averaged Rectangle)	Real 3D US volumes Pre-selection - 20 Train/Test - 20/21	0.82±0.08 -Recall 0.83±0.09 -Precision	Auto	Feature pre-scoring and selection
	Rahmatullah et al. (2011b)	Abdominal	AdaBoost	Intensity (Haar, Average)	Real 2D US images Training - 2000 Testing - 2384	0.81 -Stomach 0.64-UV	Auto	Separate training

Table 2.3: Historical listings of the work done in fetal ultrasound detection and/or segmentation. The table shows the following columns: reference, region of interest (ROI), algorithm *i.e.* techniques and features used, dataset and modality, the result(s), if the work is semi or fully automatic and some notes for references.

	Reference	ROI	Learning Algorithm		Dataset and Modality	Results	Semi/ Auto	Notes
			Techniques	Features				
Fetal Ultrasound	Sofka et al. (2010)	Brain, Face (& Left atrium)	Hierarchical Detection Network & Monte Carlo	NA	Real 3D US volumes Brain: Train/Test - 589/295 Face: Train/Test - 962/48	(*See reference)	Auto	Multiple object detection
	Marhaban et al. (2010)	Head (HC & BPD)	Morphological operators, Ellipse fitting	Gradient Intensity	20 Real 2D US images	(*See reference)	Auto	Elliptical shape assumption
	Tolay et al. (2009)	Spine	Morphological operators, K-means	Orientation, shape, spatial proximity, size	Real 2D US images Testing - 16	87.5 % Accuracy	Auto	
	Feng et al. (2009)	Face	Probabilistic Boosting Tree, Marginal Space Learning	3D Haar features, Steerable features, Context Info	Real 3D US images Train - 962, Test - 48 2D profile detector Train - 500, Test - 48	(*See reference)	Auto	
	Yu et al. (2008a;b)	Multiple fetal biometry	Edge detection, Fuzzy C-means, Hough transform, Snakes	Gradient intensity (contour)	215 fetuses	(*See reference)	Semi	

Table 2.4: Historical listings of the work done in fetal ultrasound detection and/or segmentation. The table shows the following columns: reference, region of interest (ROI), algorithm *i.e.* techniques and features used, dataset and modality, the result(s), if the work is semi or fully automatic and some notes for references.

	Reference	ROI	Learning Algorithm		Dataset and Modality	Results	Semi/ Auto	Notes
			Techniques	Features				
Fetal Ultrasound	Carneiro et al. (2008b)	Head (cerebellum, cisterna magna and lateral ventricles)	Probabilistic Boosting Tree	3D Haar features, Streeable features, Context info	Real 3D US volumes Training - 200 Testing - 40	(*See reference)	Auto	Semantic query for plane standardisation
	Carneiro et al. (2008a; 2007)	Multiple fetal biometry	Constrained Probabilistic Boosting Tree	Intensity (Haar)	Real 2D US images Training: 1,293 - abdomen, 1,426 - head, 1,168 - femur. Testing: 183 - abdomen 177 - head, 171 - femur.	(*See reference)	Auto	
	Lu et al. (2005)	Head (HC & BPD	K-means & Hough transform	Intensity (brightness)	Real 2D US images Testing - 217	0.12% - BPD 0.52%- HC **Mean signed difference	Semi	

Table 2.5: Historical listings of the work done in fetal ultrasound detection and/or segmentation. The table shows the following columns: reference, region of interest (ROI), algorithm *i.e.* techniques and features used, dataset and modality, the result(s), if the work is semi or fully automatic and some notes for references.

	Reference	ROI	Learning Algorithm		Dataset and Modality	Results	Semi/ Auto	Notes
			Techniques	Features				
Fetal Ultrasound	Hanna & Youssef (1997)	Head (HC & BPD	Morphological operators & Hough transform	Gradient intensity (contour)	NA	0.994 (BPD) 0.985 (HC)	Auto	
	Chalana et al. (1996)	Head (HC & bpd)	Active contour model	Gradient intensity (contour)	35 digitised sonogram	1.4% (BPD) 2.9% (HC) ** Mean absolute difference	Semi	Smooth unbroken boundary assumption
	Thomas et al. (1991)	Femur	Morphological operators	Size, shape	24 Real ultrasound films	0.8133 (MSE)	Auto	
Others	Donner et al. (2011)	Hand	Random Forests, Markov Random Field, Mean-Shift	Haar-like 3D descriptors	8 hand's CT volumes Leave-One-Out cross validation	Mean/median/std residual: 10.13/5.59/16.99 voxels	Auto	Incorporating Subpart Interdependencies
	Criminisi et al. (2011)	Multiple anatomical structures	Regression Forest	Intensity (Haar)	100 real CT volumes Train/Test : 55/45	(*See reference)	Auto	
	Zheng et al. (2009a;b)	Liver, Left Ventricle (LV)	Marginal Space Learning	Haar	Liver : 226 of 3D CT volumes LV : 795 of 2D MRI	(*See reference)	Semi	

3

Datasets and Evaluation

Well-characterised image research datasets are an essential component of object recognition research but rather lacking in supporting current medical image analysis research. Despite the ever-increasing amount of annotated medical data for tasks such as automatic segmentation and localisation of anatomical structures (Carneiro et al. 2008a, Rueda et al. 2014), there is often a gap between research image datasets and datasets reflecting real-world clinical data. For instance, in ultrasound (US) research, US scans might be collected of the fetus where images are of a certain size, quality and orientation. It remains unclear how the prepared datasets relate to the complex images encountered in routine clinical practices. In particular, questions arise as to whether the same machine-learned model that might be derived and tested successfully on ultrasound data collected from limited orientations and of certain high-quality, will be able to perform well on much more challenging datasets, in which the images are of more variable quality and taken from arbitrary orientation, which would be more typical of routine clinical use.

This chapter summarises how the datasets are selected and used throughout the thesis. The datasets have been carefully chosen so that they cover a wide range of applications under consideration. We aimed to ensure that the selected images contain no selection bias in terms of quality, contrast, *etc* that would perturb the results. Two substantial databases were used to create annotated databases for our research. These are described later.

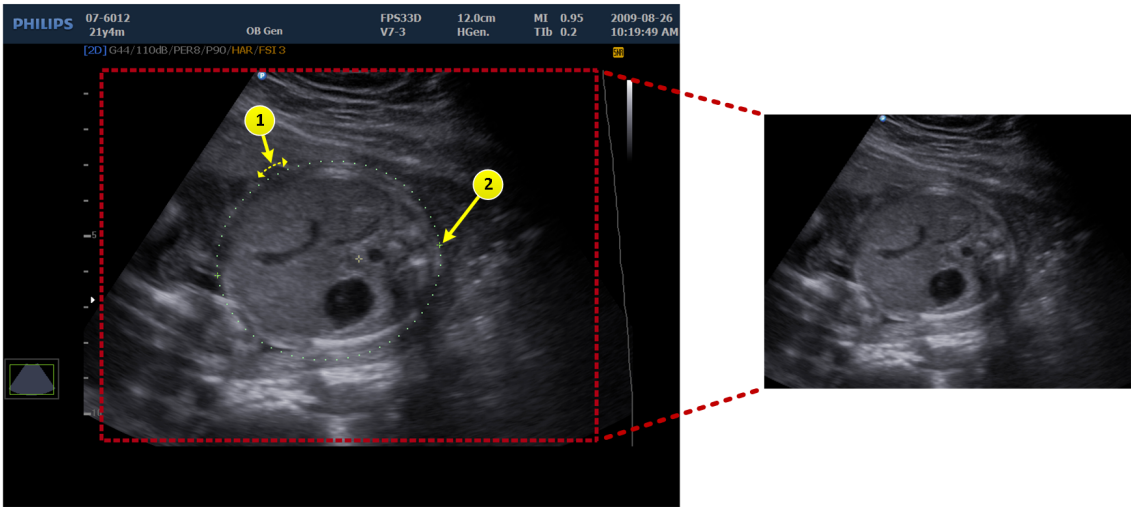
This chapter will also describe the annotation methodology and the evaluation criteria used to measure the performance of the proposed classification, localisation, and segmentation algorithms for each dataset as presented in subsequent chapters.

3.1 INTERGROWTH-21st Study

The INTERGROWTH-21st study is a multicentre, multi-ethnic, population-based project, which has developed a new prescriptive growth standard for fetal and newborn infants. The study's design and implementation of this project have been described in details in Villar et al. (2013), and major findings of the study have been reported in 3 Lancet papers in 2014 (Papageorghiou et al. 2014, Villar et al. 2014a;b). In this study, the US examinations were carefully carried out by highly skilled sonographers using a Philips HD9 US machine. US image acquisition followed a standardised protocol and was performed only on healthy pregnant women (Papageorghiou et al. 2013, Villar et al. 2013). All images were acquired using a $2-5$MHz transducer.

Standard fetal biometrics measurements were taken from three main fetal body parts, namely: the head, the abdomen and the femur. In measuring the fetal abdomen for example, there are certain criteria to be observed such as having a cross-sectional view that is close to circular, having the umbilical vein (UV) visible in the anterior third of the abdomen, having the stomach bubble (SB) visible, and having the abdomen fill at least 30% of the whole US image (Papageorghiou et al. 2013).

We obtained 860 US images of 2D fetal abdominal scans from the INTERGROWTH-



(a) (b)

Figure 3.1: Dataset preparation from the Intergrowth-21st data repository. Given the original data with image header and already marked abdomen as in (a), we removed the header by automatically cropping the DICOM images to a size of 750 × 650 pixels (as marked by the *red* rectangle). We further resized the abdominal scan into half the original size as in (b). The calliper marker served as our ground truth for the abdomen. See text for details.

21st data repository, acquired over a period of 2 years (2009 — 2011). These images were carefully selected, in which the image quality was assessed by an experienced sonographer (*T.N.*) to confirm that the quality meets the standardised quality control criteria (Sarris et al. 2013), as illustrated earlier in Figure 1.1. In particular, we reviewed and inspected the size and shape of the abdomen, as well as the presence and absence of the stomach and the umbilical vein.

As illustrated in Figure 3.1, we first resized the image to fixed image dimension ($s = 280 \times 375$ pixels). This standardises the images in the dataset to a fixed image size. No intensity normalisation was applied to the original US images. Rahmatullah (2012) highlighted that the results gained upon intensity normalisation are of poorer quality due to the fact that the speckle patterns and significant variation between US images follow a non-Gaussian distribution, making it unsuitable for normalisation.

In each image, biometric measurement was performed by manually drawing an ellipse over the abdominal structure. Thus, each abdominal US image contains an abdominal circumference estimate, which includes both ellipse points and calliper markers as highlighted in Figure 3.1 (a). We developed an ellipse fitting algorithm

and applied it to the US image to extract the circumference measurement points,which later served as our ground truth for abdomen biometry. We subsequently applied a median filter of size $(5, 5)$ *pixels* on the extracted points to remove the measurement markers. Since the measurement markers are small in size, we chose a filter of size 5×5 to ensure it covers these markers. In doing this, each detected pixel on the individual RGB layer of the image was replaced by the median value of 5×5 neighbouring pixels on the similar layer. As the measurement markers have a high pixel intensity value, a median filtering technique demonstrates an effective smoothing performance in removing the marker by replacing the intensity of these pixels with the intensity of the surrounding pixels *i.e.* , removing them while keeping consistent appearance with the surrounding pixels.

The fetal gestational age for the images ranged from 17 to 23 weeks of gestation. The distribution of the gestational age is shown in Figure 3.2. Following the initial dating scan, each participant in the study had had up to six fetal growths scans at an interval of 5 ± 1 weeks starting as early as $14^{+0\text{days}}$ weeks of the gestational age (Papageorghiou et al. 2013). The abdomen in the images may appear to be in different size and pose, or orientation, as well as of a variable appearance and contrast range, as illustrated in Figure 3.6.

We randomly divided the 860 abdominal images into 500 training and validation images and 360 independent testing images with nearly equal distribution according to their ages as illustrated in Figure 3.3. In addition, we also manually collected 500 negative images from the INTERGROWTH- 21st dataset collection. These images belong to the patients of which we collected the positive images from, but contained no abdomen part. There are two different sources which included: 1) data taken from other fetal biometrics measurement *e.g.* the head and the femur and, 2) images captured from DICOM video collection in order to represent various possible scenarios. Out of this negative dataset, 400 images were randomly selected as testing images while the remaining 100 images were used to extract negative patches.

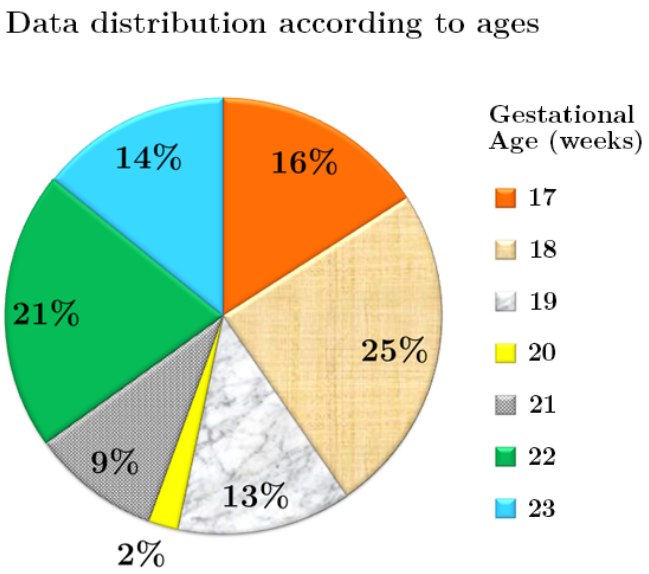


Figure 3.2: Statistics of the datasets according to the gestational ages. Distribution of abdominal images according to gestational age (weeks). (n = 860 images in total).

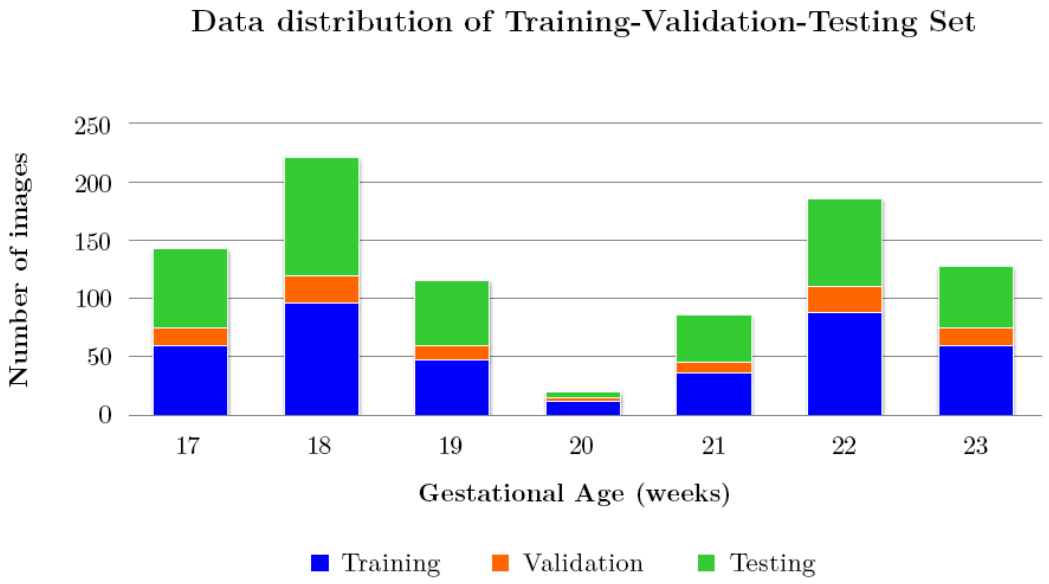


Figure 3.3: Summary of the main abdominal dataset according to gestational age. Distribution of Cross-Validation Dataset (Training and Validation data) and Testing Dataset.

Table 3.1 tabulates details of each image dataset and its extended collection. Next, the listed dataset is explained in detail.

3.1.1 Abdominal Localisation Dataset

This dataset was prepared and annotated in order to develop the classifier for a coarse patch-based localisation of the abdomen as described in Chapter 4. The following subsections describe in detail how the dataset was annotated. The performance of the algorithm evaluated on the dataset is also explained briefly.

3.1.1.1 Annotation Methodology

A common object detection pipeline based on the sliding-window framework naturally works by applying a pre-trained classifier to the subwindows of the image. This simple recipe starts by first building a dataset of labelled image windows of fixed size (say, $n \times m$). The positive examples should contain a large, centered instances of the object of interest, while those labelled negatives should not.

Upon training, the system will now pass every $n \times m$ window in the image to the classifier for an object/non-object decision. However, given that not all instances of an object will appear the same size as the classifier, there is a need to build a pyramid of the image (Lazebnik et al. 2006), so that the sliding window can scans over the image for object at multiple scales and positions. In our case, we prefer a feature approximation technique (Dollár et al. 2014) which scales the generated features rather than the image to find object of multiple scales and positions. In addition to this customary practice of following the sliding window paradigm, size normalisation is also essential to our work in order to minimise the variance among the training images available, given that the abdomen is a deformable structure of variable size and aspect ratio.

Our ground truth for abdominal localisation is based on *ellipse-shape* markers, which are extracted from the images as described earlier, and illustrated in Figure 3.4 (*first* column). It should be noted that we designated *Dataset A* to represent full-frame images from which the normalised patches (depicted as *Dataset B* and *Dataset C*) are then extracted. Subsequently, as illustrated in the second column of Figure 3.4,

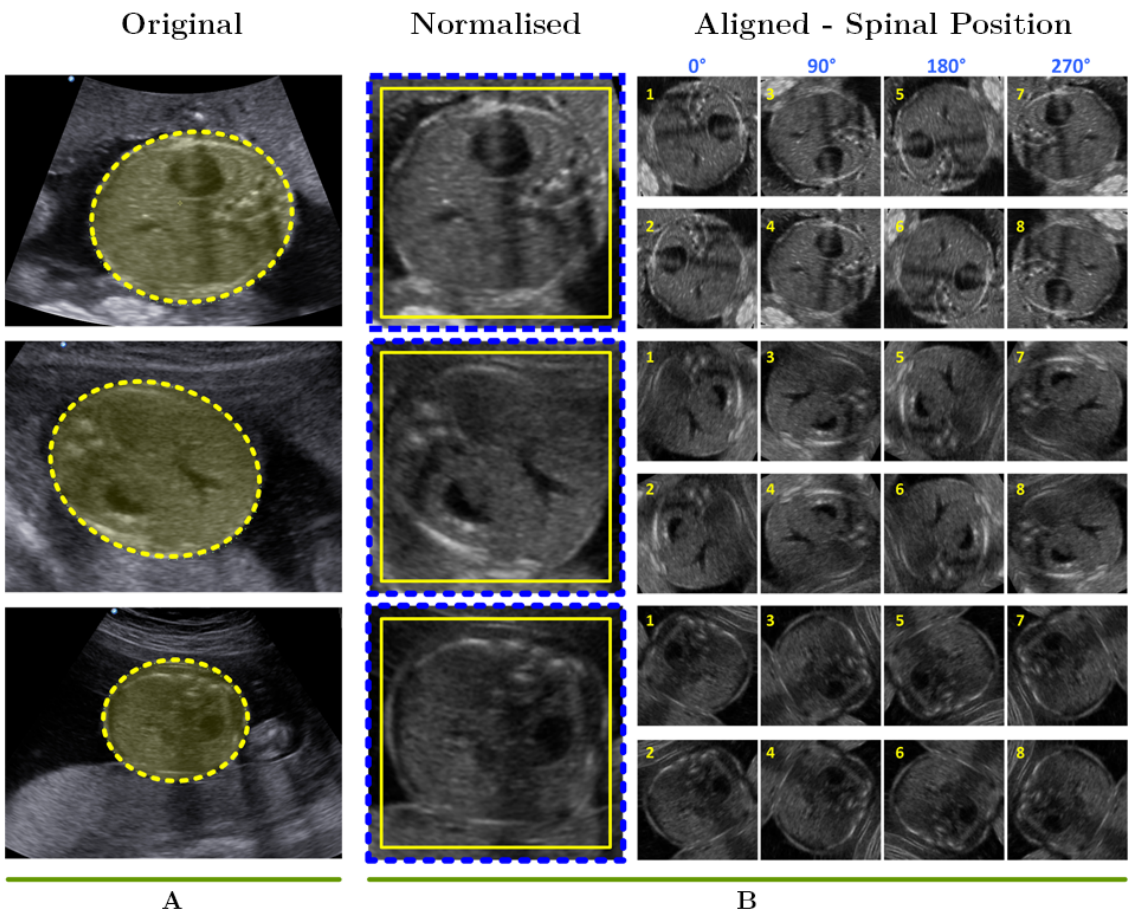


Figure 3.4: Sample preparation for the training-validation dataset. Dataset A (first column) represents full frame images of size, $s = 280 \times 375$ pixels. These are the positive images (Dataset A1) which show abdominal structure of various position, orientation, size and aspect ratio. Dataset B (second and third columns) shows some normalised positive image windows for the respective full-frame images. Note that the pose variations of these images (column 2) are later realigned with respect to the spine position (column 3). In the third column of each abdominal image, the first row (odd-numbered label) and the second row (even-numbered) are their left-right/upside-down reflections).

we normalised (by *rescale* and *crop*) each positive training image of the training-validation dataset (*Dataset A*1) to a standardised size, centred on a fixed-sized window ratio (H:W = 1:1), as shown by the yellow bounding box. We opted for a square size classifier as our observation on the abdomen aspect ratio, as illustrated in Figure 4.7(b), showing that most of our training sample is square in shape, with $\mu = 1.027$.

Further details on the effect of spatial scaling factor and aspect ratio on the object localisation will be described in Section 4.3.3.2. We also allowed some pixel padding around each side of the patch (of *w-pixels* width) to study the effect of the surrounding

Table 3.1: Summary of the abdominal dataset and its extended sets. Images (as shown in column Dataset A and D) refer to full frame size images of dimension, $s = 280 \times 375$ *pixels* while patches (Dataset B and C) refer to the number of normalised patches of canonical size (*i.e.* 120×120 *pixels*) extracted from the respective images.

	Training - Validation Dataset					**Testing Dataset**	
	A		B		C	D	
	A1	A2	B1	B2		D1	D2
	No.of positive images	No.of negative images	No.of positive patches	No.of positive patches	No.of negative patches	No.of positive images	No.of negative images
Abdominal Images	500	-	500	4,000	10,000	360	-
Non-Abdominal Images	-	100	-	-	2,500	-	400
Total	500	100	500	**4,000**	**12,500**	360	400
						760	

context and to minimise boundary effects. We called this collection of 500 positive patches *Dataset B*1.

Prior to the normalisation step for the patch generated in *Dataset B1*, we manually rotated the original images of *Dataset A*1 (if necessary), so that the respective spinal structure points to one of four corresponding reference angles (0, 90, 180 and 270 degrees). In order to represent the possibilities of the arbitrary position of the fetus, the normalised patches were then flipped (using single or multiple combination of left-right and upside-down flipping operation) to produce extended *Dataset B2*, effectively increasing the number of positive examples to 8 patches per image, as denoted by label 1 to 8 in Figure 3.4 (refer to column 3). A total of 4,000 positive patches, as depicted in *Dataset B2*, were allocated to the training-validation set.

Table 3.1 provides a summary of the two main sources of negative patches which are labelled as *Dataset C*. *First*, the negative patches were obtained from the positives images in *Dataset A1*, with a restriction to only have up to 60% overlap with the ground truth. To create these samples, we started by automatically generating 10 bounding boxes of randomly defined size on each positive image in *Dataset A1*. We

generated a random integer for the width of the bounding box, which can be any even integer from arbitrarily chosen number between 40 to 280 (*i.e.* height of the image). This is followed by a choice of starting row and column of the respective box. Both processes were executed using a uniform random number generator of the MATLAB®. Therefore, each bounding box contains information of the top-left coordinates (x, y) and the size of the extracted patch.

On top of that, we also modelled a range of poses and viewing conditions by randomly generating a pair of scale and rotation parameters for each bounding box in the list. Given an observation on the abdomen aspect ratio in Figure 4.7(b), we produced an extra set of 10 other options of possible aspect ratio on top of the default square size bounding box, $s_n \in (dx_n, dy_n)$ as tabulated in Eq. (3.1).

Random Scaling Options

n	1	2	3	4	5	6	7	8	9	10	11
dx	1.00	1.03	1.07	1.10	1.15	1.20	1.00	1.00	1.00	1.00	1.00
dy	1.00	1.00	1.00	1.00	1.00	1.00	1.03	1.07	1.10	1.15	1.20

(3.1)

For rotation parameters, the algorithm can either choose to rotate the original patch or a flipped version of it by a corresponding angle; 0° (original position), 45°, 90°, or 270°. The original patch was scaled and rotated according to the scale and rotation parameters (only if the options generated are different from original position) prior to patch extraction. We extracted and spatially standardised a total of $10,000$ negative patches from the positive images.

Using a similar procedure, we created 25 negative patches per image from an additional set of 100 negative images (*Dataset A*2). These were negative images from US scans that contain no abdomen and are of similar dimension to the original image (s = 280 x 375 *pixels*), as previously described in Section 3.1. This process resulted in an additional $2,500$ negative patches.

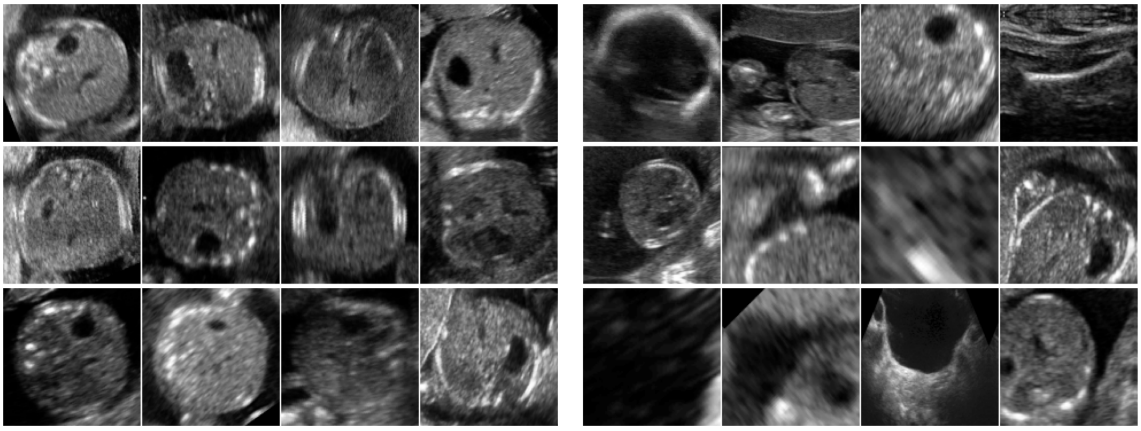


Figure 3.5: Sample patches from the training-validation dataset. Some normalised patches are from the abdominal localisation dataset. The first four columns show the positive patches in *Dataset B*2 while the remaining columns present sample of negative patches in *Dataset C*. Note that some negative patches contain abdominal images which have up to 60% overlap area to the annotated ground truth.

In summary, the training-validation dataset contains 4,000 positive patches and 12,500 negative patches as summarised in Table 3.1. It is also observed that we have an imbalance distribution of sample patches in which the total number of negative samples far exceeds the number of positive samples. As our main purpose for the abdomen detector is to achieve a *high-recall* rate, we collected more negative samples as to allow the machine learning to analyse all sort of possible variabilities of these examples.

Some typical sample patches used in our training-validation set are shown in Figure 3.5. No contrast normalisation was applied to the patches. We used K-fold cross- validation experiments for model parameterisation, with $K = 5$. Patches in *Dataset B2* and *Dataset C* were randomly partitioned into 5 equal size subsamples with equal proportions of the classes on each fold. Details of the experimental settings will be elaborated in Section 4.3.

3.1.1.2 Testing Dataset

Figure 3.6 shows some examples of both positive and negative images from the testing dataset. The abdominal location of each positive image was annotated using a similar ellipse-measurement procedure, as previously explained in Section 3.1. The negative

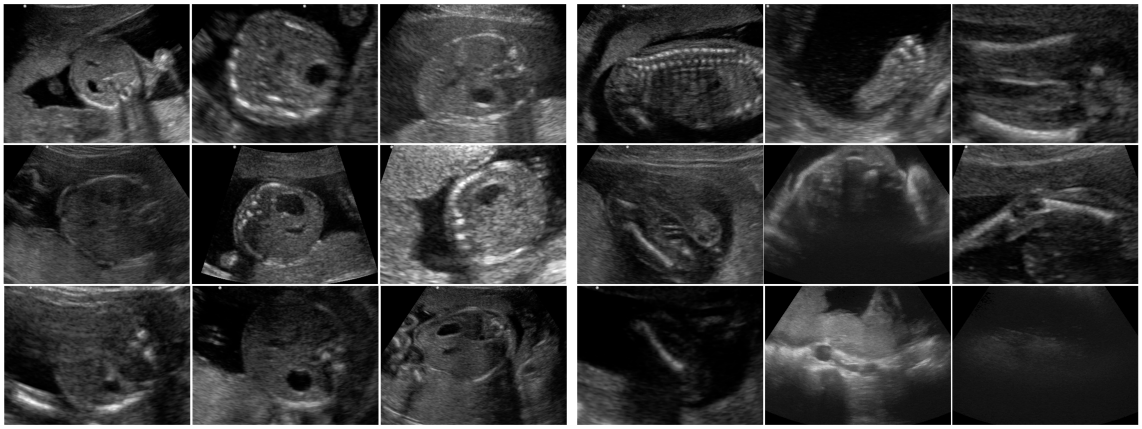


Figure 3.6: Sample of full-frame images from the testing dataset. The first three columns on the left show examples of abdominal images, and the remaining columns show the negative images from our testing dataset. Images in the positive dataset highlight issues on the challenges faced during detection, such as various sizes and shapes (top row), different contrast (middle row), occlusion, and missing boundary (bottom).

images show either an image of another part of the fetus or a background image of random textures. During detection, a test image is evaluated at all scales and locations, and results are compared against the ground truth.

3.1.2 Segmentation Dataset

For segmentation, we used the same set of 2D fetal abdominal US scans that produce a model for the abdominal localisation system as explained in Section 3.1.1. Cross-validation group was also identical. We further introduced a new set of ground truth labels for each image for segmentation task. Next, the annotation methodology employed to do this will be briefly explained.

3.1.2.1 Annotation Methodology

We manually segmented the key anatomical structures (the stomach and the UV) on the 2D US images. This manual segmentation was performed by myself, and subsequently checked by an experienced sonographer (*T.N.*). This was done using a dedicated workstation with an interactive monitor (WACOM CINTIQ 21UX). The software used to perform the manual segmentation was MITK (the Medical Imaging Interaction Toolkit MITK©, version 0.12.2, 2003-2008 German Cancer Research Center, Division of Medical and Biological Informatics).

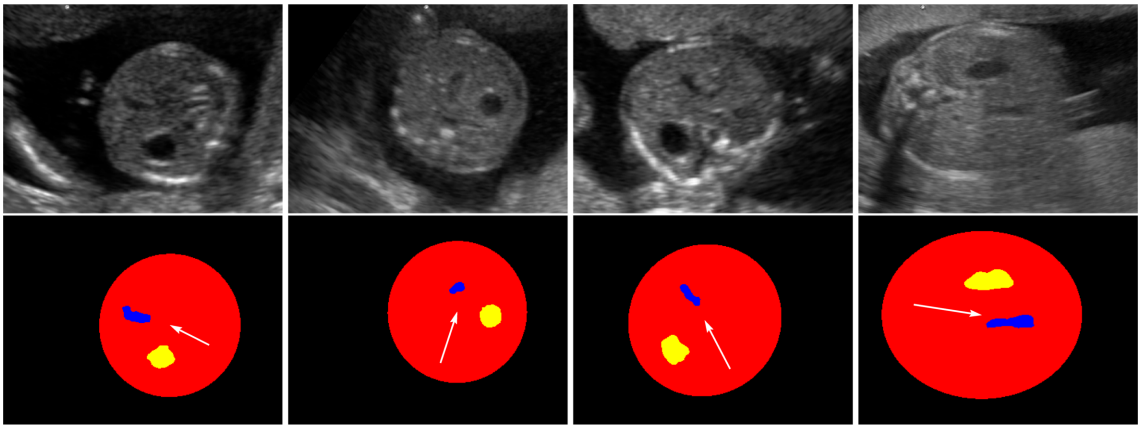


Figure 3.7: Segmentation dataset. Examples of abdominal images as used earlier in Dataset A1 (*top row*) and its respective segmentation images (*bottom row*). Abdominal (in red) is annotated from ellipsoid marker on the image, while the stomach (in yellow) and the umbilical vein (in blue) are manually annotated. The white arrow is shown pointing out from the centre of the spine to the centre of the abdomen, representing the viewing angle (pose) of the abdomen.

Figure 3.7 illustrates the pixel-wise segmentation dataset, where each structure is labelled in a different colour; red for the abdomen, yellow for the stomach, and blue for the umbilical vein. We used the previous annotated ellipsoidal ground truth of the abdomen to automatically segment all the pixels inside the ellipse as the abdominal pixels *i.e.* background), except for where the stomach and the UV were. We treated the problem as a multiclass pixel labelling, where the end goal was to label each of the pixels to belong either to the abdomen, the stomach, or the UV. We also marked the location of the spinal structure, by manually selecting a point on the centre of the spine.

3.2 Evaluation Methodology

This section highlights our choice of evaluating the performance of the proposed algorithms. It describes the evaluation criteria used to measure the localisation performance of each dataset as described in the previous sections.

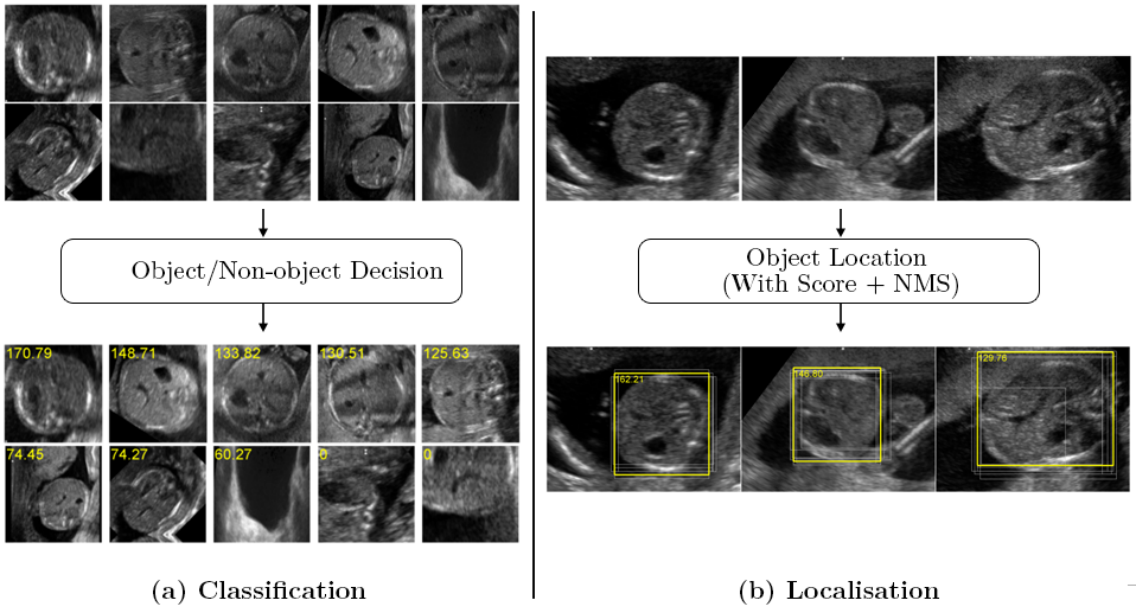


Figure 3.8: Evaluation framework for classification and localisation tasks. (a) In classification, patches will be classified as either containing object of interest or not by giving a *confidence score* (shown in top-left corner of each patch). Scores will be ranked from the highest to the lowest and progressively evaluated upon a threshold to generate a DET curve. This task was performed mainly to model the optimum parameters while designing a classifier. (b) Base on sliding window paradigm, localisation task involves not only classification on individual patch but also suppression process on multiple detections. As illustrated, using non-maximum suppression (NMS) technique, only the top-most score (*in yellow*) will be retained after supressing the overlapped lower-scoring detections. Again, scores will be ranked and the PR curves will be plotted to evaluate the localisation performance. (See text for details).

3.2.1 Classification versus Localisation

Our evaluation framework for abdominal localisation aimed to satisfy two design constraints:(a) ***high recall*** of the classifier and; (b) ***good precision*** of localisation.

We chose Detection Error Tradeoff (DET) curves (Martin et al. 1997) over the commonly used Receiver Operating Characteristics (ROC) curves (Metz 1978) to measure binary classifier performance (*i.e.* to evaluate different feature extraction algorithms/parameters) for the reasons explained later in this section, and used Precision- Recall(PR) curves (Rijsbergen 1979) to quantify the accuracy of abdominal localisation.

As illustrated in Figure 3.8, both classification and localisation tasks return a score for each image window tested, with larger values indicating greater confidence that

the window contains a structure of interest. The curve is generated from the ranked output of a method where results are ranked according to a *'confidence score'*. The respective parameters, such as number of false positives and false negatives, are evaluated while progressively varying the threshold from the highest possible score to the lowest possible score. Every single tested threshold provides a point on the curve. Each evaluation framework will now be described.

3.2.1.1 Evaluation Measure of the Classification Performance

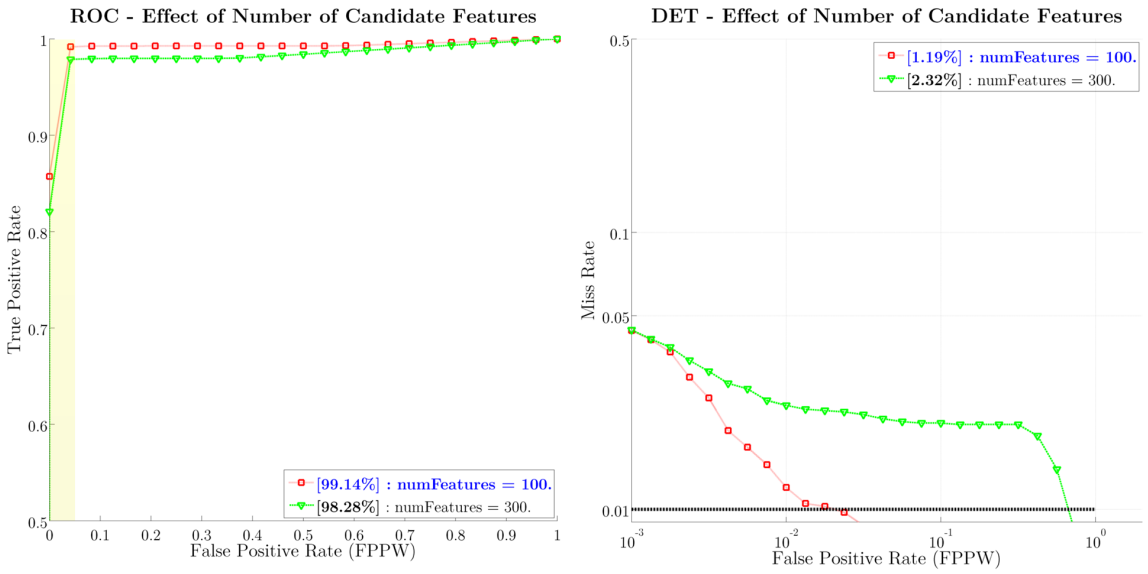
As mentioned, we aimed to build a classifier with a *high recall rate* (constraint 1), which can also be translated into a system that allows potential candidates at a very low false positive rate. Given normalised patches, we seek an answer to what makes a classifier locally discriminative. For instance, for every patch classified, we have to consider whether there is an abdomen in the window or not.

The classification performance is evaluated in terms of the correctness of labelling given to the validation patch. The principal quantitative measure used is the average miss-rate (*MR*). Using DET curves, we plotted and compared the *per-window* (*PW*) performance evaluation by measuring the proportion of miss detections as defined in Eq. (3.3) against the proportion of false positives. We used False Positives per Window tested, (*FPPW*) terminology as in Eq. (3.5) for this task.

$$Recall\ (Sensitivity) = \frac{\#TruePositivesDetections\ (TP)}{\#TotalPositivesInstances} \tag{3.2}$$

$$\begin{aligned} Miss\text{-}rate &= 1 - Recall \\ &= \frac{\#FalseNegative\ (FN)}{\#FalseNegative\ (FN) + TruePositives\ (TP)} \end{aligned} \tag{3.3}$$

$$Specificity = \frac{\#TrueNegative\ (TN)}{\#TotalNegativeInstances} \tag{3.4}$$



(a) (b)

Figure 3.9: ROC vs. DET: A comparison of different performance illustration for similar experiment. (a) Receiver Operating Characteristics (ROC) curve plots sensitivity (true positive rate) against the false positive rate while Detection Error Tradeoff (DET) curve plots the miss-rate versus the false positive rate. DET curve is plotted on a log-log scale to better distinguish the performance at low false positive rates as highlighted by yellow region of ROC curve. For DET curve, the curve closer to the south-west corner of the plot *i.e.* the red curve) is considered better.

$$\begin{aligned} \textit{False Positives Rate, FPPW} &= 1 - \textit{Specificity} \\ &= \frac{\#\textit{FalsePositives (FP)}}{\#\textit{FalsePositives (FP)} + \textit{TrueNegatives (TN)}} \end{aligned} \tag{3.5}$$

DET curves are plotted in a log-log scale to quantify the binary classifier's performance at low false positive rates. This particular attention on small FPPW sets DET curves apart from the commonly used ROC curves, which are widely found in the medical imaging analysis literature including related work of Ni et al. (2013) and Rahmatullah et al. (2011b). The y-axis of the plot represents the miss-rate, *MR*, by which a lower value denotes better classifier performance, and the x-axis represents the false positive rate per window, *FPPW*.

It is interesting to compare DET to ROC curves as exemplified in Figure 3.9. First, it can be noticed that the detection rate of the ROC plot (which typically reports performance using Area Under Curve (AUC)) remains essentially constant above a certain number of false detections. In the case where the amount of negative data

significantly outnumbers the amount of positive data, this illustration might not be the best to represent the real performance of the classifier. Secondly, it is observed that most negative training patches are generally easier to classify than positive patches, leading to a low *MR* for values larger than 10^{-1} *FPPW*. Hence in our analysis, we are interested in small *FPPW* rates (highlighted area in yellow). By contrast, the DET curve in Figure 3.9(b) heavily distinguishes between two classifiers that achieve almost perfect detection (*MR* of 0.01%) at a cost of different FPPW rates.

Instead of AUC, log-average *MR* is reported to quantify and compare the classification performance. This is computed by averaging the *MR* evaluated at 25 evenly spaced rates between the reference point of 10^{-3} and 10^{-1} *FPPW*.

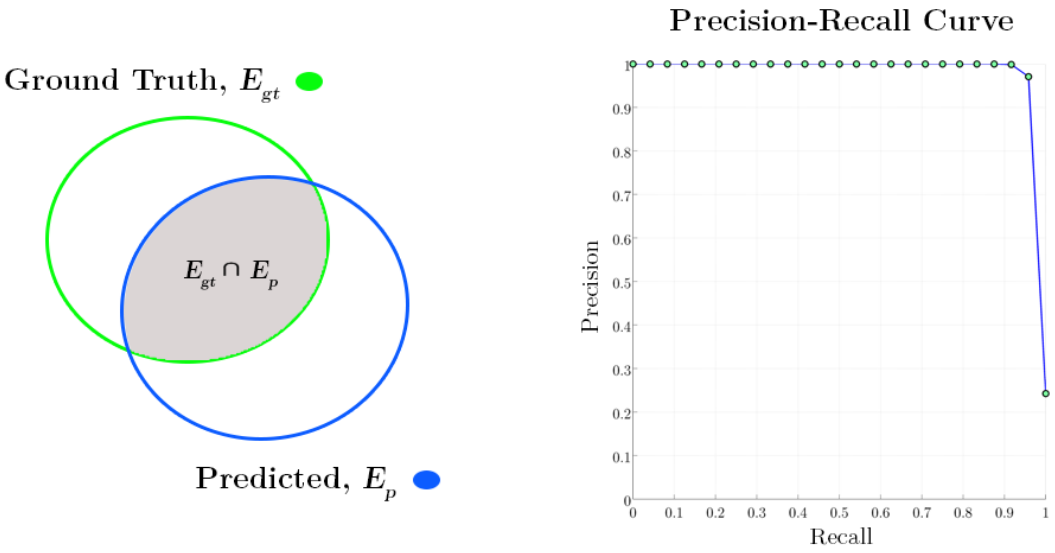
3.2.1.2 Evaluation Measure of the Localisation Performance

The former evaluation characterises the behaviour of the classifier, rather than the whole system. In the case of classification, true prediction depends only on whether the patch contains an object of interest or not; while the decision for localisation depends on whether the prediction is correct or not. Hence, on top of a good classification performance on a tested patch, the system should also be able to output the *precise location* (constraint 2) of the structure of interest *e.g.* "*where is the abdomen in the image*?"

Here, we used the definition of Everingham et al. (2015) to quantify localisation, except that the bounding box is replaced with the ellipse. The prediction is considered true if the area of overlap, $a_\circ$ between the predicted ellipse region E_p and ground truth region E_{gt} (as illustrated in Figure 3.10 (a)), is greater than or equal to a threshold, THR_A;

$$a_\circ = \frac{area(E_p \cap E_{gt})}{area(E_p \cup E_{gt})} \geq THR_A \tag{3.6}$$

where $E_p \cap E_{gt}$ denotes the intersection of the predicted and the ground truth ellipse and $E_p \cup E_{gt}$ is their union. In the case of multiple predictions per ground truth



(a) Calculation of Ellipsoid Marker Overlap (b) Calculation of Average Precision

Figure 3.10: Evaluation methodology for measuring localisation performance. Evaluation methodology for measuring localisation performance. (a) A prediction is considered a true positive if the area of overlap between the predicted ellipsoid overlaps and the ground truth ellipsoid marker are at least equivalent to a defined threshold, THR_A. (b) We use Average Precision (AP) to summarise the overall performance, computed by averaging the inter-polated precision at a set of 25 equally spaced recall rates $[0, 0.0417, \cdots, 1]$. Curves closest to the upper right-hand corner of the graph (where recall and precision are maximised) indicate the best performance.

object (if any), only the best prediction is treated as true positive and any remaining predictions are considered to be false.

As explained earlier, for each ranked detection confidence score, we compute the recall and precision are computed as follow:

$$Recall = \frac{\#TruePositivesDetections}{\#TotalPositivesInstances}; Precision = \frac{\#TruePositivesDetections}{\#TotalPositivesDetections} \tag{3.7}$$

We plot a Precision-Recall (PR) curve and report Average Precision (AP) to evaluate the overall localisation performance. Instead of using the area under the PR curve, the AP is used to summarise the shape of the PR curve by averaging the interpolated precisions at the specified 25 equally spaced recall cut-off values as illustrated in Figure 3.10 (b).

$$AP = \frac{1}{N} \sum_{r \in \{0, 0.0417, \cdots, 1\}} P_{interp}(r); N = 25 \tag{3.8}$$

The precision at the corresponding recall rate r within the interval $[r_s, r_e] \in [0, 1]$ is interpolated by taking the maximum precision measured of that interval, where $P(\tilde{r})$ is the measured precision at recall $(\tilde{r})$. This interpolation process generates a much smoother, monotonically decreasing PR curve.

$$P_{interp}(r) = \max_{\{\tilde{r}.\tilde{r} > r \in [r_s, r_e]\}} P(\tilde{r}) \tag{3.9}$$

In summary, PR curve is used to provide final comparisons of the overall localisation performance while DET curve is used to compare the performance of different feature extraction algorithms.

3.2.2 Evaluating Segmentation

Segmentation-Driven Detection: The Dice Similarity Coefficient (DSC) as formulated in Eq. (3.10) is used as a validation metric to compare and evaluate the performance of pixel-wise categorisations for both key landmarks (the stomach and the umbilical vein). DSC measures the spatial overlap between two segmentations. A^{dt} represents the automatically classified region, while A^{gt} represents the manually segmented region and $\cap$ is the intersection of the two regions.

$$DSC = \frac{2(A_{dt} \cap A_{gt})}{A_{dt} + A_{gt}} \geq THR_B \tag{3.10}$$

The structures, for which the DSC is above a defined threshold, THR_B, are considered a correct detection. The PR curve is plotted to further compare the performance of the detection algorithm based on the segmentation-driven approach.

Table 3.2: Summary of the chapter, showing key objectives and where the dataset is explained and used in the thesis.

Dataset	Used to evaluate ...	Explained in Section	Used in Chapter
Abdominal localisation dataset	• *Coarse detection*: To study the feature extraction algorithm, with two main criteria fast and precise. • *Feature comparison*: To compare and evaluate the best features for ultrasound image.	3.1.1	Chapter 4
Segmentation dataset	• *Fine localisation*: To refine the coarse detection by localising the key anatomical structures and to study different NMS approach. • *Structure segmentations*: To segment the key anatomical structures, given the abdomen.	3.1.2	Chapter 5

3.3 Summary

Table 3.2 provides a summary of this chapter, its key uses and where in the thesis, the data is used.

It should be noted that the data employed in the recent work of Ni et al. (2013) appears to have been designed to provide only the best viewpoint and image quality in guiding the selection of an ideal viewing plane. We argue that this may introduce some biases on how the measurement is taken in the real world. In this work, we developed our model from a general database, in which the sonographer was asked to perform the best measurement of whichever quality and viewing angle, without having knowledge that the measurement and expertise would be studied and evaluated using a machine learning algorithm. Thus, we expect our results to reflect more closely to the potential performance in a real-world scenario.

4

Ultrasound Object Localisation using Aggregate Channel Features

This chapter describes an original coarse objects localisation which is the second original contribution of the thesis. As explained in Section 1.4, our unified coarse to fine multiple proposal detector framework considers two closely inter-related area, key anatomical re-organisation (*Module A*) and abdominal recognition (*Module B*). It would be useful to clarify that the work produced in this chapter will first report the design of our second proposal (*Module B*). This module focuses on the abdomen detection whereas experimental works done for the first proposal framework (*Module A*) will be reported in Chapter 5. This allows us to progressively demonstrate the importance of each module in the build-up of the complete detection framework.

The approach proposed in this chapter, targeted at achieving ***fast*** and ***accurate*** object localisation in fetal US images, is motivated by recent advances in computer vision for object detection. Our approach builds on the general computer vision

framework of Dollár *et al.* which employs the use of 'Integral Channel Features' (Dollár et al. 2009) and its extended variant called 'Aggregated Channel Features (ACF)' (Dollár et al. 2014). We adapt the fast-pyramid feature approximation on a multi-resolution detection framework introduced in that work to the needs of our application. This technique has proven to be computationally efficient in the field of pedestrian detection (Dollár et al. 2014; 2012a;b) where it has demonstrated a highly accurate detection at an optimised speed.

In this chapter, we investigate the potential of ***channel features*** in providing the basis for a fast method of localising key anatomical structure in ultrasound images. Specifically, we have designed a coarse detection solution inspired by the framework of Dollár et al. (2014) which contains three important and necessary adaptions as follows:

1) We thoroughly investigate various feature encoding schemes and subsequently select an appropriate feature set for the unique appearance of US images.

2) We explore the implementation of soft-cascading threshold and score recalibration in ensuring a high recall for US object localisation.

3) We extensively explore the choice of parameters in the resulting algorithm to ensure it performs well in the intended application.

Section 4.1 introduces relevant background on the main challenges that should be taken into account in designing the learning algorithm, or computing the image features, as we aim to detect and localise the abdomen. Section 4.2 provides an overview of the complete architecture based on the well-known 'sliding-window' paradigm, including the learning and localisation phases. This includes a brief review of Aggregated Channel Feature (ACF) descriptors and also the theoretical concept of a boosting decision tree framework. Section 4.3 summarises the technical schema of the proposed feature extraction and classification algorithms, including the

experimental setup and implementation details of various choices available during modelling. We present and discuss experimental results, including an in-depth study of various feature encoding schemes and the optimal values of our designed parameters in Section 4.4. We also describe the implementation of a soft cascade rejection threshold with score calibration and show its impending practicality on ultrasound images. Furthermore, the results section also includes evaluation and quantification of the overall detection performance as well as the fusion of multiple detections over multiple scales and positions on unseen test images. Finally, a summary of our findings concludes this chapter.

4.1 Introduction

An efficient detection of key anatomical structure(s) is a fundamental task and often serves as a prerequisite for higher level analysis and understanding of medical images, for instance segmentation and registration, and subsequent quantitative analysis.

In clinical practice, sonographers typically follow a strict protocol (often called a "standardised protocol") when they acquire an image of a specific imaging plane. Examples include the Fetal Anomaly Scan Programme (FASP) standards and guidance for England by NHS Fetal Anomaly Screening Programme (NHS FASP) (2010) and the guidelines of the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) (2013) in which biometrics assessments should only be taken from standardised image planes. However, US is a heavily user-dependent imaging modality. Hence, there is interest in developing automated localisation algorithms to potentially reduce sonographer dependence as well as variations in sonographer's scanning techniques.

As considered in this work, localisation of the abdomen is considered vital during fetal anatomical evaluation *i.e.* estimating gestational age. In a real-time real world scenario, we observe that sonographers are only likely to realign their transducer to gain a better view once they have managed to obtain the first glimpse of where the

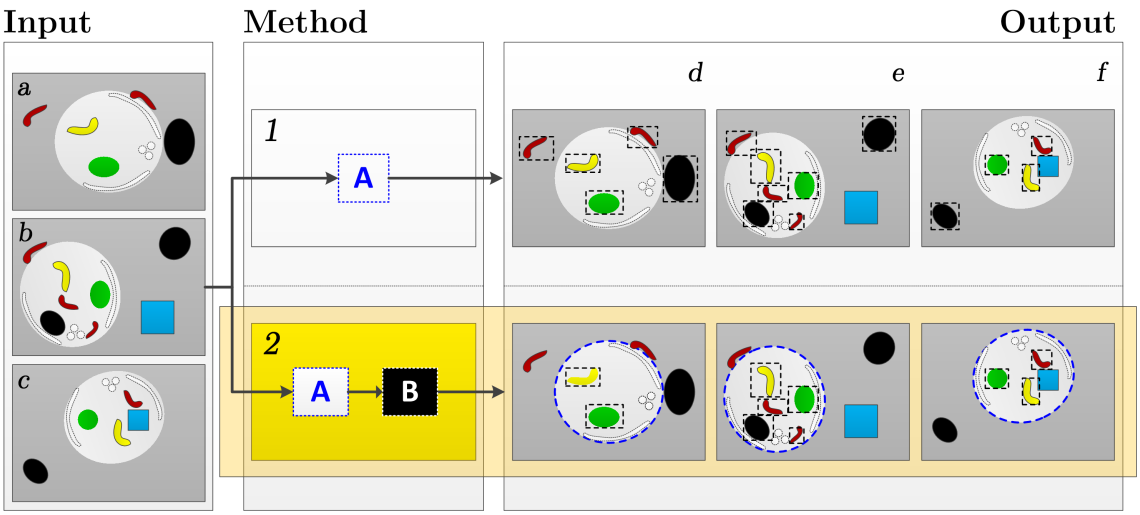


Figure 4.1: Problem formulation for coarse object localisation framework. Proposal B of Method 2 as highlighted in the diagram is proposed to provide a coarse object localisation of an ultrasound object *i.e.* abdomen, in order to reduce the searching space for respective key structures localisation. In this chapter, we provide further details of the modelling process to select the optimal parameters for the respective proposal module.

fetal abdomen is anticipated to be. Subsequent refinement is typically carried out by observing the existence of key structures (*e.g.* ribs, stomach and UV) and spatial relationship between those structures. Although these two-stages are inter-related as further highlighted in Figure 4.1, we limit our discussion in this chapter to the modelling process for the coarse localisation framework of an abdomen, while the latter refinement is discussed in Chapter 5.

The main purpose of the proposed module B is to build a system that would return a *precise location* of the structure of interest *i.e.* to answer two fundamental questions; 1) "is there an abdomen in the image?", and if yes, 2) "where is the abdomen?" In doing so, we propose an architecture that demonstrates the following properties:

1) ***Efficiency***: high computational efficiency which allows fast abdominal localisation, mimicking the sonographer's experience. This is influenced by the choice of both the features and the classifier.

2) ***High recall***: sensitive enough to accurately identify any possible candidate while discarding the vast majority of unpromising proposals, as any undetected objects at this stage cannot be recovered at a later stage. This provides an effective alternative in reducing search spaces, whilst allowing

the proposed future work of having extended set of classifiers aimed at improving accuracy as well as estimating the viewpoint (see Chapter 6).

There exists several fast and robust machine learning algorithms for detection and segmentation such as Adaptive Boosting (AdaBoost) (Viola & Jones 2001), Random Forests (Breiman 2001) and Probabilistic Boosting Trees (Tu 2005). AdaBoost has been popularly known for its simplicity, combining many weak learners to form a single strong one, while Random Forests (RF) is typically acknowledged for its efficiency in training and testing, as each tree (*i.e.* weak classifiers) could easily be parallelised. In this work, we opted for a shallow decision tree which utilises the powerful combination of AdaBoost and RF in a single framework. In particular, we implemented the Boosting Decision Tree approach of Appel et al. (2013), which guarantees an equivalent performance to the classical AdaBoost at much reduced computational cost, both in training and testing.

In conjunction with this, feature selection plays an important role in improving computational efficiency. Solving detection of objects over scale-space and multiple aspect ratios requires a system with simple and fast feature computation. Yet, finding a perfect trade-off between 'simple-(and fast)-to-compute' and robust features that also offers good detection accuracy is a non-trivial task. In searching for this, Dollár et al. (2014) proposed aggregated channel features (ACF), where features are simply a single pixel of the ACF lookup table. Together with a fast feature pyramid construction, the ACF, which is derived from the original Integral Channel Features (ICF) framework, demonstrated that a carefully designed feature based on single pixels is powerful enough to outperform the more expensive features (*e.g.* ICF are based on sums over rectangular channel regions computed via integral images *i.e.* Haar features).

In a bid to achieve a high recall during coarse localisation, a model should be robust enough to be able to detect objects of arbitrary viewing pose with a minimal false detection rate. We adopted the soft-cascade rejection framework introduced in Bourdev & Brandt (2005) to handle large variance in the abdomen viewing pose

effectively. A soft-cascade rejection architecture; 1) elegantly recalibrates the weight of a classifier in the validation stage to achieve a high recall percentage for the coarse classification (*i.e.* indirectly implying that the system is tolerating the pose variance) and 2) offers processing speed-up by having irrelevant image windows rejected early on. We will now elaborate on the underpinning strengths of each of these approaches as we seek to understand the value of combining them in a medical imaging analysis solution.

4.2 Overall Architecture

Our abdominal localisation architecture is built around a sliding window approach, which can be divided into two phases. The ***learning phase*** creates a binary classifier which provides object/non-object decisions given a normalised image window (or patch); while the ***localisation phase*** uses the designed classifier to perform an exhaustive search over possible locations, scales, and aspect ratios of an unseen image, assigning a *confidence score* to each extracted image region - a score which represents its *'objectness'*. These preliminary predictions are then fused to obtain the final object localisation. The overall architecture is summarised in Figure 4.2 and the two phases are elaborated in the next two subsections.

4.2.1 Phase 1 – Data-driven Learning Process

The first phase, the learning process, contains three stages; data preparation, feature encoding, and classifier modelling. Sample preparation of the training-validation has been described in Section 3.1.1. We will now describe the feature sets used to model the appearance of image patches and the machine learning technique employed to learn these features.

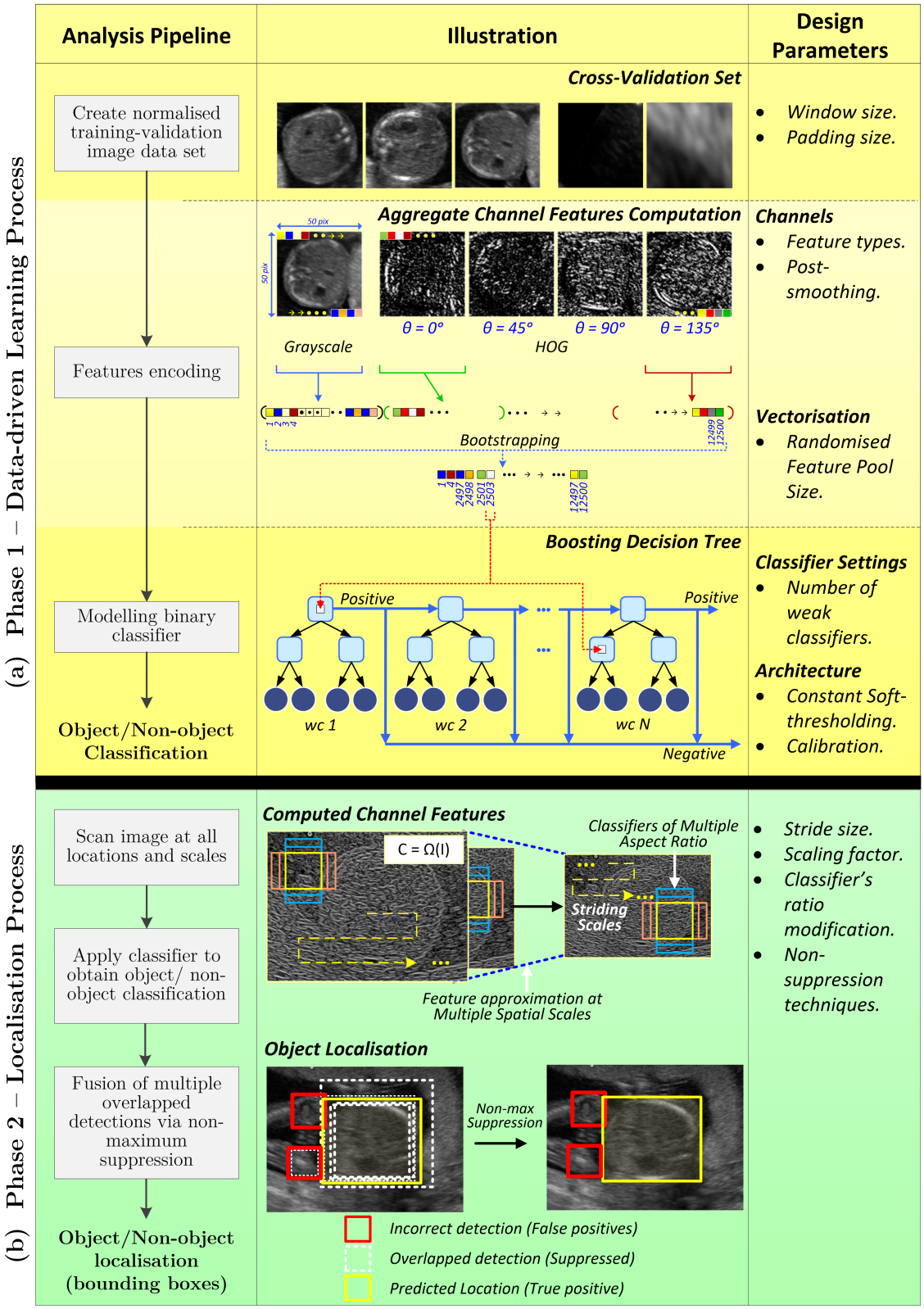


Figure 4.2: Overall object detection architecture. (a) *Phase 1 - The learning process*: Visual features are extracted from fixed-size training data. Parameters of interests are learned via cross-validation to form the final binary classifier. (b) *Phase 2 - The localisation process*: The designed classifier scans through an unseen test image at all locations and scales for object/non-object detection. These preliminary bounding-box detections are later fused to produce the final abdominal localisation.

4.2.1.1 Overview of Feature Sets

Recently, a family of channel features (Dollár et al. 2014; 2009) has been introduced and has recorded excellent performance in the domain of pedestrian detection. A "channel" feature is simply a straightforward pixel-wise feature map of an input image that undergoes transformation(s) either via a linear or non-linear computation. For instance, given an input image I with dimension $m \times n$ pixels, several corresponding channels, $C = \Omega(I)$, such as grayscale and oriented gradient features of similar size are directly computed (hence preserving overall image layout). The simplicity and effectiveness of its feature computation makes it possible to extract multiple, diverse channels from an image henceforth facilitating a rich feature representation of an image window for possible feature selection and analysis.

Here, we consider and test the performance of several types of channels. We partly experimented with the same channels as suggested in the original implementation (Dollár et al. 2009) as well as other channels commonly used in the ultrasound literature. We briefly provide an overview of the feature channels, as further illustrated in Figure 4.3.

Grayscale (Intensity): As illustrated in Figure 4.3 (a), the grayscale channel, $\Omega_{GRAY}(I)$, is the intensity channel that directly represents the original image.

Gradient Magnitude: The gradient magnitude channel, $\Omega_{GM}(I)$ as illustrated in Figure 4.3 (b), characterises edge strength (but not orientation). Given an input image $I\ (x,y)$, a gradient magnitude channel, Ω_{GM} represents the gradient magnitude, $M(I) = \sqrt{(\partial I/\partial x)^2 + (\partial I/\partial y)^2}$ is computed, where $\partial I/\partial x$ and $\partial I/\partial y$ denote the discrete derivatives of I.

Gradient Histogram: The gradient orientation of an image is defined as $\phi(I) = \arctan((\partial I/\partial y)/(\partial I/\partial x))$. In order to compute the gradient histogram, the gradient magnitude of each pixel location is quantised into Q orientation bins according to its respective gradient orientation (angle) so that $\phi(I) \in [1, Q]$. Hence, the qth bin of the histogram is defined by: $h_q = \Sigma_{(x,y)} M(I) 1[\phi(I)=q]$ where 1 is the indicator

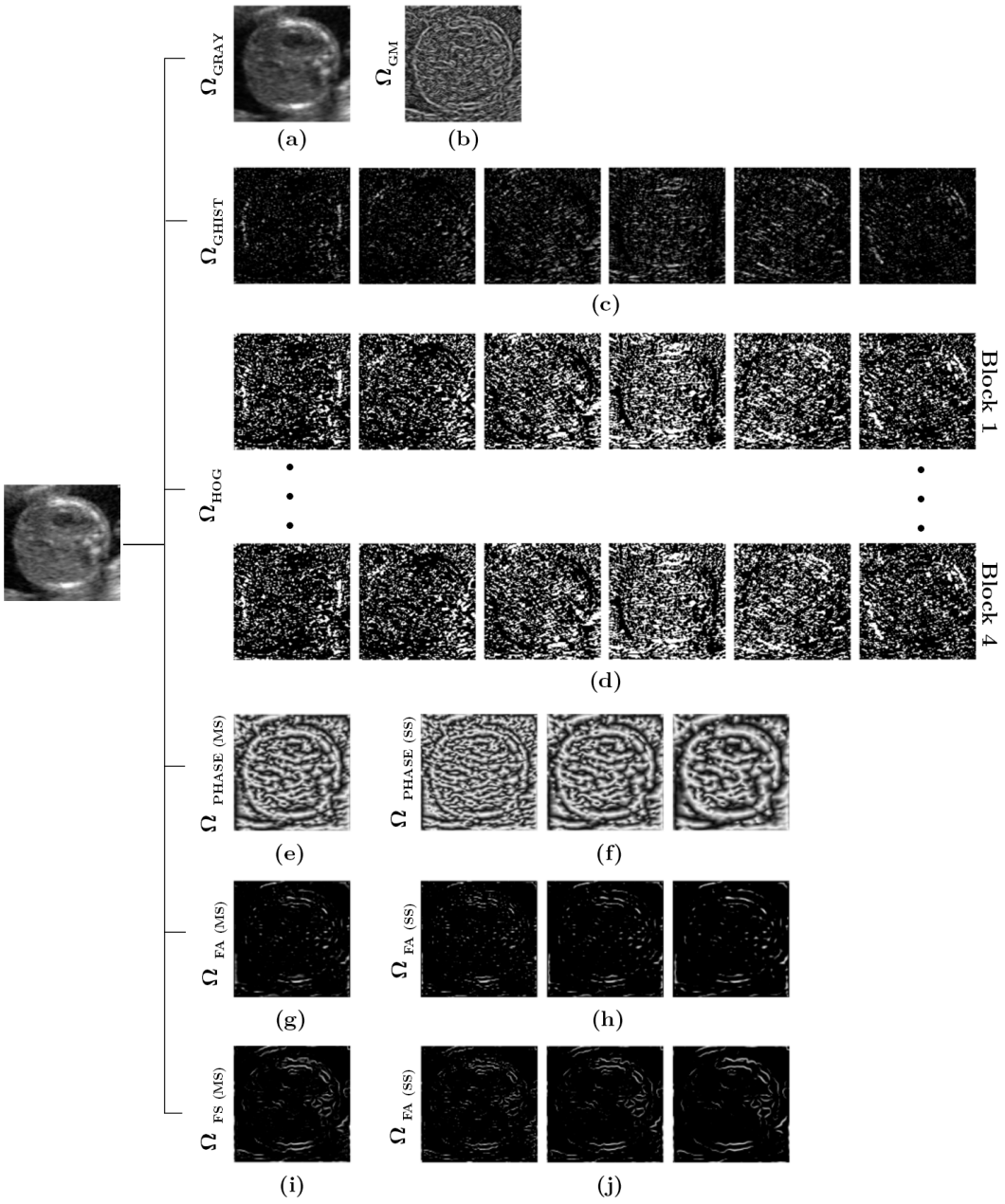


Figure 4.3: Visualisation of channel features samples. Given an input image, multiple image channels are generated using various transformations. Among the features that we experimented in this works includes the original grayscale channel, gradient magnitude, gradient histogram, HOG, and local phase based filtered channels, which considered both single-scale (SS) and multiple scale (MS) frequencies. These features will be vectorised to form a long feature look-up table consists of zero-order features (*i.e.* pixel values). (Note: All features have been normalised to [0,1] value range for a better visualisation. See text for details on how features are generated)

function. An example of a gradient histogram channel, Ω_{GHIST} which is quantised to 6 orientation angles, $\theta_Q \in \{0°, 30°, \ldots, 150°\}$ is shown in Figure 4.3 (c).

Histogram of Oriented Gradient (HOG): From the extracted gradient histogram channels, we generate approximate HOG features using the method of Dollár et al. (2014). For each resulting histogram (of Q orientation bins), four normalisations are computed using the L2 norm on 2×2 adjacent histograms, resulting in a four times expansion of the number of HOG channels, Ω_{GHIST}. Hence, each histogram (except at boundaries) thus has 4 different normalisation values associated with it. We illustrate the first and the last block of the HOG channels, Ω_{HOG} in Figure 4.3 (d).

Local Phase Based Filter: Following some successful applications of local phase-based analysis on ultrasound images (Mulet-Parada & Noble 2000, Rahmatullah et al. 2012a, Rajpoot et al. 2009), we decided to experiment with features generated using this approach in our application. We followed the implementation of Rahmatullah et al. (2012b) to generate channels based on local phase computation as illustrated by the flow chart in Figure 4.4. Given an image *I(x,y)*, we produced a monogenic signal, $I_M(x,y)$ as proposed by Felsberg & Sommer (2001) via a series of convolutions of the image with an isotropic band-pass filter and the Riesz filter. The image *I(x,y)* was first convolved with an even isotropic band-pass filter, *b(x,y)* to produce the even (symmetric) component of the monogenic signal. The bandpassed image, $I^b(x,y)$ was then convolved with the Riesz filter to produce the two odd (antisymmetric) components: $odd_1(x,y) = h_1(x,y) * I^b(x,y)$ and $odd_2(x,y) = h_2(x,y) * I^b(x,y)$. The monogenic signal can be expressed as in Eq. (4.1);

$$I_M(x,y) = [even(x,y),\ odd_1(x,y),\ odd_2(x,y)]; \quad \textit{where the even and odd signals are the filtered components of the monogenic signal} \tag{4.1}$$

We extracted and tested several derived features from this local phase representation, including a single and multi-scale local phase image (which are

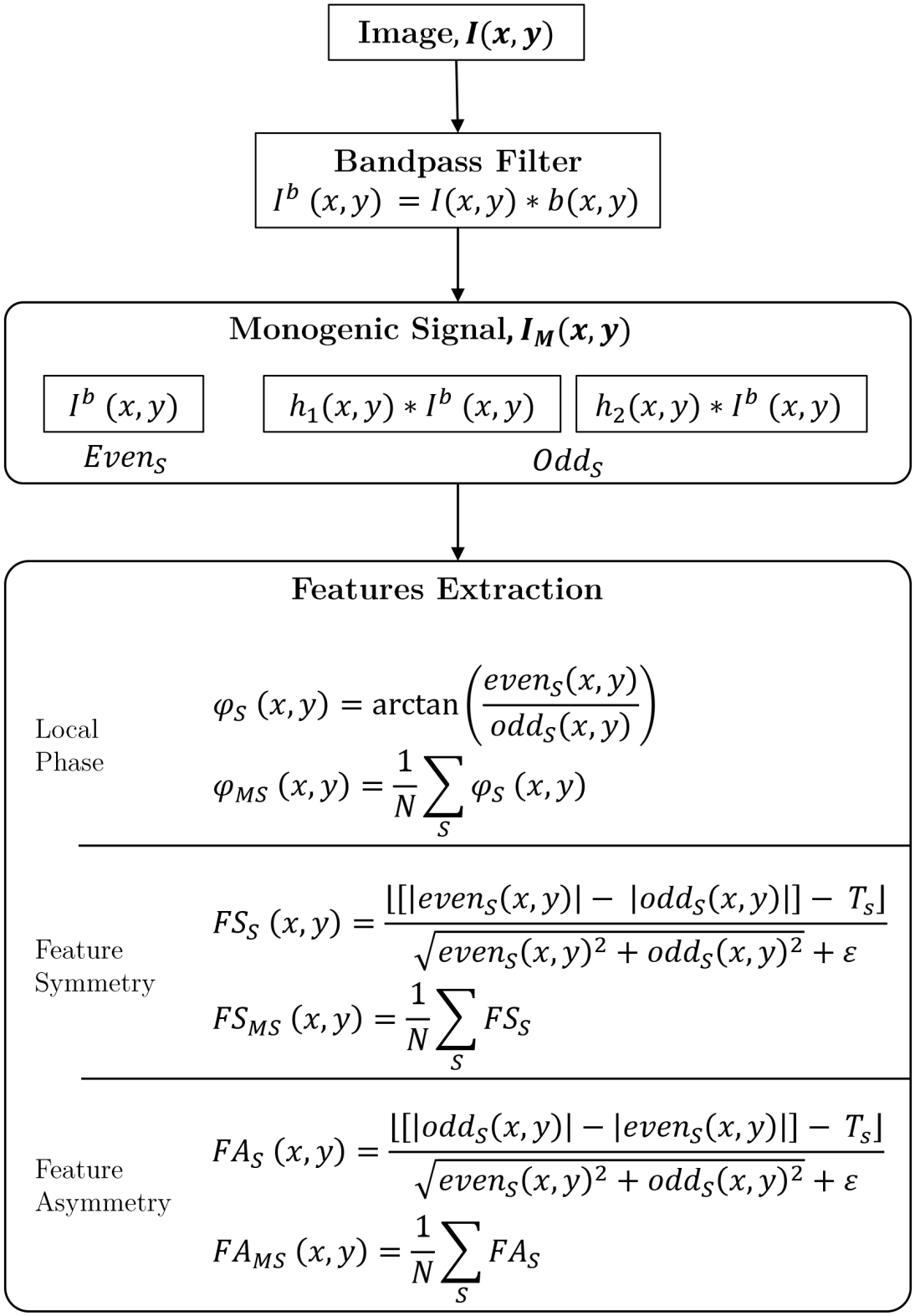


Figure 4.4: Feature extraction from local-phase information. Flow chart shows the procedure to generate relevant features from local-phase information. Subscript ***S*** and ***MS*** represent single-scale and multi-scale respectively. (See main text for definition of symbol)

represented by subscript *s* and *ms* respectively) as well as feature symmetry (FS) and feature asymmetry (FA) as introduced by Kovesi (1997). The measure of symmetry and asymmetry include the following term:

$$T_s = \exp^{\left(\text{mean}\left(\log\sqrt{even_s(x,y)^2+odd_s(x,y)^2}\right)\right)} \tag{4.2}$$

where s represents the scale of the band-pass filter, T_S is a scale specific noise compression term, and ε is a small positive constant to prevent division by zero, as used in Kovesi (1997). We also compared the performance of different type of bandpass filter, which include the selection of various frequency bands and scales. Samples of the channels are illustrated in Figure 4.3 (e) - (j).

Feature vectorisation: The features described above are vectorised into a single lookup table. For example, given an image window size of 120×120 pixels, a model which implemented a grayscale (intensity) feature alone will produce $14,400$ features (*e.g.* $120 \times 120 = 14,400$ individual pixels). If this feature is to be combined with the gradient magnitude feature and histogram of oriented gradient (24 channels) feature, it will then produce a total of $\Omega_C = 374,400$ features to be trained from,

$$\Omega_C = \left[\Omega_{GRAY,[1:14,400]}\ ;\Omega_{GM,[14,401:28,800]}\ ;\Omega_{HOG,[28,801:374,400]}\right] \tag{4.3}$$

where the subscript indicates the respective feature index after concatenation. We evaluated various combinations of the aforementioned features and compared the performance using the detection toolbox[1] provided by Dollár *et al.* Dollár et al. (2014). Some of the parameters tested per feature will be explained later in the experimental setup.

4.2.1.2 Boosting Decision Tree

Boosting offers a convenient approach to learning, given a large number of candidate features. This is done by combining many weak learners to form a single strong one. The technique has been popular since the introduction of the Viola-Jones detection

[1]An updated version of Piotr's Computer Vision Matlab Toolbox is freely available from https://pdollar.github.io/toolbox. We use an old version 3.22 which should be sufficient for this work.

framework (Viola & Jones 2001) which trained a simple single decision stumps per weak classifier. However, this approach has been reported to lack of discriminative power when dealing with an object detection of a complex and unconstrained settings (Yang et al. 2014a). A popular method to tackle this issue introduces weak classifiers within a tree containing multiple nodes (*i.e.* single decision stump per node). Instead of having only a single decision per weak classifier, a combination of multiple features (of multiple nodes) per weak classifier offers more complex features which indirectly present a better discriminative power (Carneiro et al. 2007, Yang et al. 2014a).

As we desire a quick decision-making framework for coarse localisation that also offers simplicity and robustness in practice, we limited ourselves to only use a shallow tree *e.g.* depth-2 decision tree ($L = 2$) to characterise a single weak classifier. We followed the implementation as given in (Appel et al. 2013). In this work, we trained a single stage classifier. Training a boosting decision tree is similar to training a decision tree or Random Forest, except that a boosting classifier (*i.e.* AdaBoost) is employed at each node to split the data.

Specifically, a boosting classifier having the form $H(x) = \sum_t \alpha_t h_t(x)$ was trained by selecting a weak learner $h_t(x)$ that optimises a weight scalar α_t at each iteration t. Each data sample, x_i was assigned an initial non-negative weight, w_i , in which at each iteration, the misclassified samples are weighted more. A decision tree $h_{tree}(x)$ is composed of a stump $h_j(x)$ at every non-leaf node j. A stump was tested, one at a time starting from the root, working all the way through to the lower nodes. Given an input $x \in R^K$ which parameterised with a polarity $p \in \{\pm 1\}$, a threshold $\tau \in R$, and a feature index $k \in \{1, 2, \ldots, K\}$, a stump produces a binary decision,

$$h_j(x) \equiv p_j \text{sign}(x[k_j] - \tau_j) \tag{4.4}$$

where $x[k]$ indicates the kth feature of x. At each node, the goal is to select the single best feature k^* from all the features that minimise the error, ε:

$$\{p^*, \tau^*, k^*\} = \arg\min_{p,\tau,k} \varepsilon^{(k)} \tag{4.5}$$

in which the weighted classification error can be rewritten as:

$$\varepsilon^{(k)} \equiv \frac{1}{Z}\Big[\sum_{x_i[k]\leq\tau} w_i 1_{\{y_i=+p\}} + \sum_{x_i[k]>\tau} w_i 1_{\{y_i=-p\}}\Big] \qquad ; Z \equiv \sum w_i \tag{4.6}$$

[where $1_{\{\ldots\}}$ is the indicator function.]

The boosting process continues until a number of weak classifiers have been generated or an early termination criterion is reached *i.e.* minimum error.

4.2.1.3 Soft-Cascade Threshold

Localisation speed is determined by the simplicity of the features, as well as the classifier architecture. Viola & Jones (2004) presented a *cascade* strategy in their AdaBoost implementation to attain a computational speed-up by rejecting massive amounts of arguably easy negative patches early on in the classification process, which are much simpler to discriminate and only require few classifiers. While having a decomposed cascaded classifier with increased feature complexity presents an opportunity for better accuracy, this approach has been criticised by several papers, for instance, (Bourdev & Brandt 2005), due to its premature information removal as well as its inefficiency and complex training procedure.

We opted for a '*soft-cascade*' architecture introduced by Bourdev & Brandt (2005) for what we believe to be the simplest, yet the most applicable solution to achieve our goals. Using this approach, we trained a single-stage boosted classifier, H consisting of K weak classifiers, as explained in the previous section.

$$H(x) = H_K(x) = \sum_{t=1}^{K} \alpha_t h_t(x) \tag{4.7}$$

where h_t is a weak classifier (*i.e.* depth-2 decision tree of 3 decision stumps) selected at step t and scaled by its associated weight, α_t. *H(x)* represents a confidence score of an image window of feature vector x, and classified as positive if $H(x) > 0$.

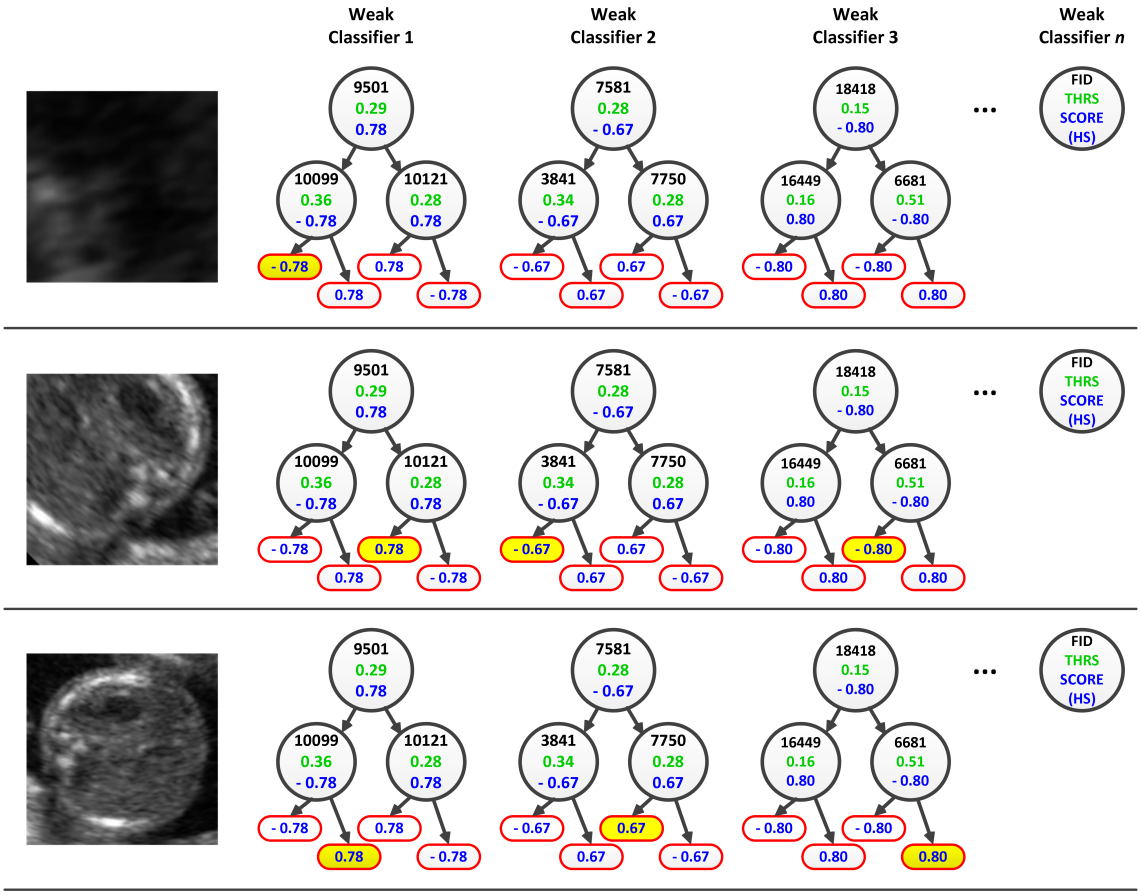


Figure 4.5: Sample of boosting decision tree with soft-threshold rejection. In this figure, multiple cases were provided to illustrate how an image window is evaluated using a calibrated constant rejection threshold, *e.g.* $\theta^* = -0.5$. The weak classifiers show the feature id (FID), the stumps value (THRS) and the weighted score (HS). Once a decision has been made, the score at the leaf node (highlighted in yellow) will be evaluated against the threshold, θ^*. Consecutive weak classifiers will only be evaluated if the cumulative score at stage k is above the rejection threshold, *i.e.* $H_k < \theta^*$. For example, a negative sample is rejected at the earliest possible instance and this speeds-up the evaluation process by not having to test the rest of the weak classifiers (top row). A potential hard negative sample might be evaluated up to certain number of weak classifiers and their evaluation stops once its cumulative score, $H_k < \theta^*$ (middle row). Only positive images will have all the weak classifiers evaluated (bottom row). The first few weak classifiers are selected to ensure that only hard negative examples will be rejected early on while consecutively building up a more complex classifier as more weak classifiers are added.

When training a boosting tree, weak classifiers are chosen subject to a classification score, $H(x) > 0$; if and only if x is positive. In an analogous manner, this observation is also true for all intermediate classifier, $H_K(x)$ for $k<K$ which is trained to attain a similar objective function. Having said that, it is possible to design a constant rejection threshold, $\theta_K = \theta^* \leq 0$ for all k, exploiting the fact that it is highly unusual to have $H_K(x) < \theta^*$ (with the likelihood decreasing with decreasing θ^*) in order to meet a requirement for $H(x) > 0$.

We illustrate this using the example in Figure 4.5. During evaluation, a soft-cascade will test the cumulative score per stage, H_K against a defined rejection threshold, $\theta_K = \theta^*$. This will determine whether the evaluation should be continued for the next weak classifier, or whether the samples get declined immediately. Given $\theta^* = -0.5$, for instance, the sample in the top row is shown to be rejected as early as the first evaluation as $H_1(x) = -0.78 \leq -0.5$. If the window passes the first test, we can test the next feature with the same constant threshold procedure so that it produces few or no false negatives on the output of the first test. This is illustrated by the second sample (middle row) where it was only rejected after the third stage *i.e.* once its cumulative score, $H_3(x) < \theta^*$, while samples with $H_K(x) \geq \theta^*$ for every stage k will accumulate the score up to the last weak classifier as illustrated in the last sample (bottom row). This simple constant rejection threshold architecture allows early elimination of the most obvious negative samples, hence saving computation time.

In conjunction with the constant rejection threshold, we also applied a *classifier recalibration* approach with the aim to achieve a certain target miss-rate, MR_{opt} (*i.e.* equivalent to reciprocal of the detection rate) within a specific target time, T_{opt}. Constant rejection thresholds can be extended to reject more possible negative windows by *recalibrating H*, such that $H(x)) < \phi$. By assigning $H'_k(X) = H_k(X) - \phi$; where $\phi = k/K$ for all k; the calibration could be designed such that, $H'_k(X) > 0$ if and only if $H_k(X) > \phi$.

Using the constant threshold, $\theta_K = \theta^*$ with H' is equivalent to using $\theta_K = \theta^* + \phi$ with H where the calibration parameter, ϕ controls the trade-off between the reduction in false positives (*recall rate*, which should be higher) and the decrease in detection rate (speed, which should be lesser). By using this approach, we ensured that a list of possible candidates is maintained and not rejected early on, up to a certain evaluation stage. We found this property fairly useful given an ultrasound patch of *random* nature, where the system may well tolerate for the missing features up to a certain extent.

4.2.2 Phase 2 – Localisation Process

Referring to Figure 4.2, while the trained classifier learned in Phase 1 has been designed to meet the objective of being able to discriminatively classify a positive and negative patch (classification task), there are still some additional settings to be considered in order to achieve the ultimate objective of a localisation algorithm which outputs the ***location*** of the abdominal structures within their original image context (localisation task). Among them are the choices of ***spatial*** and ***stride*** scale, and the use of a ***non-maximal suppression*** technique.

Fast Feature Pyramid Scheme: In the computer vision community, fast feature pyramid scheme refers to the technique of reproducing an image at multi-scale level. Standard image processing pipeline will normally construct a densely sampled image pyramid of multiple sizes before features can be computed at each scale. Here, given a full-size test image, respective features will be computed once for the whole image. Rather than computing features explicitly per individual scales (*i.e.* recomputing features over rescaled image), we implemented fast feature pyramids (Dollár et al. 2014) to generate approximated multi-resolution image features via feature extrapolation. This approach has demonstrated significant speed-up with minor loss in accuracy in solving object detection problems (Dollár et al. 2014).

Given an image I, a channel, $C = \Omega(I)$ is computed from corresponding input pixels. Let I_s denote I at scale s, where the dimension $h_s \times w_s$ are s times the dimensions of I. We followed the implementation of Dollár et al. (2014) in generating the approximated multi-scale image features, I_s. Using similar approximation principles, the learned feature is also re-estimated to model a classifier of multiple aspect ratio. Details of this approach will be described in Section 4.3.3.2.

Full Image Evaluation: Once the multi-scale image features have been computed, a binary classifier which has been modelled via the learning process in Phase 1 will now slide over the scale-space images and process them to produce an object/non-

object decision for the respective image windows. A localisation system needs to take an image and return a bounding box with its confidence score for each detection.

In doing so, this approach will typically produce multiple overlapping detector responses at close vicinity, which should be fused to produce a single coherent decision. For example, initial evaluation on multiscale detection will first produce a list of bounding boxes (BBs) with a confidence score over a pre-defined threshold, as illustrated in Figure 4.2(b). Next, a necessary non-maximal suppression (*NMS*) technique should be applied for merging nearby detections. Going through a list of detected BBs, the top-score detection will suppress (if any) BBs should their areas of overlap exceed 50%. The process is repeated if there exist another BB which does not overlap or has an overlapping area of less than 50% with the previous merged detection. In this example, an image could end up with several detections; either it is a true positive case (in *yellow*), or a false positive case (in *red*) after suppressing nearby detections (in *white*).

Hence, the overall localisation accuracy depends very much on how finely-grained the image is scanned, and how the final detection is fused or selected. A discussion on the *stride scale* as well as the criteria used for *NMS* isprovided in Section 4.3.3.2 when the experimental settings are described.

4.3 Experimental Setup

Our experiments utilised the channel feature toolbox written by Dollár et al. (2014). Other than the parameters being experimented in the context of our research for ultrasound image analysis as explained later in the following subsections, we kept the original settings as given by that author.

Table 4.1: Summary of the 5-fold cross-validation dataset. Data shows the distribution of sample proportions for each fold. Each set contains a total of 800 positive patches and 2,500 negative patches. *Dataset* $C^{+/-}$ refer to negative samples collected from positive abdominal images and non-abdominal images respectively.

	Train				Valid
K-Fold Set	Set 1	Set 2	Set 3	Set 4	Set 5
Positive Set	800	800	800	800	800
Negative Patch (C^+)	2000	2000	2000	2000	2000
Negative Patch (C^-)	500	500	500	500	500
Total	13600				2600

4.3.1 *K*-Fold Experiments

We run a K-fold cross-validation technique with, K=5, to assess and validate the parameters of the model. We used the normalised patches from *Dataset B*2 and *Dataset C* as previously summarised in Table 3.1 (page 44). The samples were randomly partitioned into 5 equal size subsamples with equal proportions of the two classes as illustrated in Table 4.1. Cross-validation was repeated 5 times, with each subsample used exactly once as the validation set. We computed the log-average miss-rate performance as described in Section 3.2.1.1 for each fold and reported the average MR over the 5-fold experiments. We report the average miss-rate performance over the experiments. We compare the classification performance of the binary classifier learned and select the model parameters with the best average predictive accuracy (*i.e.* lowest average miss-rate).

4.3.2 Model Parameterisation

During cross-validation, we kept all design parameters constant when we parameterised the optimal settings of the classifier (*i.e.* window size, number of classifiers *etc.*) except those explicitly being assessed. We will now describe some design parameters tested as illustrated in the last column of Figure 4.2 (page 62).

i) **Window Size:** We quantitatively evaluated the impact of the size of the normalised window and its extended padding. Figure 4.6 illustrates the

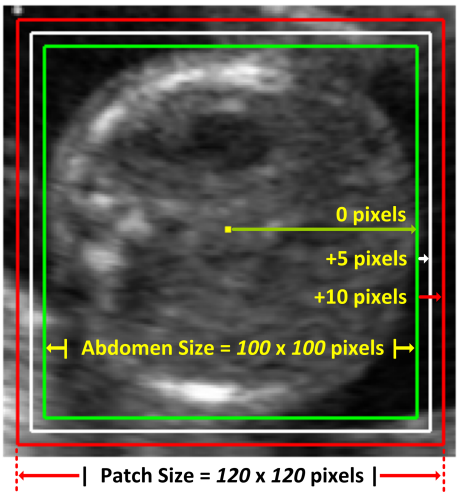


Figure 4.6: Illustration of Window Size Parameters. We refer window size of a normalised patch as the overall dimension of a patch inclusive of the abdominal size and padding pixels (if any). For example, an abdominal of size 100x100 pixels and 10 padding pixels (refer *red* bounding box) resulting a window size of 120x120 pixels.

parameters tested. We define window size as the enclosing bounding box which includes the object of interest and its extended padding (if any). Zero pixel padding refers to a window with only the abdominal area included while +5 pixels padding represents an additional 5 pixels (around each side of the patch) that were included in the learning process. We hypothesised that an image window with inclusive padding might improve detection. As shown in the illustration given in Figure 4.6, a normalised window with an abdomen of size 100 × 100 pixels and an additional +10 pixels padding is said to be of size 120 × 120 pixels (refer to image window bounded in *red*).

ii) **Feature Encoding**: Two recent works (Cheng et al. 2014, Zitnick & Dollár 2014) hypothesised that the likelihood of a box containing an object is higher if there is a well-defined closed boundary or contour. To explore this hypothesis, several features and their combinations were extracted and compared to study which features are more relevant to ultrasound images. We evaluated the effect of other feature parameters, such as the orientation and the smoothing factor of HOG features and different frequency bands and types of local phase filter features. We also considered the role of

randomisation and the effective number of features candidates given a pool of features in optimising the performance.

iii) **Classifier Settings**: We evaluated the optimal number of weak classifiers to ensure that a model does not overfit. We also studied how a soft-constant rejection threshold coupled with score recalibration could potentially be useful in a challenging US image environment while also offering an efficient processing time.

For each experiment, we plotted average miss-rate versus false positives per window (*FPPW*), where lower curves indicate better performance. Legend entries display log-average miss-rate as a common reference value for summarising the performance.

4.3.3 Overall System Architecture

4.3.3.1 Key Statistics Training Dataset

We observed and analysed some statistical patterns of the 500 positive abdominal images in Dataset A1, in particular to study the relationship between the key landmarks available.

Figure 4.7 shows histogram distributions which exemplify additional information in designing the full image localisation model. Figure 4.7 (a) illustrates how a number of key object geometric size ratios vary over the dataset (as a measure of dataset variability). Of note, the mean object-to-object ratio is 33% and approximately Gaussian. The aspect ratio of the abdomen *i.e.* ratio of the width over the length of the abdomen as in Figure 4.7 (b) which is used to set the spatial scale of our detector is slightly skewed and has a mean of 1.027. The size ratio of the stomach and the UV with respect to the abdomen are given in Figure 4.7 (c) and (d) respectively. This ratio is used to design our fine structure segmentation method as explained later in Section 5.3.1.

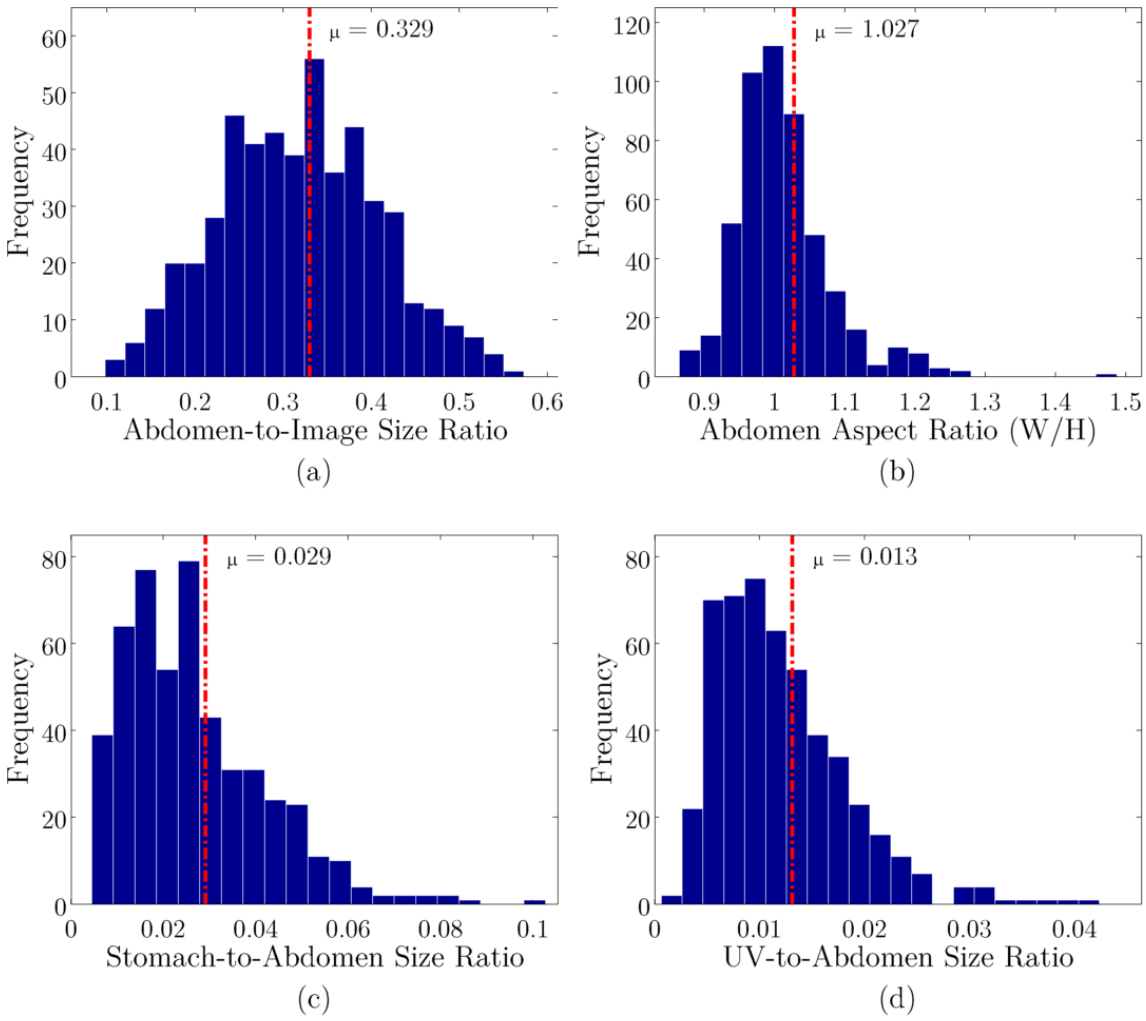


Figure 4.7: Histogram distributions of size ratios for various pair of structures in the training-validation dataset. We calculate the pixel ratio of different pair of structures for each image in the training-validation dataset. Histogram represents number of images tabulated over 20 interval bins. This information serves as a reference while we design the detection system.

4.3.3.2 Full Image Localisation

Given a fully developed binary classifier and full-frame validation image, we conducted a number of experiments to find the optimal choices of the *spatial scaling factor* (and *aspect ratio*) and *stride size* parameters. Experiments included how the choice effects the execution time where our objective was to achieve a certain target miss-rate, MR_{opt} within a specific target time, T_{opt}.

Spatial scaling factor and aspect ratio: In according to the statistical distribution presented in Figure 4.7 (a) and (b), there are two main design principles need to

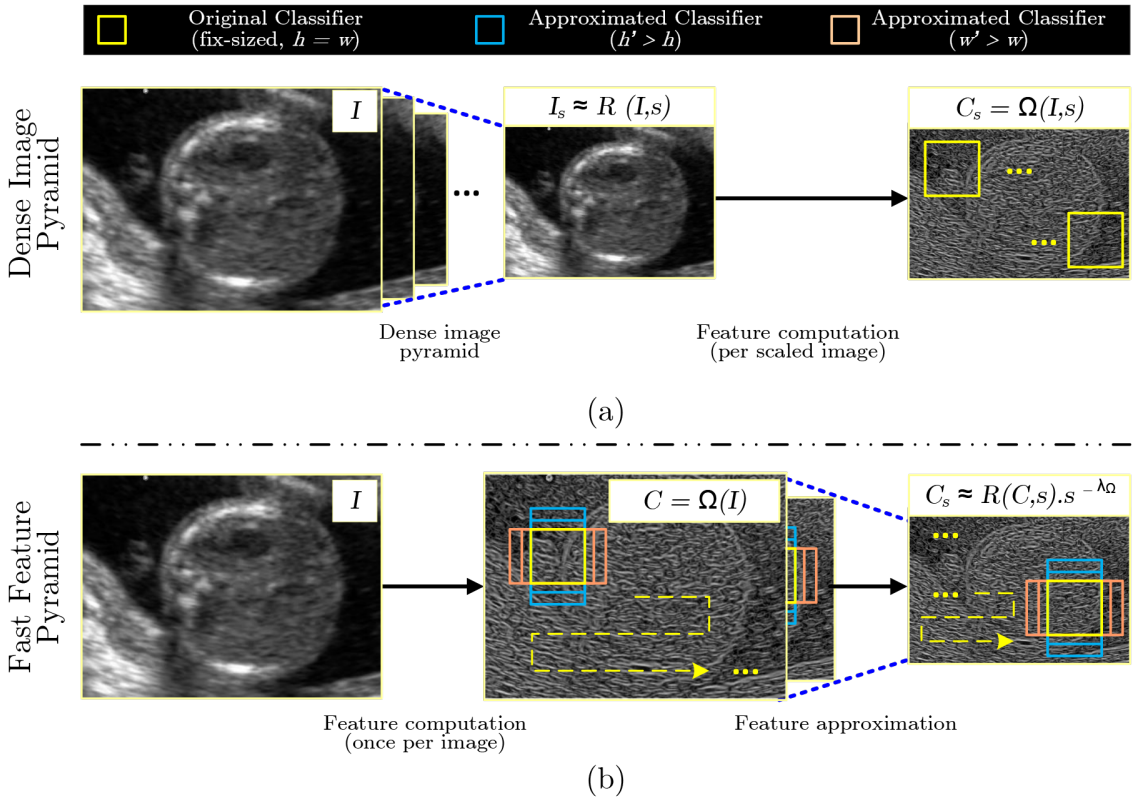


Figure 4.8: Fast feature pyramids scheme consists of feature channel scaling and approximation. (a) Standard pipeline approach normally constructs a densely sampled image pyramid $I_s = R(I,s)$ first, before computing features at each scale, $C_s = \Omega(I_s)$. Sliding window classification (with a fixed-scale model) is then executed for performing multiscale detection. The construction of the feature pyramid will mostly dominate the cost of detection, leading to slow multiscale detection. (b) Instead, fast feature pyramids scheme (Dollár et al. 2014) proposed to compute $C_s = \Omega(I_s)$ for only a sparse set of s (once per octave) and then computed C_s at intermediate scale using the approximation $C_s \approx R(C,s) \bullet s^{(-\lambda_\Omega)}$. In our experiments, we also apply the same approximation principles to reproduce the original trained classifier into multiple-size classifiers (of different aspect ratios) before performing sliding-window classification task.

be addressed under this topics; 1) *scale-space size* - to detect abdomen of various size across the images, and 2) *irregular size ratio* - to detect abdomen over a range of aspect ratios.

We adopted the fast-feature pyramid approximation as introduced by Dollár et al. (2014) to compute features at multiple scales. We illustrated and briefly described this strategy as compared to the standard pipeline in Figure 4.8(a). Let Ω be any low-level function (*e.g.* intensity, gradient *etc.*) which takes an image I and creates a new channel image, $C = \Omega(I)$ where a channel C is a corresponding per-pixel feature map (thus preserving overall image layout). Typically, the standard approach

constructs a feature pyramid by computing features, $C_s = \Omega(R(I,s)$, where $R(I,s)$ denotes an image I, resampled at every possible scale, s. In contrast, by utilising and extrapolating feature computations carried out at a coarsely sampled set of scales, the fast feature pyramids scheme as illustrated in Figure 4.8(b) efficiently speeds-up the process of computing the features across scale-space images by using the following feature approximation:

$$C_s = R(C,s) \bullet s^{(-\lambda_\Omega)} \tag{4.8}$$

Specifically, this approach benefits from the findings by Ruderman & Bialek (1994) which discovered the ratio of expectation $E[*]$ of image statistic $\phi(I)$, computed over two images captured at scales s_1 and s_2 respectively depends only on the ratio of s_1 / s_2 and follows a power law:

$$E[\phi(I_{s1})]/E[\phi(I_{s2})] = (s_1/s_2)^{(-\lambda_\phi)} \tag{4.9}$$

which further implies that every statistic ϕ will have its own corresponding λ_ϕ. We refer the reader to the original work of Dollár et al. (2014) for extra technical details on this. In summary, in the context of our work for channel feature scaling, it can be concluded that for any type of channel feature Ω (*e.g.* intensity, gradient *etc.*), there exist their own corresponding λ_Ω which can be determined empirically. Hence, supposed that we have computed a channel feature $C = \Omega(I)$ for an image I, we can now approximate the channel features at a new scale s as illustrated by Eq. (4.8).

Following the implementation of Dollár et al. (2014), a pair of $I_s = R(I,s)$ and $C_s = \Omega(I_s)$ is computed for a sparse set of $s \in 2^{(0:(nScales-1)}$; $nScales = 10$. Next, at intermediate scales, approximate features were computed as in Eq. (4.8) *e.g.* $C_s \approx R(C_{s'}, (s/s') \bullet (s/s')^{(-\lambda_\Omega)}$ where s' $\in \{1, \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^{(0:(nScales-1)}}\}$ is the nearest scale for which we have $C'_s = \Omega(I_{s'})$.

We use the same approximation principles to reproduce a multiple-size detector (*i.e.* of different aspect ratio) from the detector learned on fixed-size image windows.

We rescale either the height or the weight of the original classifier (while kept the other side in its original dimension) and re-estimate the features learned for each of them. This is to model the statistical knowledge of the abdominal shape as shown in Figure 4.7 (b). We create another 10 pair of detectors alongside the original fixed-sized classifier with the dimension of the window size is upscale to approximately 0.2% from original window dimension.

Finally, we performed multi-scale detection using all the detectors on each feature layers generated using the method discussed earlier.

Stride scale: Stride refers to how coarse or fine the step taken by each classifier is, in evaluating subsequent image windows in multiple scale-space analysis. A stride scale of 1 pixel means each classifier will scan and evaluate every possible image windows while stride scale of size 4 pixels means that each classifier will have a much coarser scan and skipping image windows for every 4 pixels. As it suggests, the larger the stride, the faster the evaluation would be, but trading with a likely loss in detection accuracy.

Detection merging via Non-Maximum Suppression (NMS): As mentioned, we used sliding window over multiple scales. Typical non-maximal suppression (NMS) was used to suppress multiple nearby detections and return the highest score bounding box. The top-score detection E_{TS} will suppress any predicted ellipse region with much lower detection score E_{LS}, if the area of overlap (defined as ratio of intersection over union) meets the following criteria:

$$Overlap\ (E_{TS}, E_{LS}) = \frac{area\ (E_{TS} \cap E_{LS})}{area\ (E_{TS} \cup E_{LS})} \geq 0.5 \tag{4.10}$$

Finally, we provide comparison on the effect of varying the spatial scale factor and stride in terms of computation time spent and the accuracy of the algorithm.

4.4 Results and Discussion

We first discuss cross-validation experimental results for designing the classification model in Section 4.4.1 which include observation and analysis on feature selection. Subsequently in Section 4.4.2, we assess the soft-cascade and recalibration architecture especially in the context of US image analysis for achieving the target miss rate, MR_{opt} within a specific target time, T_{opt}. Next, in Section 4.4.3, we test the learned classifier on full-frame validation images to find the optimal setting for the final localisation system which take into consideration stride and scale as well as suppression of multiple detections. Finally, we demonstrate a qualitative and quantitative analysis on the unseen test images given the designated model with verified selected parameters in Section 4.4.4.

4.4.1 Design Parameters

Following the evaluation methodology explained in Section 3.2, we summarise performance in the following subsections using the log-average miss rate *(MR)* evaluated at 25 evenly spaced rates between the reference point of 10^{-3} and 10^{-1} *FPPW*.

4.4.1.1 Windows and Padding Size

Window size: We fixed the padding size to +10 pixels and tested several window sizes. Figure 4.9 (a) shows that a smaller-size patch performs worst, with a patch of size 60×60 pixels and 80×80 pixels scoring 2.95% and 2.42% miss-rate respectively. Also observe that a patch of total size 120×120 pixels (100×100 base patch and 10 pixels padding) scored the best performance at 1.57% miss-rate (*MR*).

Padding size: Next, we fixed the abdomen to a size of 100×100 pixels and tested the impact of extra padding. As presented in Figure 4.9 (b), having some information surrounding the object helps the object/non-object decision with 8% improvement

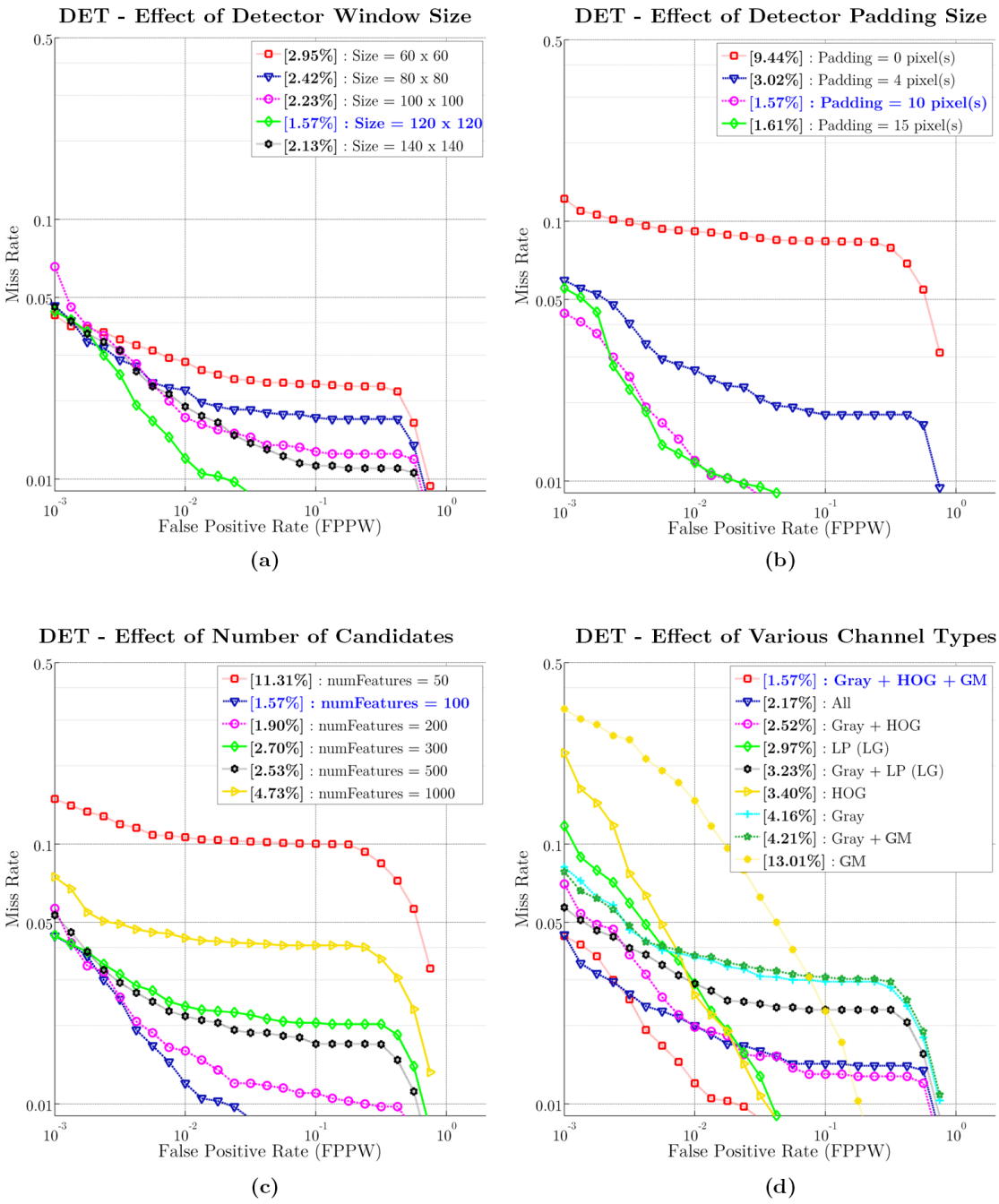


Figure 4.9: Evaluation of several principal designed parameters. Results show the log-average miss-rate performance over *FPPW* for (a) Window size, (b) Padding size, (c) Feature types, and (d) Number of randomised feature candidates. Legend entries in bracket show the average miss-rate.

between non-padded patches as compared to when 10 pixels were padded across all dimension of the patch. This finding further confirms that a non-padded patch is sensitive to boundary effects. There is no significant visual difference between 10 and 15 pixels padding suggesting that adding more padding pixels does not

improve the performance instead it becomes much worse. It can be concluded that adding more pixels will only introduce greater and diverse statistical variance of the surroundings pattern, hence does not improve the result. Larger window sizes also indirectly mean greater storage requirement and processing time needed to process a patch. Throughout the rest of the experiments, we used a patch of window dimension 120×120 pixels (inclusive of 100×100 abdominal region and 2×10 padding) for performance comparison.

4.4.1.2 Feature Parameterisation

In this subsection, we present experimental results of several feature evaluations which include randomisation factor, types of features as well as several internal feature properties such as orientation selection and channel smoothing.

Feature Randomisation: As demonstrated in Figure 4.9 (c), feature randomisation plays a role in optimising the classification performance. We varied the number of features randomly selected per node and found that 100 features performed best at 1.57% *MR*. The study also confirms that less selected feature *e.g.* 50 features, would not be able to model a good classifier while adding more features only injects uncontrolled randomisation effects which in turn impacts the classification performance.

Feature Types: We present a full feature comparison in Figure 4.9 (d). In this experiment, we compile the result for features with already optimised properties while subsequent discussions on relevant properties of some of the features will be followed afterwards.

As presented in the comparison, the gradient magnitude features (GM) alone perform the worst individually with 13.01% *MR* while the grayscale feature (Gray) alone score 4.16% *MR*. Oriented gradient based features such as local phase features (LP) and histogram of oriented histogram features (HOG) score a better performance at 2.97% and 3.40% *MR* respectively. Extended features such as a combination

of gray, gradient magnitude and HOG features which mixed several heterogeneous information such as intensity and gradient information (*i.e.* edges/oriented contours) show that it possess enough richer representation of the abdomen classification by scoring the least *MR* at 1.57%. Though, it is an interesting pattern to be analysed that the local phase (LP) features did not appear to work well with other features *i.e.* grayscale feature. This probably shows that more likely there is only a handful of them are significant while the rest do not uniquely add much information. The result demonstrates that the feature selection is being dominated by other channel features.

We performed further inspection of the best feature combinations. We plot the percentage of time each feature is selected per weak classifier (*i.e.* 3 stumps decision) in Figure 4.10, with an additional plot of the first hundred, the middle hundred and the last hundred of the weak classifiers. The features selection pattern shows that there is no dominant feature type at least up to the first 50 weak classifiers. This indirectly indicates that the feature chosen at the initial stage were designed to reject any easy negative patches while ensuring that none (or very minimal) of positive patches drop early on. As the number of weak classifiers increases, HOG feature-based classifiers are more often being selected.

Also highlighted in Figure 4.10 (b) is which feature appears most in the top node (*i.e.* topmost weighted feature). Our findings further confirm that the oriented gradient type features dominate the top-node feature selection. On the other hand, despite the gradient magnitude scoring the worst individual performance at 13.01% *MR*, the feature heatmap suggests that a few of this feature convey some important information which improves the overall performance of the algorithm.

These results make sense given our dataset, since given an arbitrary view or unconstrained object orientation, contours or edges are a rich source of information for a coarse classification task. This further confirms the importance of features carrying information of an oriented gradient such as HOG in informing the *objectness* of an image. Further, the analysis shows that other channel features such as information of grayscale and gradient magnitude features, are consistently showing up as information

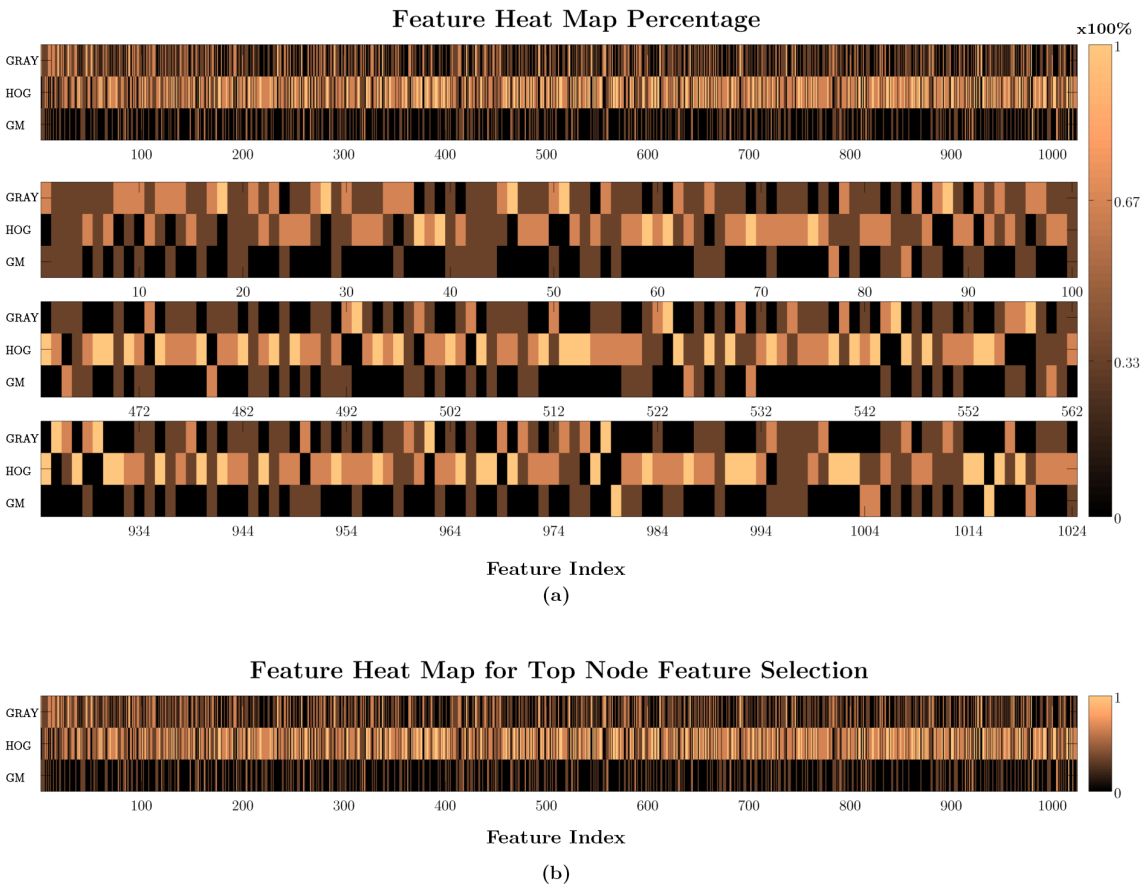


Figure 4.10: Feature heat map of the best 1024**-weak classifier model consisting of grayscale, HOG and gradient magnitude features.** The heat maps show: (a) Percentage of each feature over total number of decision (*i.e.* 3 stumps) for a total of 1024 weak classifier (top) and the first, the middle and the last hundred features for better visualisation (the following heat maps). As example, if all stumps contain only single feature type, the respective feature will attain a score of 100% and the rest will score 0% while in another case, each feature will score 33.33% if each of them was selected once per weak classifier. (b) Probability of each feature being selected at the top node of the weak classifier *i.e.* most weighted feature (denoted by 1 if selected and 0 otherwise)

rich for classification, which again, makes intuitive sense, as edges alone do not define image content in ultrasound images.

Internal feature evaluation: Estimation of some features such as HOG and Local-Phase require setting internal parameters. Figure 4.11 illustrates quantitatively how performance varies with the choice made.

i) **HOG types**: Using block normalisation in producing HOG features manages to achieve almost 0.5% *MR* reduction. The normalisation offers a more

effective elimination of false positive image windows at higher detection rate as shown in Figure 4.11 (a).

ii) **HOG orientation bins**: As the number of orientation bins increases from 2 to 6 orientation angles, we observe a reduction of up to 63% *MR* as shown Figure 4.11 (b). While increasing the number of bins means that there will be more detail on the information of orientation angle for the respective gradient, it also means increasing storage requirements. Hence, we used 6 bins in our model as this performance seemed to lead to a stable model.

iii) **Local Phase filter based analysis**: We compared the performance of local phase (LP), feature asymmetry (FA) and feature symmetry (FS) at multiple scales (see Figure 4.11 (a)). The channel features were generated using different filters such as Gabor, log-Gabor (LG) and Gaussian-derivative (GD) filters where the spatial frequency scales are given within the bracket. Log-Gabor filtered channels performed the best compared to other filtered channels with a score of 2.97% *MR*. Additionally, we compared the performance of single scale and multiple scale LP channels using a log-Gabor filter with spatial frequency scales of $[10, 15, 20]$ pixels. Results are shown in Figure 4.11 (b) in which multiple scale combinations of LP, FA and FS top the results with 2.97% *MR*. Given the nature of a sliding window approach which analyses information at multiple ratios and scales of the object of interest, the process of selecting the best spatial frequency scale for LP based analysis is getting more complex. Certain frequency scales might work best given certain size of image windows but fail given others. This leads to a suboptimal performance especially when combined with other features such as shown earlier.

iv) **Channel smoothing**: Using the best channel feature combination as experimented earlier, $\Omega I = [\Omega_{GRAY}; \Omega_{GM}; \Omega_{HOG}]$ we studied the effect of channel smoothing, both pre and post-smoothing on the cross-validation

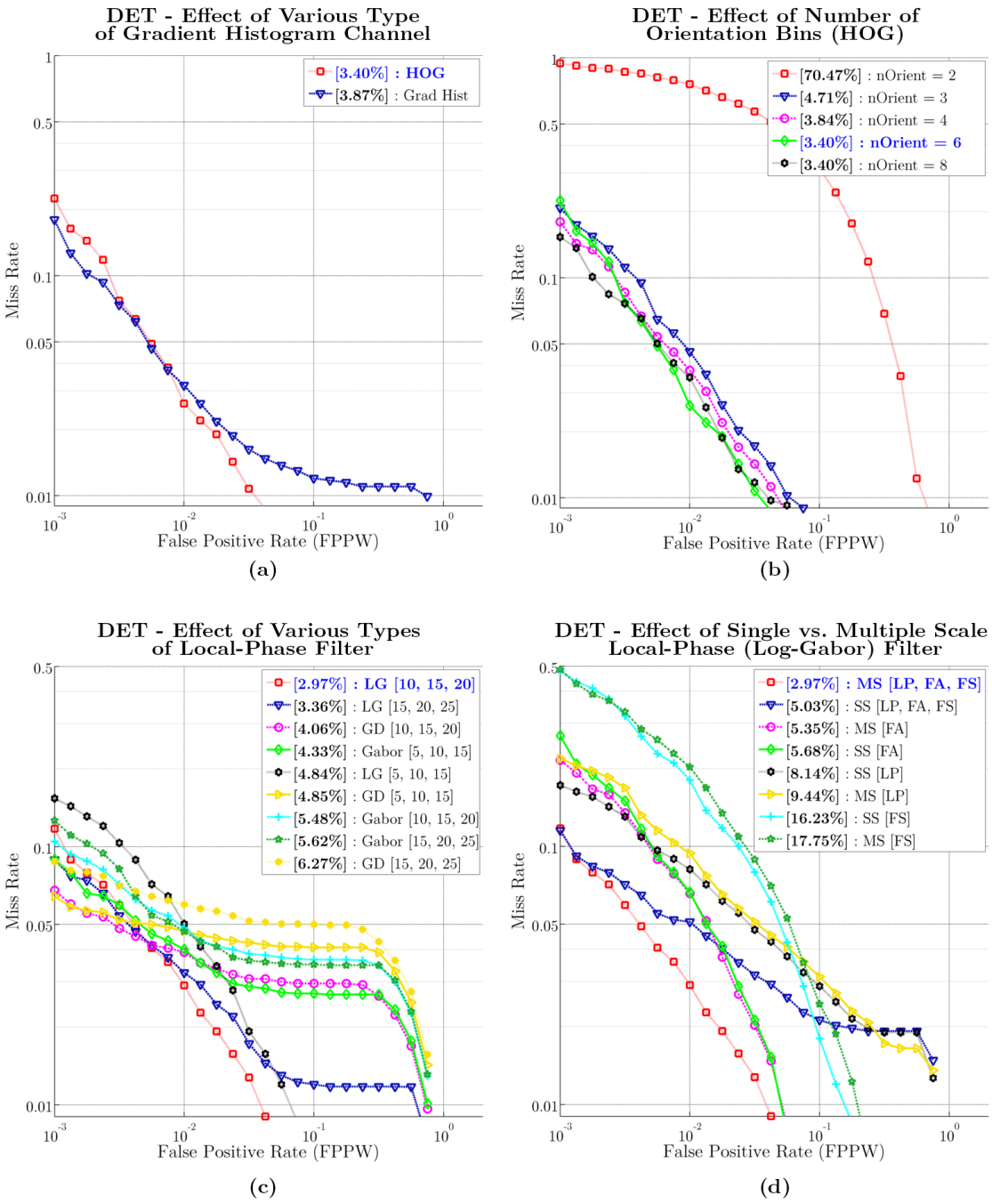


Figure 4.11: Evaluation of channel features properties. The log-average miss rate performance over FPPW for multiple channel properties; (a) Different HOG types, (b) Number of orientation, (c) Types of LP filter and (d) Spatial frequency scales. Legend entries in bracket show the average miss-rate.

dataset. In the original settings of the ACF description, smoothing refers to pixel normalisation using information from a local neighbourhood. There are two types of smoothing; pre-smoothing and post-smoothing. In our case, pre-smoothing refers to the normalisation process done to the image

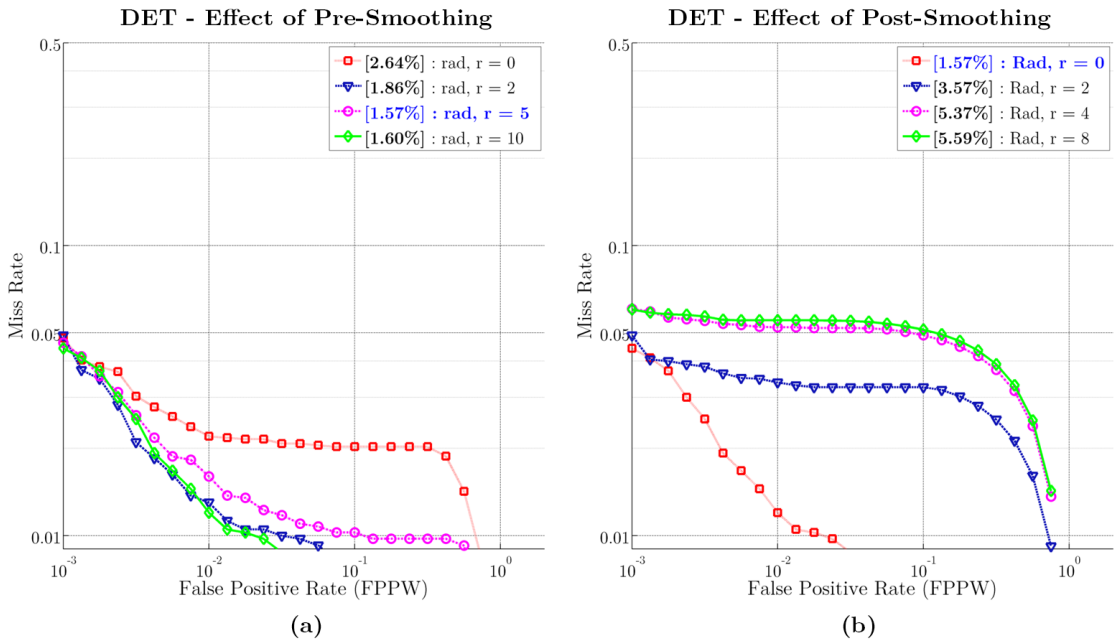


Figure 4.12: Evaluation of pre and post-smoothing effect on the performance of channel features. The log-average miss rate performance over FPPW for (a) pre-smoothing, (b) post-smoothing where rad, r is the radius of the smoothing filter. Legend entries in bracket show the average miss-rate.

before feature computation being executed. Post-smoothing refers to when the process was applied on the extracted features instead.

As shown in Figure 4.12, pre-smoothing is demonstrating more impact compared to post- smoothing. This is because pre-smoothing helps to suppress fine-scale structure as well as noise as the radius increases, though too much suppression (*i.e.* a large radius) can destroy some important information. On further inspection, gradient information, which has been shown to be very important in determining the *'objectness'* of the image is strongly affected by post-smoothing. In our experiments, we empirically found that a combination of a pre-smoothing filter with radius, $r = 5$ pixels and no post-smoothing filter gave the best performance.

4.4.1.3 Classifier Settings

Number of Weak Classifier: We studied the effect of varying the number of weak classifiers ranging from 4 to 2048. Figure 4.13 shows the average miss-rate for various number of weak classifiers computed over a cross-validation set and visualised on a

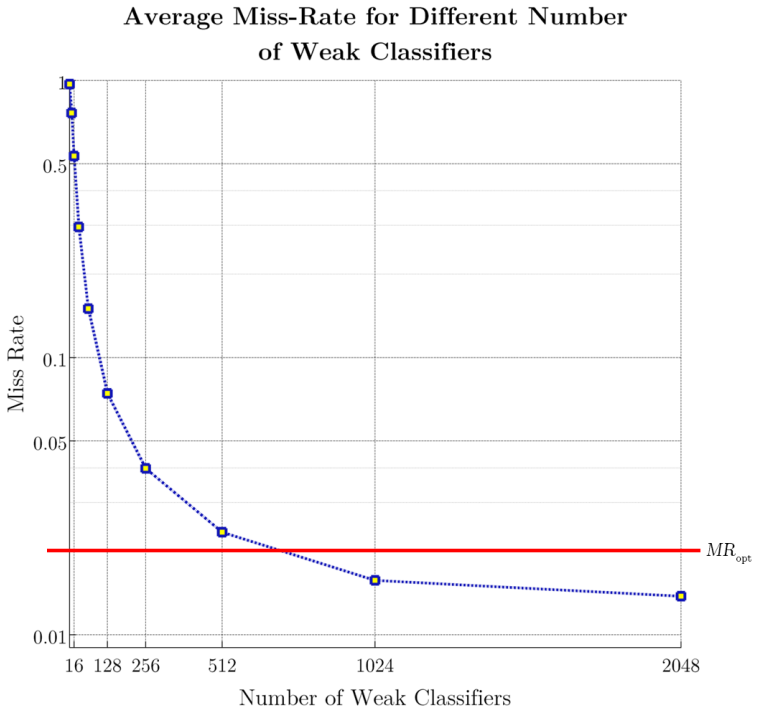


Figure 4.13: Comparison of different number of weak classifiers in the boosting decision tree. We vary the number of weak classifiers per boosting tree model from 4 to 2048 weak classifiers and measure the log-average miss-rate performance. (Note: Graph is plotted in a log-log scale. Target miss-rate, MR_{opt} is marked with the red-line)

log-log plot. This experiment illustrates that adding more classifiers (or trees) appears to improve performance. However, the performance begins to saturate and show little improvement with a larger number of classifiers. While more classifiers offer improved accuracy, this is at a cost of slower detection (as there will be more nodes to be evaluated per image window).

A model consisting of 512 weak classifiers scored 2.37% *MR* which is close to the target miss-rate, $MR_{opt} = 2\%$. Adding more weak classifiers, up to 1024 classifiers improved the performance slightly by reducing the *MR* up to 0.77%. There was a further 0.19% reduction for a model with 2048 weak classifiers. Further elaboration on the selection of the best number of weak classifiers with respect to the computation time is presented next.

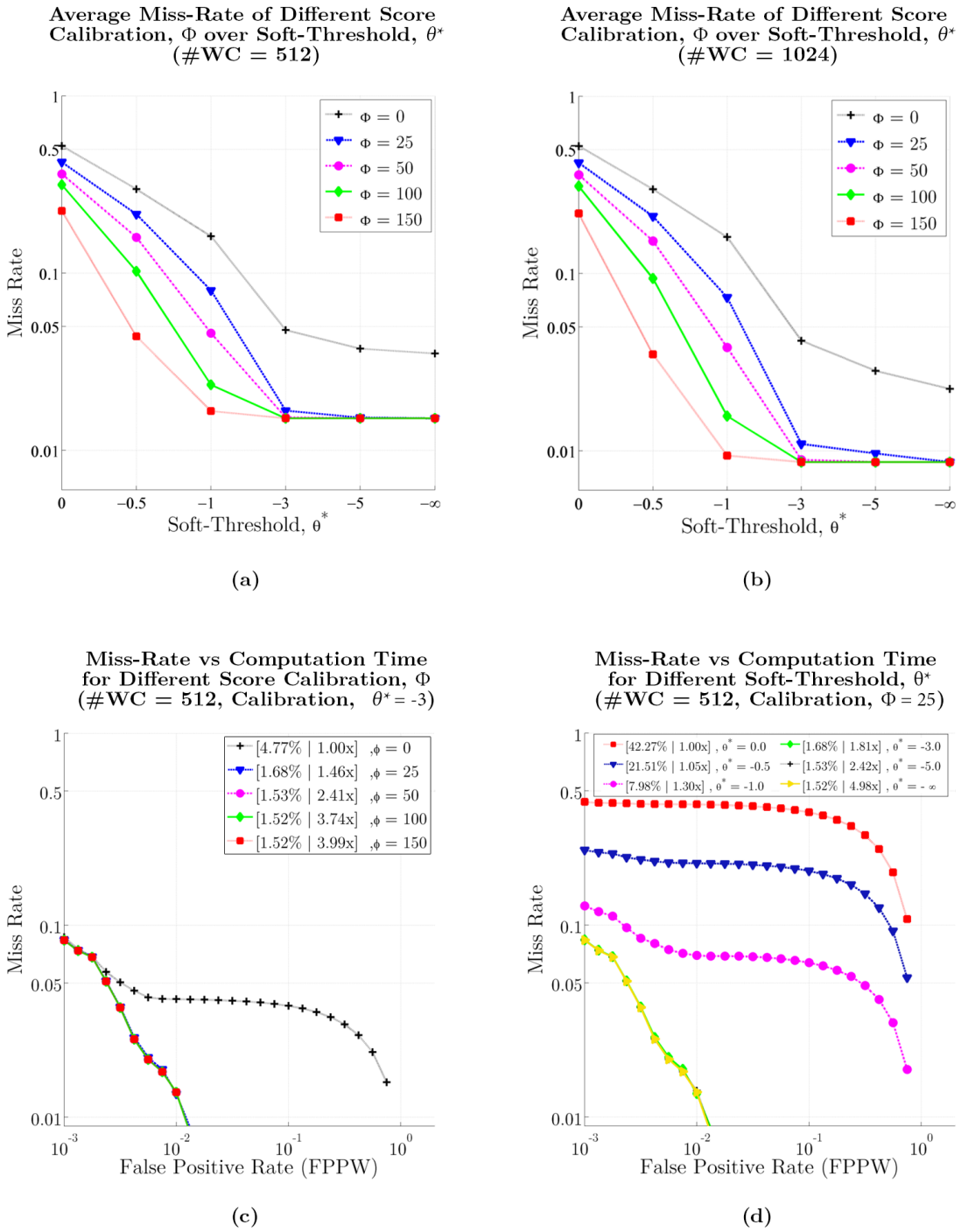


Figure 4.14: Evaluation of the impact of various calibration parameters ϕ and soft-threshold rejection values, θ^* for different number of weak classifies over computational time. Results in (a) and (b) show the log-average miss rate performance for various calibration parameters over various soft-thresholds values for classifier with 512 and 1024 weak classifiers respectively. Taking a model with 512 weak classifiers, we then plot the log-average miss-rate performance for selected parameters while also illustrating how it affects the execution time. (c) Log-average *MR* performance for (c) various calibration parameters with threshold values, $\theta^* = -3.$, and (d) various threshold values with calibration parameters, $\phi = 25$. Legend entries in bracket show the average miss-rate and computational time ratio [*MR*/computational time ratio] obtained from recalibrating the relevant parameters. Non-calibrated classifier [$\phi = 0$, $\theta^* = 0$] is used as reference.

4.4.2 Calibration and Soft-Cascades Architecture

Up to this point, we have discussed experimental results with a fixed calibration parameter, $\phi = 100$ and constant rejection threshold, $\theta_* = -1$. In Figure 4.14, we illustrate the relationship between the miss-rate performance and execution time given various calibration and soft-threshold settings.

First, it is demonstrated in Figure 4.14 (a) and (b) that a model with more weak classifiers performs slightly better. However, having said that our target miss-rate is $MR_{opt} \approx 2\%$, a model of 512 weak classifiers appears adequate. Figure 4.14 also illustrates that the performance converges to the desired *MR* with $\theta^* = -3$ for most calibration parameter settings. Figure 4.14 (c) shows that increasing the calibration parameters, ϕ, up to 150, will increases computational time by almost 4 times when compared to the case where no recalibration is applied.

The soft-threshold parameter, θ^*, plays an important role in providing a high recall rate as shown in Figure 4.14 (d). Reducing the rejection values mean more image windows need to be evaluated, and hence offers much better candidates and performance. Despite improving the *MR*, the computational time is greatly increased by almost 5 times in the case where all weak classifiers are evaluated (*e.g.* $\theta^* = \infty$).

We report the computational time taken over various combination settings of calibration parameter and soft-threshold rejection values in Figure 4.15 and Table 4.2. From these results, we choose to have a model with calibration parameter, $\phi = 25$. and constant rejection threshold, $\theta^* = -3$. A model with 512 classifier score 1.68% *MR* in 0.56 secs/frame while the system with 1024 weak classifiers score 1.09% *MR* in 0.76 secs/frame. This evaluation is executed and measured on a single-core system of an Intel Core i7 $-$ 870 computer with 16GB RAM. Hence, much faster computation time can be done using a GPU or multi-core processing system.

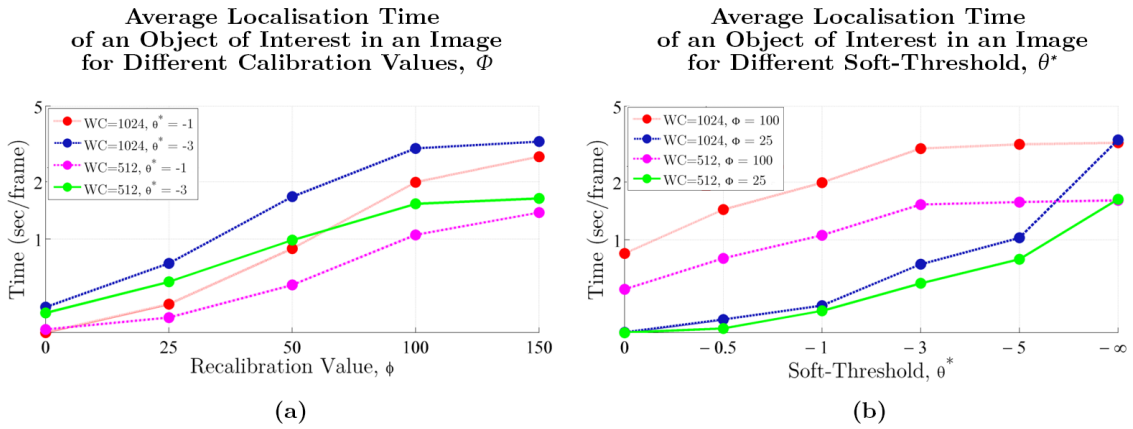


Figure 4.15: Evaluation of execution time (sec/frame) for various calibration parameters, ϕ and soft-threshold rejection values, θ^* on a single-core machine. Results in (a) and (b) show the average execution time (sec per frame). Increasing the value of both parameters will consume more times (see text for details).

Table 4.2: Time taken (sec) and average *MR* performance comparison for various parameter setting. Data shows average time taken to complete full image evolusion.

Calibration parameter, ϕ		25				100			
Soft-Threshold, θ^*		0	−1	−3	$-\infty$	0	−1	−3	$-\infty$
Time (s) [MR(%)])	#WC 512	0.33 [42.27]	0.43 [7.98]	0.56 [1.68]	1.63 [1.52]	0.55 [31.49]	1.06 [2.34]	1.53 [1.52]	1.61 [1.52]
	#WC 1024	0.33 [42.12]	0.45 [7.34]	0.75 [1.09]	3.33 [0.86]	0.85 [31.07]	1.99 [1.57]	3.00 [0.86]	3.22 [0.86]

4.4.3 Scaling and Striding Effects

Given a full size test image, we used a sliding window over multiple scales with a step size of n-pixels and 10 discrete scale step sizes, s as discussed in Section 4.3.3.2. Typical non-maximal suppression (NMS) was used to suppress multiple nearby detections and return the highest score bounding box. As formulated in Eq. (3.6), detection is considered a true positive (TP) if the detection overlaps more or equal to $THR_A = 80\%$ over the ground truth.

Next, we provided cross-validation results to study the impact of having a multiple-size detector of various aspect ratio as well as the steps taken. We then concluded the

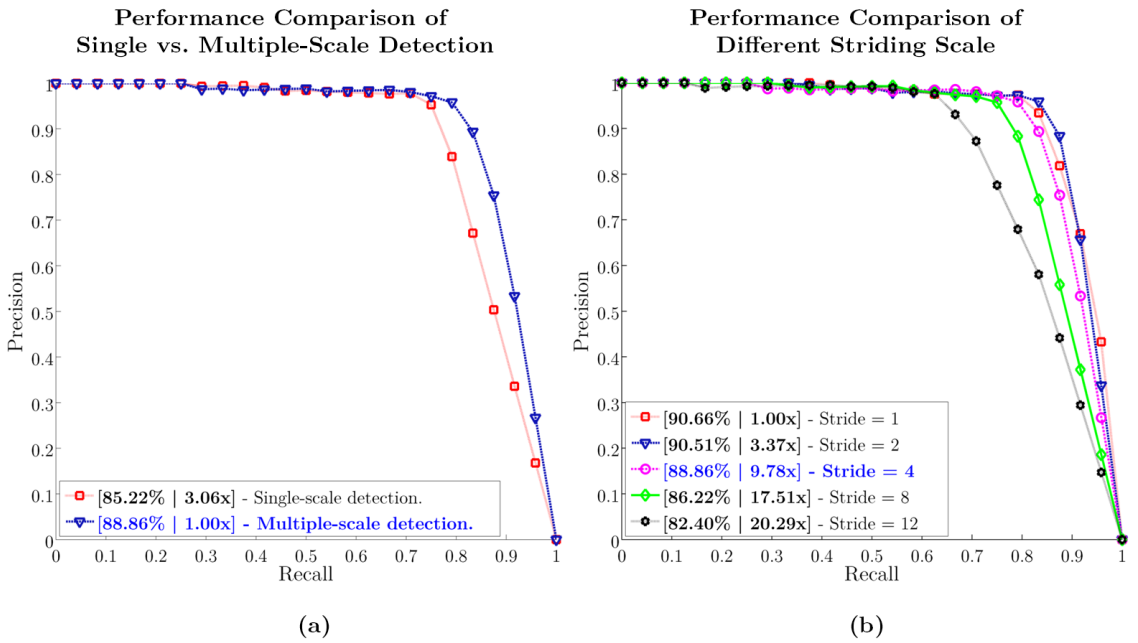


Figure 4.16: Evaluation of spatial and striding scale effect on full frame image. Results in show the precision-recall performance for (a) single vs. multiple-scale detection, and (b) various striding scale (pixels). Legend entries in bracket show the average precision and speed-up factor [*MR*/speed-up].

Table 4.3: Average time taken (sec/frame) and Average Precision (AP) performance comparison of various scaling and striding options for cross-validation set. Data shows mean average precision and average time taken per frame to complete full image evaluation in cross-validation set

Detector Scaling Type	Striding (n—Pixels)	Time (sec/frame)	Average Precision (%)
Multiple-Scale (variable aspect-ratio)	1	5.50	90.66
	2	1.63	90.51
	4	0.56	88.86
	8	0.31	86.62
	12	0.27	82.40
Single-scale (fix-ratio)	4	0.18	85.22

performance of the algorithm on the unseen test images which include 360 positive images and 400 negative images.

As shown in Figure 4.16 and Table 4.3, the average precision (in %) and computational time taken (secs/frame) were computed to summarise the performance of the algorithm on the impact of single versus multiple-size detector and striding steps taken while performing sliding window classification.

It is obvious that taking coarser steps will speed-up the classification process as there will be less windows to evaluate. As we aim for a fast algorithm, a stride of $n = 4$ is found to be adequate to provide a satisfactory performance under a second mark. Using the same step size, utilising the multiple-scale detector is shown to gain over 3% improvement at a cost of 3 times slower detection rate than using a single fixed-size classifier.

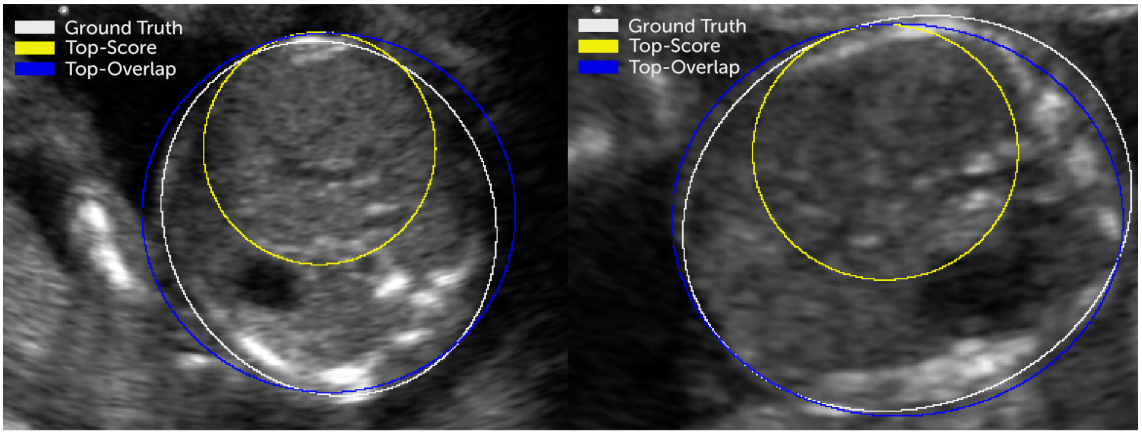
4.4.4 Evaluation of Abdominal Localisation Framework

Upon finalising the designed parameters for our abdominal localisation module, we applied the framework on a set of 760 unseen test images. The test set consists of 360 positive images and 400 negatives images, with a full dimension of 280×375 pixels for each image. In this assessment, we simply picked the top-scoring candidate as our final detection (after suppressing other detections via NMS) before evaluating against the overlap threshold criteria as described in Eq. (3.6).

We also provided a simple comparison on the anticipated outcome for different scenarios, such as top-scoring versus top-overlapped detection, as illustrated in Figure 4.17. A summary on relevant information with regard to the number of images belonging to possible detection classes was provided as well.

4.4.4.1 Quantitative Analysis

As commonly practised in the work on object detection problem, the area of overlap between the predicted region and the ground truth should meet the minimum 50% for a prediction to be considered a correct detection. In medical image analysis and especially in the context of our work, it is significantly crucial to record a much better quality detection. One of the means to ensure a better prediction quality is to set a higher threshold overlap criterion. We first ran a simple experiment on our positive test images by varying the overlap threshold to assess its effect on the quality of the detection. As illustrated in Figure 4.18(a), the recall rate records a respectably good performance beyond the 50% marks but drops dramatically after the 80% mark.



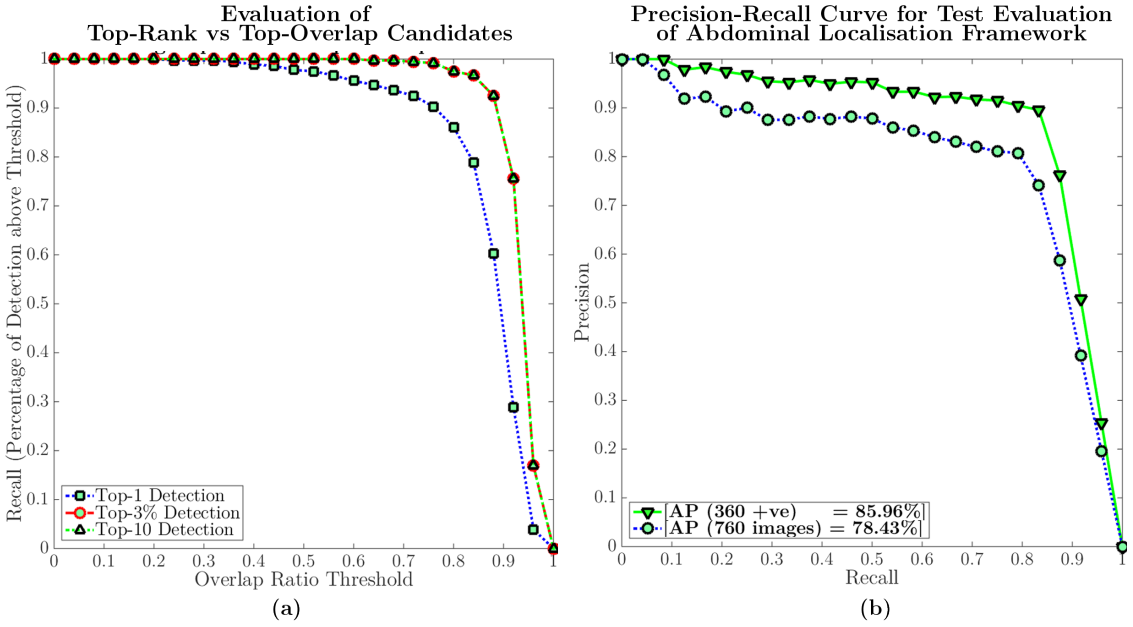
Rank 1 - Score: 34.69 | Overlap = 44.30 %
Rank 2 - Score: 34.48 | Overlap = 87.10 %

Rank 1 - Score: 27.56 | Overlap = 38.57 %
Rank 6 - Score: 25.10 | Overlap = 91.45%

Figure 4.17: Illustration of Top-Scoring Detection versus Top-Overlapped Detection. Given an unseen image, the abdomen detector (*Module B*) will localise the best location of an abdomen based on the detection output which has the highest classification score (Rank 1). As shown here, we are interested to investigate into cases where the highest-scoring candidate (in *yellow*) is considered as a false candidate as it recorded a low overlapping score to the ground truth (in *white*). We are keen to understand how much we could possibly achieve if we are able to improve the system to select the highest-overlap candidate (in *blue*) which in most cases, has slightly lower score to the the top-scoring detection as illustrated here.

In addition, we also sought to study the maximum recall bound, for example, if given the top-overlap candidates instead. **Top-overlap detection** refers to a detection with the most overlapped area to the ground truth but was ranked lower for its slightly inferior classification score. Here, we varied the overlap threshold (refer x-axis) and counted the true positive detection given two scenarios for comparison: 1) considering top-overlap candidate which recorded a score within 3% from the top-scoring detection; and 2) given top-10 ranked candidates from the final detection list prior to suppression. Both cases had similar performance and were superior compared to the top-ranked detection.

Two important observations could be drawn for further analysis and put into consideration whilst designing our fine localisation module at a later stage: 1) A decent tune to the candidates within the top 3% detections can produce a superior performance compared to simple selection of the top-scoring object. Retaining few top candidates and re-scoring them in subsequent analysis stage would perhaps re-correct the top-score detection decision, hence improving the overall performance;



Summary of the object localisation

Items	Number of images
True positive (TP)	**308**
False positive (FP)	**146**
.. Insufficient overlapping ratio	42
.. Detection on negative samples	104
False negative (FN)	**10**
True negative (TN)	**296**

(c)

Figure 4.18: Quantitative assessment of abdominal localisation framework on the test set. (a) We executed a simple experiment on the positive images to assess the quality of our detection by varying the overlap-threshold criteria (intersection-over-union, *IuO*) Given three different scenarios, we recorded the recall rate. Rather than implementing the commonly used 50% *IuO* criterion, it is observed that the system still manages to record a respectable precision score given a much higher 80% overlap threshold. Added to that, we are also interested to study how good is the top-overlap candidates from the top−3% or the top−10% detection list (*i.e.* retain the candidate that scores within 3% and −10% range from the top-scoring candidate). This serves as indicator of the maximum recall bound that could be achieved if the most overlap candidate is selected as the correct location instead. (b) Using 80% overlap threshold, we compare a precision-recall performance for the positive images (360 images) versus the whole 760 test images. An average precision (*AP*) of 85.96% which can be achieved while considering only positive images drop to 78.43% *AP* when the classification includes the negative images. This highlighting the challenges that exist in the US image analysis. (c) A summary of the result achieved in (b) was presented by highlighting number of images belonging to relevant detection classes.

and 2) 90% recall rate is recorded at 90% overlap threshold. A highly-overlapped detection exemplifies a more accurate localisation.

Next, using a detection overlap threshold $THR_A = 80\%$ over the ground truth, the designated abdominal localisation framework demonstrated an average precision of 85.96% on the dataset containing only positive images, as illustrated in Figure 4.18(b). However, the performance dropped to 78.43% given a full set of 760 images. This is mostly contributed by 104 cases of FP detection on negative images.

While this coarse detection module was designed primarily as the first stage towards achieving a fast and efficient abdomen detector with the target for a higher recall rate in return, it is anticipated that additional analysis can be performed at a later stage (*i.e. Module C* will be proposed on top of this module as extension work in Chapter 6). Hence, it is vital to explore cases where images are classified as false positives (*FP*). Out of 146 *FP* cases, we have 42 positive samples where these top-scoring detections record an unsatisfactory overlap to the ground truth The remaining cases involved negative samples that were falsely identified as containing an abdomen. Most of these 42 images (as shown qualitatively in Figure 4.19) have the best detection candidates that met the 80% overlap requirement among their top-3% scoring list. We also noted that most of the falsely identified negative samples had a detection score just above the threshold, which might be simply removed given a proper feature in later modules.

4.4.4.2 Qualitative Analysis

Figure 4.19 provides a qualitative analysis of our localisation results on a test set. Samples are sorted in three different rows where each row represents a different scenario. The first row shows 5 examples of highest-scoring positive detection. Most of the highest-scoring samples possess a perfect circular shape with overlap accuracy over 90% except the third examples which have an overlapping score of 87.50% and look slightly elliptical. In most cases there are other highly-overlapped candidates among the top−10 scores for each sample as highlighted in light blue.

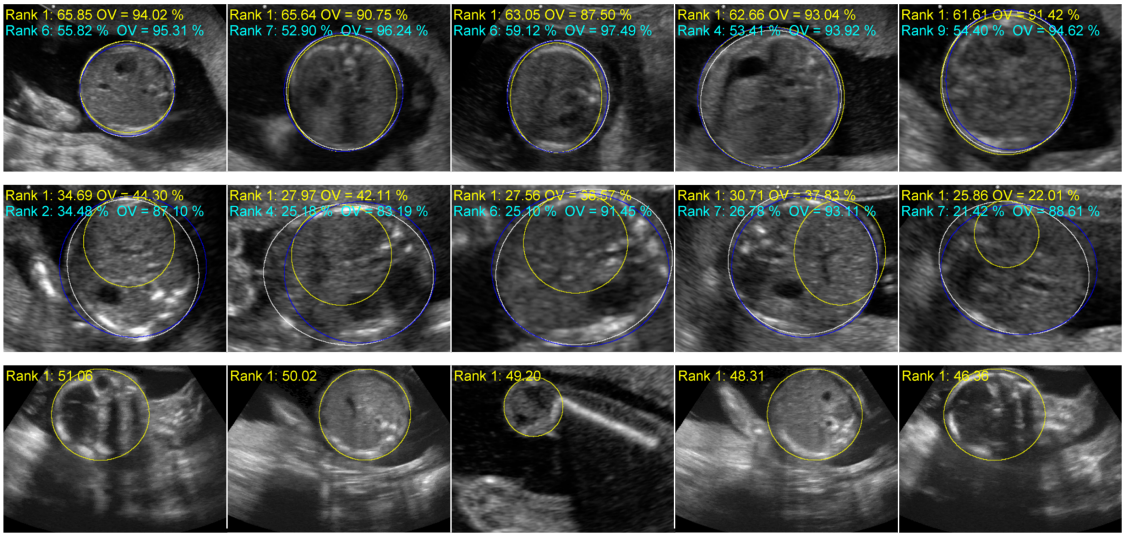


Figure 4.19: Qualitative analysis of test set. Results show localisation results with ground truth annotated by white ellipsoid marker and highest-score detection in yellow. In the case that the highest-ranked detection is not the most-overlapped detection (in comparison with the top 10 candidates), we offer an insight on how close the most-overlapped detection with the ground truth (depicted in blue). Ranking of each detection together with scores and overlapping percentage are given on top-left corner. First row illustrated the top 5 true positive detections (TP). The remaining rows show false detection (FP) cases. The middle-row show cases where overlapped region of the top-score detection on positive samples is less than overlap threshold given while the examples in last rows indicate top-5 negative samples.

The second row illustrates some typical samples of failure cases by which the top-ranked detection demonstrates insufficient overlap to the ground truth. In most cases, the abdomen is found to occupy a large space of the viewing field and to look slightly elliptical in shape. The false detection is found to rely weakly on strong boundaries which is part of the abdominal structure. Note that there exists a better overlap of candidates at much lower-rank (*top-overlap detection*) which were neglected for their slightly lower score. The last row presents typical falsely identified samples from a negative test image. In most cases, the 'abdominal-like' object possesses a circular shape with a strong presence of edges in its boundary. Some of the samples are even of an abdomen (column 2 and 4) but were regarded as negative images as either the stomach or the UV were missing.

It could be observed that given a training image of arbitrary orientation, edges and boundaries were the most prominent features learned from this coarse localisation stage. Thereupon, this finding suggests that there should exist a fine localisation

module afterwards which could analyse the pattern within the region of interest and further refine the decision on whether the object detected is abdomen or not.

4.5 Summary

Extensive experiments and validations have been performed to create a coarse localisation framework for an abdominal image. We have found that the oriented contour type features such as HOG carry a significant amount of information with regard to the *'objectness'* of the structures. However, the local-phase feature which was popularly employed within the US domain was surprisingly found to record a poor performance when combined with other features. This might be due to redundancy while it does not uniquely add much information compared to others.

Another important observation is that the classification of individual positive image windows often gets rejected early in the boosting algorithm architecture as a result of the first few important features being missing. Hence, model '*tweaking*' using a soft-cascade rejection threshold and recalibration approach provides a perfect combination in tolerating hard positive images in the build-up of the classifier model. By ensuring high-recall at a coarse localisation stage, this offers a promising opportunity to reconsider a list of potential candidates for further analysis and improvement. To conclude, we summarise our choice of selected parameters for *Module B* in Table 4.4 which is the first proposal module of our multiple-proposals framework.

Table 4.4: Selected parameters of our 'Proposal A' coarse localisation framework to detect an abdomen. Data shows the best performing parameters of our localisation module for performing fast multiscale detection under a secs/frame and within target miss-rate, $MR \approx 0.2\%$. This summarised values of each parameters of interests that build up our overall object detection architecture as presented earlier in Figure 4.2.

	Design Parameters	**Description**	**Best Settings**
Feature Encoding	Patch size	Normalised size of image windows which include base-patch size (containing abdomen) plus padding pixels.	$120x120$ pixels
	Padding size	Padding pixels around structure of interest.	+10 pixels
	Feature types	• Set of computed channel features used to provide the localisation • Internal feature parameters	• Gray + Gradient Magnitude (GM) + Histogram of Oriented Gradient (HoG) • HoG ◦ nBins = 6

	Channel smoothing	Pixel normalisation using information from neighbouring pixels.
		• Pre-smoothing
		• Post-smoothing
			• Pre-smoothing
			  ◦ $r = 5$ pixels
			• Post-smoothing
			  ◦ None

	Feature pool size	Number of available random features over a node in each round of training session.	100 features/node
Classifier Settings	Number of Weak Classifiers	Number of trees (of depth 2) that form a strong classifier.	512
	Soft-threshold, θ^*	Constant rejection threshold which adjusting acceptance tolerant of any image windows based on first few weak classifiers.	$\theta^* = -3$
	Calibration parameter, ϕ	Control the performance to converge into desired *MR* and within certain amount of execution time.	$\phi = 25$
	Striding scale	No of steps taken (in pixels) to evaluate consecutive image windows in an image	4 pixels

5

Segmentation-Driven Detection of Multiple Key Anatomical Structures

In Chapter 4, we presented the implementation of a sliding-window method based on fast-feature pyramid approximations (Dollár et al. 2014) to localise the abdomen in an ultrasound image. In this chapter, we introduce an original framework for the fine localisation of multiple key structures in ultrasound images. The main aim is to localise two main structures of interest, namely the stomach and the umbilical vein (UV). As a matter of fact, the existence of the stomach and the UV in a fetal abdominal image is considered vital to ensure a proper diagnostic plane.

We address the problem of class-driven object recognition, *i.e.* determining the position and the class of objects within an ultrasound image by designing a fine localisation framework in a bottom-up fashion which further instigates the following inter-related principles:

1) We reassess the multi-class probabilistic outputs of a pixel-wise semantic segmentation as a basic instrument to provide the best possible location for the intended structures.

2) We explore the idea of perceptual grouping in a post-processing framework. Building on mean probabilistic superpixels, we perform high-level grouping using bag of visual words histograms (commonly known as the Bag-of-Words (BoW) approach), in an attempt to extract coherent regions in a bottom-up fashion.

In the following sections, we begin by reviewing the problem formulation, including how the algorithm presented in this chapter will sit within the full unified multiple proposals framework. Next in Section 5.2, we describe each of the components of our proposed framework and relevant technical details. In Section 5.3, we summarise the complete setup and procedures for finding the best solution to solve the multi-class object localisation problem. Further, we evaluate our framework, and discuss the strengths and weaknesses of our approach in Section 5.4. Finally, in Section 5.5, we summarise our findings which include observable limitations of our work.

5.1 Introduction

The presence of two main anatomical markers, the stomach, and the umbilical vein is key to standard abdominal view plane definition (Salomon et al. 2006). The earliest research within this problem domain (Rahmatullah et al. 2011b), used the conventional sliding-window approach to detect these structures independently. That approach ignored the spatial relationship between both structures as two independent classifiers are learnt separately, hence making it not time efficient. Also, a bounding-box output that often simplifies the presence of the object is a simple and crude approximation of an articulated object. In particular, it does not capture the true

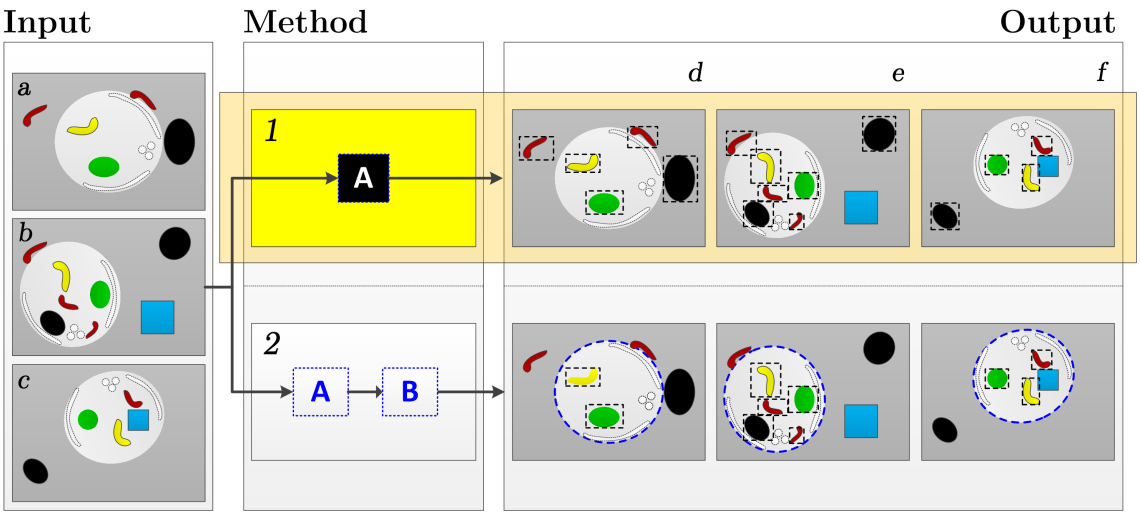


Figure 5.1: Problem formulation for fine localisation of multi-class structures via segmentation-driven framework. As highlighted in the diagram, the first-stage proposal (**Module A**) is proposed to localise multiple key structures in abdominal image, namely the stomach (*in green*) and the umbilical vein, UV (*in yellow*). Challenges include potential similar looking structures (in terms of shape and textural properties) which were represented in black and red respectively, resembling the stomach and the UV. In this work, we developed a model for proposal *Module A* and later on experimented to see how robust the method is, with and without the abdomen detector (*Module B*).

shape of the object especially for deformable objects such as the UV. In addition, the stomach and the UV structures can look similar in a fixed-size normalised patch, thereby introducing more false positive cases as highlighted in Figure 5.1.

The need to generate multi-scale sliding windows across the image leads to an inefficient computation time for the sliding window method. In contrast, a pixel-wise classification approach assigns a label to each pixel in an image based on local context around the pixel, hence preserving relations between neighbourhood pixels. Based on low-level feature extraction *i.e.* taking raw pixel values from several image statistics such as texture and intensity, this approach is usually fast to compute and offers simple class-priori information for every single pixel. However, in the case of an ultrasound image being considered here, a decision based on simple class priori information might lead to an unsatisfactory result. For example, while the presence of the stomach and the umbilical vein is generally heterogeneous with respect to an object's scale and size, it is locally homogeneous in appearance due to similar tissue characteristics. A decision made based on observation of individual pixels alone is

less meaningful and susceptible to noise, and tends to separate the object into many different regions rather than recognising the structures as a single, coherent object.

On a different note, our dataset contains a vast collection of 'objects' of varying shapes and appearances as a result of imaging conditions as well as a constantly moving target, *i.e.* the fetus. These unconstrained scenes create a challenge, which additionally makes the process of testing the presence of the intended structures under all possible imaging conditions intractable.

In short, there is a clear need for segmentation and recognition to work together. While 'low-level' features provide a good input to start with, useful higher-level image representations which combine and integrate these features and information in some way to perform object inference are important as well. Synergy between a mid-level feature and low-level pixel-based information enables one to encode high-level concepts, such as object and scene level information. ***Perceptual grouping*** — a term referring to the human ability to organise image components into coherent collections without relying on object-specific knowledge (Wertheimer 1938) is essential to facilitate fast object recognition and image analysis, hence needed to deal with this complexity.

Having established the relationship between pixels and the need to re-organise image information into meaningful assemblies *e.g.* region, and the importance of perceptual grouping in dealing with the complexity, an automated segmentation-driven detection technique based on semantic pixel-level classification and probabilistic superpixels has been developed. Using this approach, a superpixel-based Bag-of-Words (BoW) technique has been implemented to iteratively optimise a probabilistic multi-class Random Forest (RF) classification output, thus refining the location of these two key structures by combining knowledge of the neighbouring regions. The following section elaborates on each of the components mentioned.

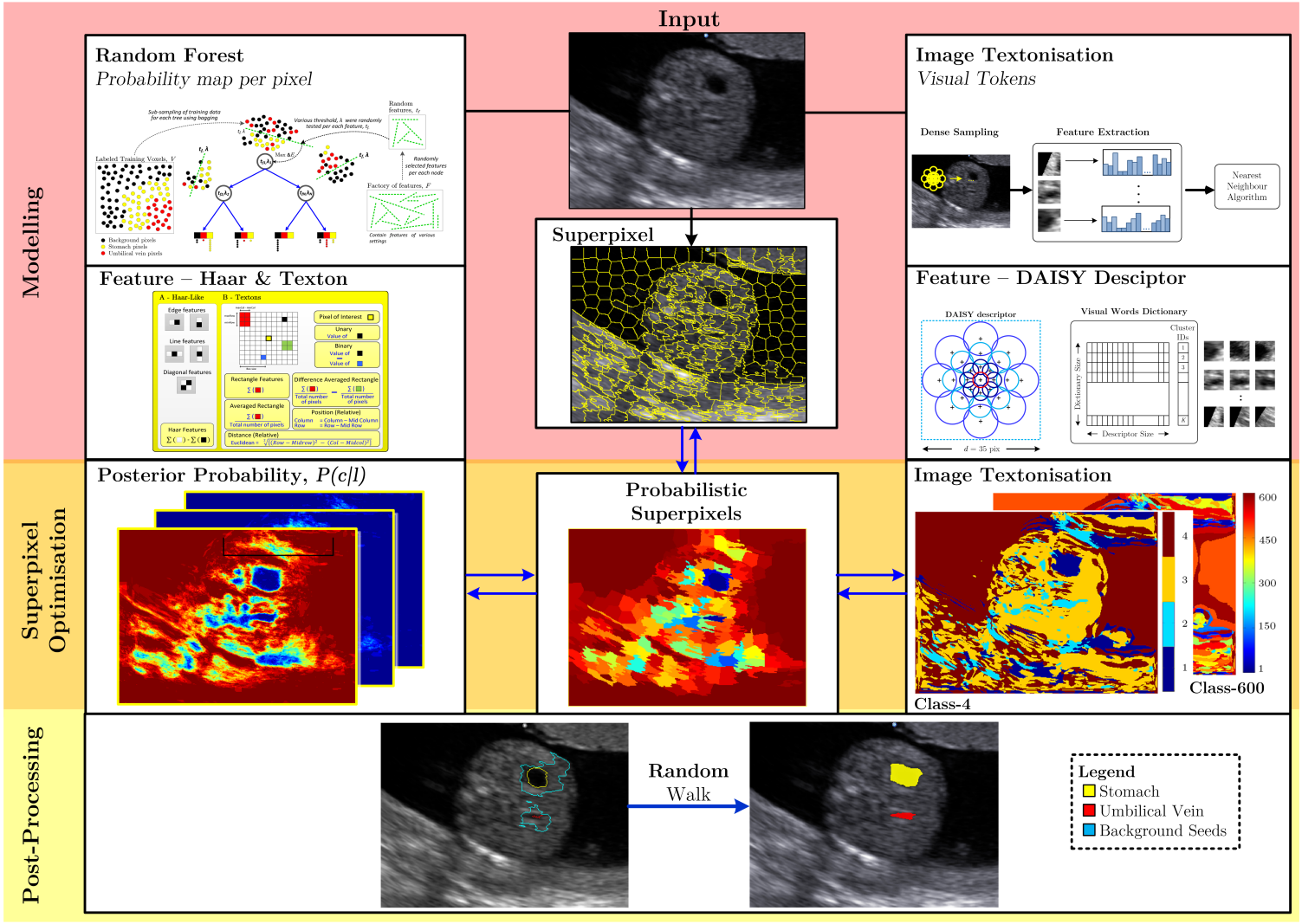


Figure 5.2: General overview of the processes involved in fine localisation of the stomach and the UV. There are three interlinked major tasks which can be jointly executed at test time. During modelling stage, a per-pixel probabilistic map based on Random Forests was built. We also constructed a vocabulary of visual tokens based on DAISY descriptor and k—Means clustering algorithm. Upon generating a superpixel map, we used these two models to develop a probabilistic superpixel. Finally, we refined the bottom-up segmentation output using a Random Walk algorithm. [See text for details]

5.2 Proposed Algorithm

In this section, we present some background and a technical review for each component of the proposed fine localisation framework. As illustrated in Figure 5.2, there are three interlinked major tasks; semantic pixel-level categorisation, image textonisation and superpixelisation, which can be jointly executed at test time. During modelling stage, a per-pixel probabilistic map based on Random Forests (RF) will be built. Section 5.2.1 will elaborate this multiclass pixel-based detector framework which provides a semantic pixel-level categorisation.

Next, we construct a vocabulary of visual tokens based on DAISY descriptor and k—Means clustering algorithm. Upon generating a superpixel map, we use these two models (RF probabilistic output and visual tokens) to develop a probabilistic superpixel. Using perceptual grouping cues such as proximity and similarity, we then combine neighbouring superpixels to highlight the probable location of the respective key structures. The whole process is described in detail in Section 5.2.2. Finally, we refine the bottom-up segmentation output using a Random Walk algorithm as described in Section 5.2.3. In summary, we present a unified segmentation-detection driven model which effectively combines both the segmentation and the recognition approaches into a single framework.

5.2.1 Multiclass Pixel-Based Detector

As illustrated in the schematics of Figure 5.3, our training module for the pixel-based classification system comprises three major components. The annotated dataset has been introduced in Section 3.1.2 (refer to page 47), while the process of feature extraction and selection is discussed next.

Feature sets: A set of strong discriminative features is required to capture all variations of appearance as illustrated in Figure 5.4. We used textons (Shotton et al. 2008) which include unary, binary, and rectangle, Haar feature sets (Carneiro et al.

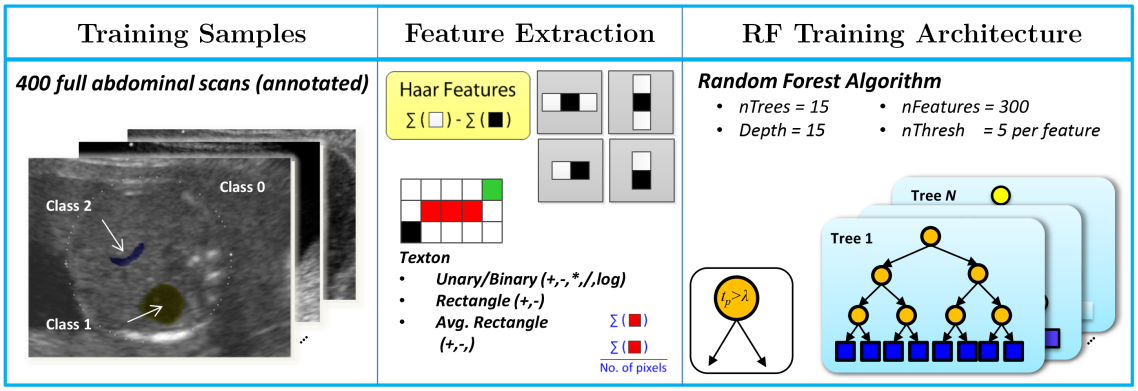


Figure 5.3: Training module for pixel-wise classification using the Random Forests (RF) framework. Selected training data were annotated and given a class label, $c_p \in p = 1, 2, \ldots, P$. Features such as Texton, Haar and others were extracted from the images and became essential input elements for the training session. Finally, a forest consisting of N number of trees and relevant parameters (*e.g.* depth, number of features, *etc.*.) was constructed and tested.

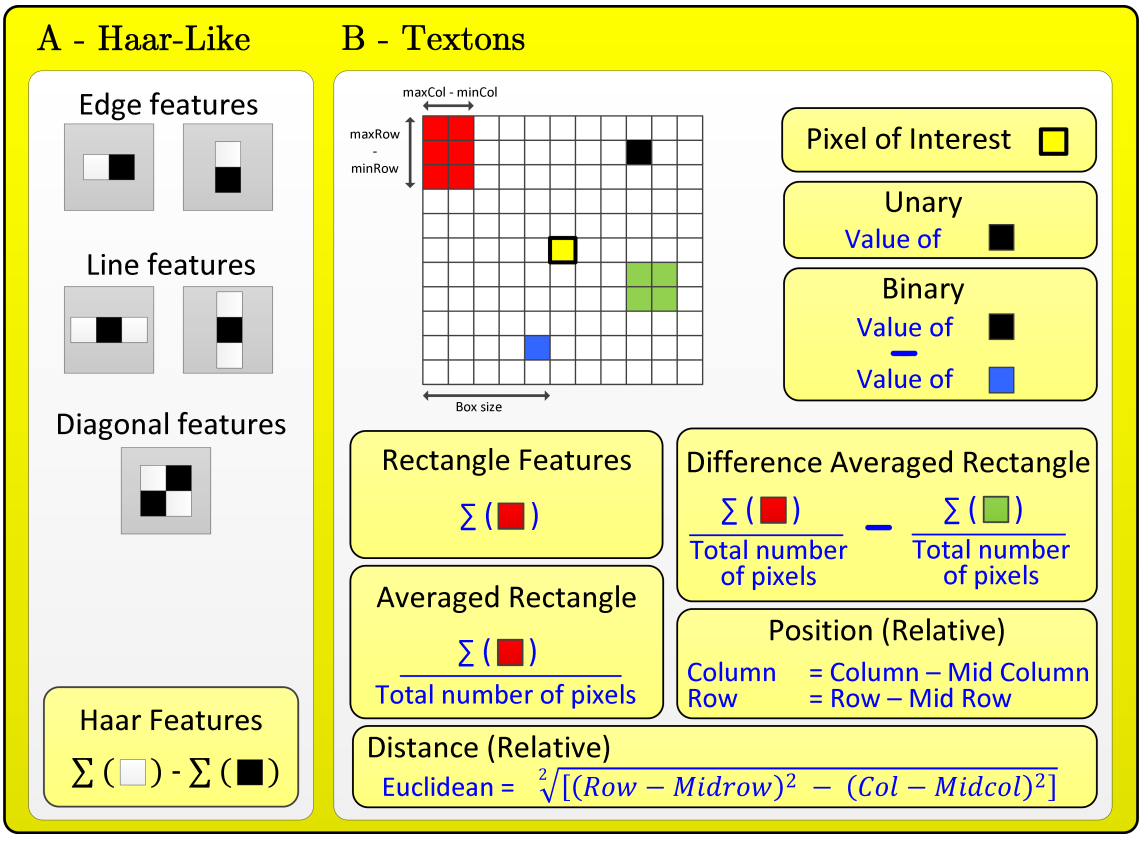


Figure 5.4: Illustration of feature sets used for Random Forest training session. The features are extracted from image patches of size $= 2 \times (2 \times \text{Box size} + 1)$ with respect to the pixel of interest.

2007, Viola & Jones 2004), and average and difference of average rectangles (Yaqub et al. 2011). The features were extracted from image patches around the pixel of interest as in (Shotton et al. 2008). These features have been used in many biomedical

imaging applications such as cellular image segmentation (Jin et al. 2014), 2D fetal cranial segmentation (Namburete & Noble 2013), 3D femur ultrasound segmentation and brain MRI segmentation (Yaqub et al. 2014).

We also introduced a relative position and a distance feature which were computed with respect to the pixel of interest in order to indirectly exploit the structural properties of the umbilical vein structures, which mostly appears as a thin-like structure compared to the stomach.

Randomised Learning: We considered randomised binary trees as in (Shotton et al. 2008), to form the decision forests with their structure and decision nodes learned discriminatively in a supervised setting. Figure 5.5 simplifies the process of feature selection per randomised tree.

In the training phase, we used a collection of labelled pixels, in the form of,

$$V_i^j = \left\{\left(I_j, x_k, c_p\right)\right\} : \quad \begin{aligned} i = &\quad 1, ..., (j \times k); \\ j = &\quad 1, ..., N; k = 1, ..., M; \\ p = &\quad 1, ..., P \end{aligned} \tag{5.1}$$

where V_i^j is the i-th training pixel at position x_k of image I_j, whereas c_p is its class label provided by an expert, which can be either 0 (background), 1 (stomach) or 2 (umbilical vein). We learned from a total of $N = 400$ training images and validated the learned model on 100 validation images. The structures that we intended to localise vary in terms of size from image to image and are quite small as compared to the number of pixels in the background region. Therefore, for each image j, we collected all non-background pixels (*i.e.* class 1 and class 2) and randomly selected background pixels which total up to the sum of foreground classes per image. Then, we normalised the total training example by the inverse class frequency as used in Shotton et al. (2008).

Trees are recursively built in a top-down manner starting from the root node. For each tree, the training data is sub-sampled using bagging with replacements (Breiman

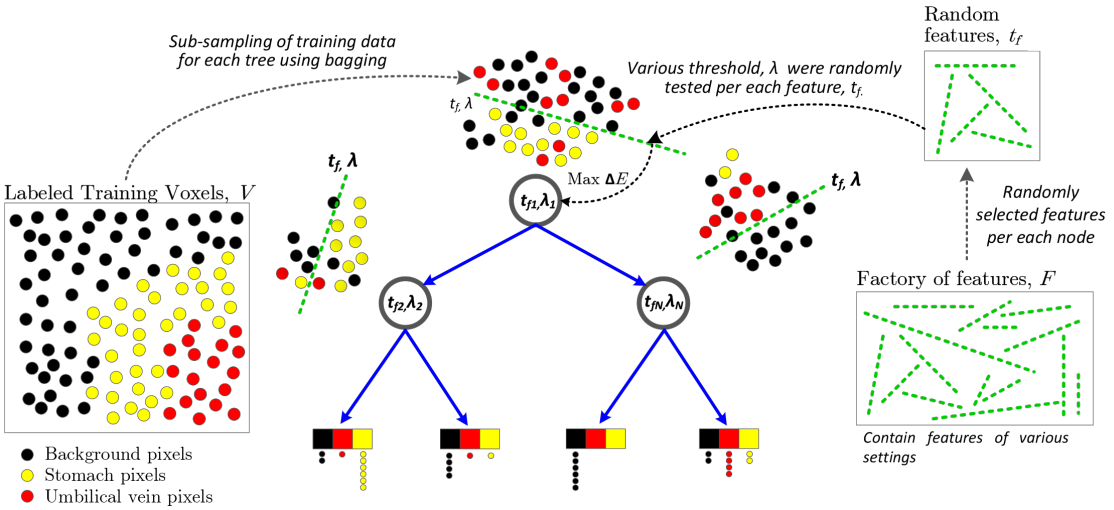


Figure 5.5: Node training in the Random Forests framework. Illustration of the independently trained decision tree within the Random Forest classifier technique. The training data for each node is randomly sub-sampled from the pool of all labeled pixels by the bagging (with replacement) process. During training of a specific node, a number of features, t is subsampled from the pool of features F, and a number of thresholds, λ were tested over the features. Both processes were executed in random manner. The pair (t_f, λ) that maximises the information gain, ΔE is selected as node-features. The leaf nodes will record the empirical class distribution.

1996). In our experiment, we drew $t_f = 300$ features randomly from the feature pool which can be different in terms of *type, dimension, size,* and *location* with respect to the pixel of interest. Added to that, multiple uniformed sampled thresholds, $\lambda = 5$, are tested for each feature. The best threshold (λ) for the selected feature (t_f),that maximised the expected gain of information, ΔE, was chosen (Lepetit et al. 2005, Shotton et al. 2008). This splits the training data, V' at node n into left and right subsets V_{left} and V_{right}.

$$\Delta E = -\frac{|V'_{left}|}{|V'|}E(V'_{left}) - \frac{|V'_{right}|}{|V'|}E(V'_{right}) \tag{5.2}$$

where,

$$E(V) = -\sum_{c=1}^{P} \frac{|V :_{class=c}|}{|V|} log_2\left[\frac{|V :_{class=c}|}{|V|}\right] \tag{5.3}$$

The recursive training continues to a maximum tree depth, $D_{max} = 15$ or until no more information gain is achieved. The class distributions $P(c_p|l)$ are estimated empirically as a histogram of the class labels c_p of the training examples p that reached node l.

Classification: The classification was performed on the remaining 360 testing images. Given a random forest, *i.e.* an ensemble of $T = 15$ decision trees that were independently trained as discussed previously, we classify a new pixel V at the position x in an unseen abdominal image I by traversing the pixel, $V = (x,I)$ over all trees.

Given a set of leaf nodes, $L = (l_1, ..., l_T)$ that the pixel ends up in different trees, we compute the mean posterior probability $P(c|l)$ as follows:

$$P(c|l) = \frac{1}{T}\sum_{t=1}^{T} P(c|l_t) \tag{5.4}$$

5.2.2 Learning to Combine Mid-Level Cues for Object Localisation

Despite the fact that Eq. (5.4) already formulates how each pixel is given a probability of belonging to certain classes, object recognition is still missing a final step where a decision on object/non-object classification are pending. The simplest way to do this is to pick a threshold value and decide the presence of an object if the probabilistic output of a pixel is above the threshold value selected. Arguably, there is no guarantee that this will generate only a single object decision with a well-aligned boundary given its hard decision on each pixel. Furthermore, this simple post-processing steps neglect the very important fact that the pixel by itself belongs to a more meaningful region.

Proximity and **similarity** are perhaps two of the most basic visual perception rules in grouping parts together into a whole. In this work, we consider a superpixel-based approach, exploiting these two key principles in providing the exact object regions. In defining the anticipated object, the concept of proximity is reflected via exploration of adjacent superpixel pairs. A superpixel (*SP*) capitalises on the local

statistics about the pixel in a given neighbourhood to reduce redundancy by grouping pixels belonging to the same object. Analysis on pairwise potentials of interconnected neighbourhood regions is vital to encode grouping relations in a build-up process of an object. Moreover, well-aligned object boundary properties of a superpixel will aid in representing accurate object shape in the image.

Encoding similarity between adjacent regions in an ultrasound image is a challenge due to various factors such as the presence of artefacts, weak boundaries between heterogeneous tissues and more importantly, diverse image quality (*i.e.* varied intensity, texture patterns, contrast etc.) due to the high-dependency on machine settings and operator capability. Hence, relying on a limited number of cues such as solely depending on low-level features (*i.e.* raw intensity value, edge etc.) to resolve bottom-up ambiguity across images of diverse quality may result in a segmentation that is overly sensitive to slight changes in image appearance. Furthermore, low-level representations are susceptible to small variations in pose and style. Alternatively, a more complex 'mid-level' representation which builds upon low-level features like bag of visual words (Sivic & Zisserman 2003) and Fisher vectors (Perronnin et al. 2010) combines and transforms low-level pixel-based information to form a richer representation that provide invariance to image deformation.

Most recent papers on perceptual grouping have emerged in the direction to provide a better object region proposal. Early work within this area include superpixel merging using low-level features (Alexe et al. 2010, van de Sande et al. 2011), region merging over multiple scales of adjacent superpixels (Arbeláez et al. 2014) and region re-ranking using mid-level features such as eccentricity and convexity (Carreira & Sminchisescu 2012).

Our approach is motivated by superpixel grouping (Rantalankila et al. 2014) which has been used to produce object segmentation proposals, coupled with the work on multi-cue mid-level grouping (Lee et al. 2015) that generates regions and resolves bottom-up ambiguity using a combination of low-level and mid-level cues. While we partly use a similar processing pipeline as (Rantalankila et al. 2014), we choose a

different seeding criterion as well as choice of low-level and mid-level features to quantify the likelihood of grouping.

We introduce a probabilistic superpixel-based Bag-of-Words (BoW) model to iteratively refine the output of the pixel-wise classification of our Multiclass Pixel Based Detector. We will now explain our framework, starting from the way superpixels are generated, then selection of features used to signify proximity and similarity, and finally the complete model.

5.2.2.1 Superpixelisation

We utilise the visual coherence of the structures by generating a semantically meaningful 2D superpixel lattice using the simple linear iterative clustering (SLIC) algorithm (Achanta et al. 2012). In particular, we use the MATLAB® implementation[1] of this algorithm (Vedaldi & Fulkerson 2008).

For an image consisting of M pixels, the algorithm takes an input of the desired superpixel number, Q, such that the approximate size of superpixel clusters is equivalent to M/Q. To produce roughly equally sized superpixels, the grid sampling size is determined as $S = \sqrt{M/Q}$. Pixels are associated to the nearest cluster centre within a region of $2S \times 2S$ via minimisation of a modified Euclidean distance metric, D:

$$D_{jk} = \sqrt{\frac{(i_k - i_j)^2}{m^2} + \frac{(x_k - x_j)^2 + (y_k - y_j)^2}{S^2}} \tag{5.5}$$

where (x_j, y_j) and i_j are the spatial coordinates and intensity of pixel j, respectively; while (x_k, y_k) and i_k are belong to the cluster centre. The compactness factor, m regulates the compactness of the superpixels where a higher values of m result in a more compact superpixel lattice, while lower values yield less compact and irregular superpixels that closely adhere to the image boundaries. Samples of superpixels

[1]VL_Slic is part of VLFeat Library Toolbox which is freely available from http://www.vlfeat.org

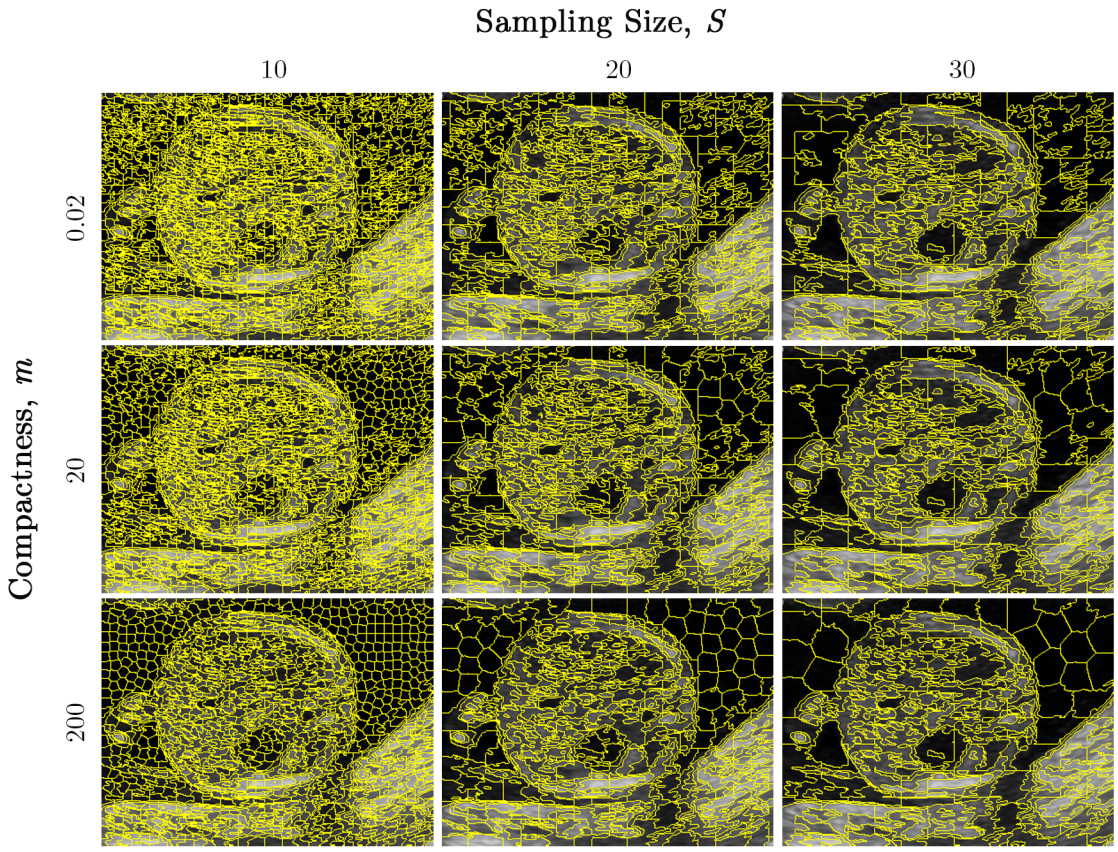
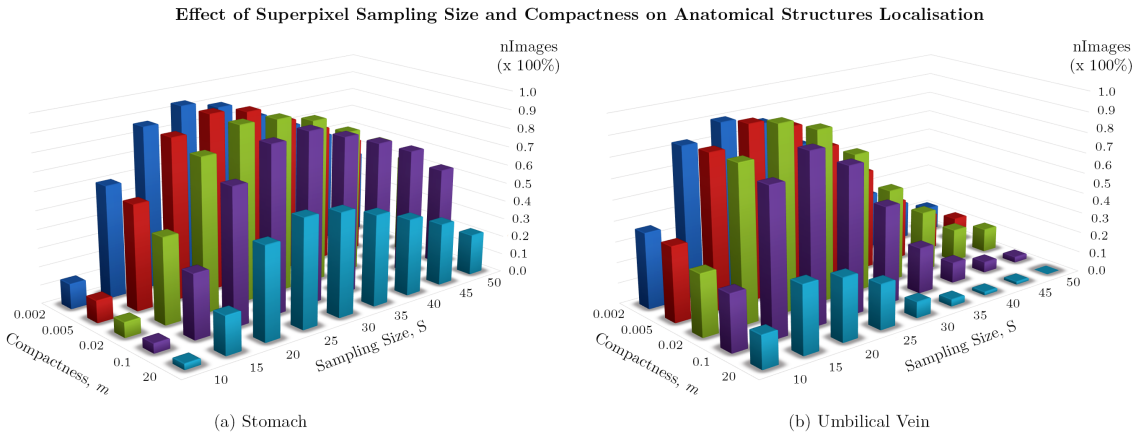


Figure 5.6: Visual comparison of superpixels produced by various combinations of sampling size and compactness. Illustration shows a more compact and regular superpixels' shape by the increasing value of compactness factor, m at a cost less sensitive to the image boundary. Increasing the sampling size, S (refer column-wise) means smaller number of superpixels is generated per image. We elaborate the impact of these selections in Figure 5.7

.

produced by various combinations of sampling size S and compactness factors m are visualised in Figure 5.6.

As mentioned, we consider a superpixel as the smallest atomic unit from which our refinement and regrouping process starts. In order to illustrate how the superpixel size and compactness affect localisation, we generated superpixels using various combinations of $\boldsymbol{S}$ and $\boldsymbol{m}$, and carried a simple test on each validation set in the training dataset. For every generated superpixels map, we observed if there was at least a single superpixel which overlaps at least 50% with the ground truth for each structure (*i.e.* Dice Similarity Coefficient ≥ 0.5). This experiment was conducted to simulate a case where the structures might be classified as a true positive case (if detected).



(a) Stomach (b) Umbilical Vein

Figure 5.7: The effect of sampling size and compactness parameters in the initialisation of key structures localisation. Histogram shows average number of images having their ground truth of (a) stomach and (b) umbilical vein overlap with any single superpixel lattice. Given an image, we generate superpixel lattices using various combination of sampling size and compactness terms. For each setting, we count and calculate average number of images which has its respective ground truth intersects with any superpixel lattice for at least 50%. We run this on a cross validation set to find out the best parameter combination to start with to initiate our superpixelisation process.

Figure 5.7 shows a histogram of the average number of validation images (over 5-fold cross validation set) which met the criteria mentioned for either structures. We learned a couple of important clues from this plot;

1) Both structures demonstrate different effective sampling sizes. The stomach performs better with a bigger-sized superpixel (depicted by higher value of S) while the umbilical vein settles for a much smaller-size, *e.g.* $S = 20$. As further observed, there are no additional gain obtained with respect to the size increment as shown by the deterioration of the segmentation overlap.

2) Accuracy of the segmentation-driven method is inversely proportional to the compactness, m, of the superpixel. A more compact superpixel performs worse, especially for the umbilical vein.

Using this as the basis of our selection, which later is confirmed via cross-validation experiments, we empirically picked $\boldsymbol{S = 20}$ and $\boldsymbol{m = 0.02}$ as the initial combination to start with for both structures. This preliminary output will then be regulated and

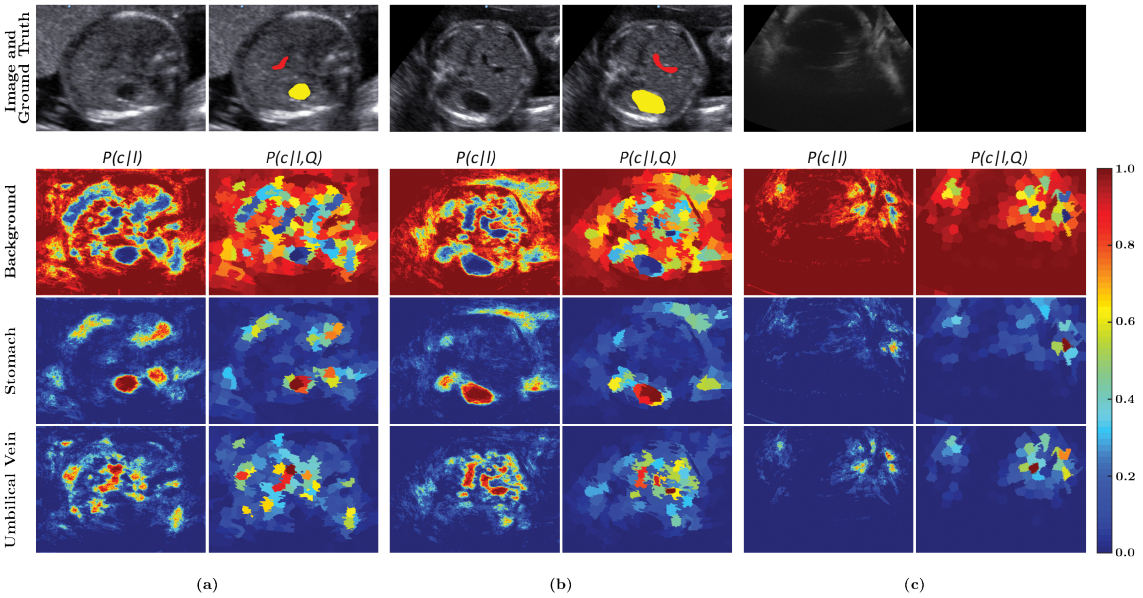


Figure 5.8: Sample illustration of mean probabilistic superpixel map for the respective classes. *Top row*: Three samples inclusive of their ground truth. *Remaining rows*: RF probabilistic output, $P(c|l)$ and its corresponding mean probabilistic superpixel map for each class, $P(c|l,p)$. Positive samples are shown in column (a) and (b) while negative sample is illustrated in column (c). Examples of positive sample demonstrated two common problems; (1) multiple candidate regions (mostly for the umbilical vein), and (2) brittle umbilical vein.

refined via a superpixel reorganisation technique with the aid of mid-level information cues as described in subsequent sections.

5.2.2.2 Probabilistic Superpixels

Using the chosen superpixel size, S and compactness factor m, a map of superpixels $p \subset Q$, was generated. The per pixel posterior probability of the RF was used to generate a ***probabilistic superpixel*** by computing, and assigning, the mean probability to each superpixel lattice p, as formulated in Eq. (5.6);

$$P(c|l,p) = \frac{1}{N}\sum_{\substack{n=1 \\ p\in Q}}^{N} P(c|p_n) \qquad ;c \in \{0,1,2\}, N \approx S^2, p_n = (x_n, y_n) \tag{5.6}$$

Figure 5.8 illustrates three different samples following the mentioned procedures. For each superpixel generated, three layers of confidence map, $P(c)$, were produced for each image corresponding to each classes available *i.e.* $c \in \{0,1,2\}$ representing background, the stomach and the UV respectively. This confidence map is then utilised

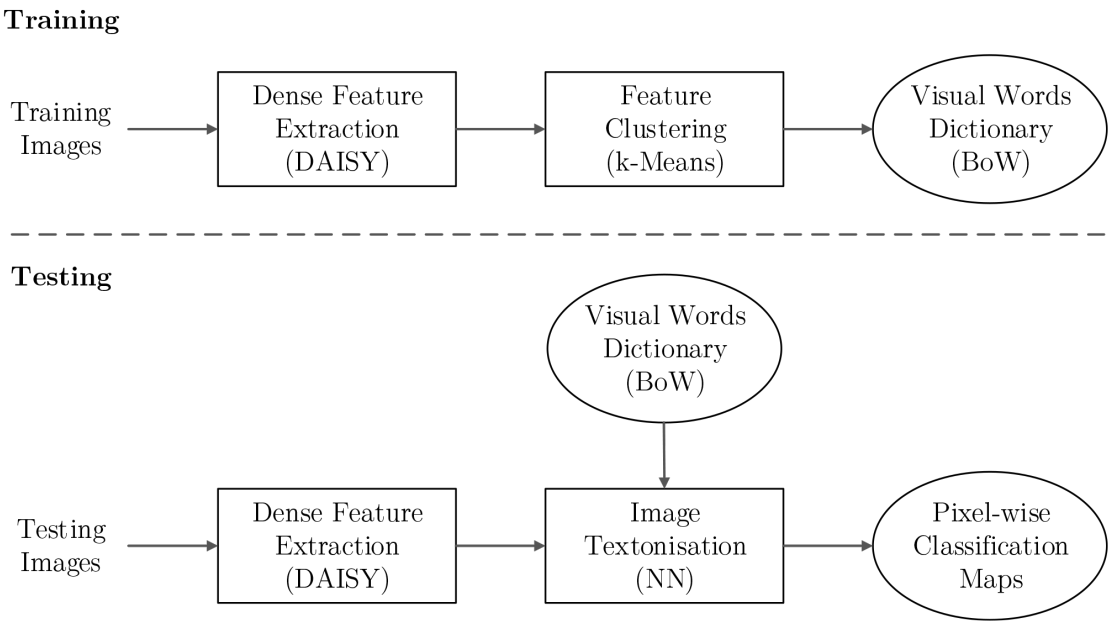


Figure 5.9: Flow diagram illustrating an example process by which a visual codebook may be constructed. In training, we estimated dense feature using a Daisy descriptor before assigning feature points into clusters via *K*-means clustering. A visual words dictionary was formed and derived from the centres upon convergence of the clusters. In the validation-testing stage, the input image was textonised and classified into respective available classes by comparing the extracted descriptor with the closer cluster centre using nearest-neighbour (*NN*) algorithm.

to provide a decision on whether an object is detected or not, via a region grouping process as explained later in Section 5.2.2.4 .

5.2.2.3 Visual Tokens - Deriving a Texton Map

Figure 5.9 presents a general diagram illustrating how features were extracted and converted into a visual words dictionary. We briefly describe the process as follows:

Training: Features were first extracted from the images. We learned from a set of 400 training images which had corresponding ground truth labels, c_p; where p can be either 1 (background), 2 (stomach), 3 (umbilical vein) or 4 (abdominal region). We extracted a maximum of 5000 patches per class from each image, where each patch, has a fixed size of 35×35 pixels and contains a labelled ground truth at the centre pixel. In the case where the total pixels/class per image is less than the maximum number of patches allowed, all pixels (for the respective class) were extracted. Next,

DAISY descriptors (Tola et al. 2010) were computed on the collected patches. The description of the feature extraction process is presented next.

Clustering was performed on the DAISY descriptors using K-means clustering to build a set of ***visual tokens***. In our experiments, we arbitrarily chose $b = 150$ clusters for each class available. We then concatenated the learned visual tokens of each class to form the final visual codebook. Hence, a total of $B = 600$ visual tokens was generated. It should be noted that a suitable degree of invariance (to rotation, scale, acquisition settings *etc.*) was learnt implicitly from the training patches. Samples of visual tokens represented in Figure 5.10(b) demonstrate the commonly occurring visual patterns for the respective structures.

Image Textonisation: Figure 5.10 visually illustrates how we built the descriptor. In our method, a given image was densely sampled at each pixel. We used the original setting of the DAISY descriptor as introduced by (Tola et al. 2008) as depicted in Figure 5.10 (a). In brief, it consisted of 3 rings, **R** (in addition to the pixel of interest) in which each ring contains 8 concentric circles centred on the location (indicated by a '+' sign) from where the features would be extracted.

For a given image, eight orientation maps, **G** were computed. Each orientation map was then convolved several times with Gaussian kernels of different standard deviation, Σ values to obtain convolved orientation maps. Let $\mathsf{h}_\Sigma(u, v)$, be the vector at location (u, v);

$$\mathsf{h}_\Sigma(u, v) = [\mathsf{G}_1^\Sigma(u, v), ..., \mathsf{G}_8^\Sigma(u, v)]^T \tag{5.7}$$

where G_1^Σ, G_2^Σ, and G_3^Σ denote the Σ-convolved orientation maps in different direction. The complete DAISY descriptor $\mathcal{D}(u_0, v_0)$ at location (u_0, v_0) was then defined as the concatenation of h vectors;

$$\begin{aligned}\mathcal{D}(u_0, v_0) = [&\mathrm{h}^T_{\Sigma_1}(u_0, v_0); \\ &\mathrm{h}^T_{\Sigma_1}(I_1(u_0, v_0, R_1)), ..., \mathrm{h}^T_{\Sigma_1}(I_N(u_0, v_0, R_1)), \\ &\mathrm{h}^T_{\Sigma_2}(I_1(u_0, v_0, R_2)), ..., \mathrm{h}^T_{\Sigma_2}(I_N(u_0, v_0, R_2)), \\ &\mathrm{h}^T_{\Sigma_3}(I_1(u_0, v_0, R_3)), ..., \mathrm{h}^T_{\Sigma_3}(I_N(u_0, v_0, R_3))]\end{aligned} \tag{5.8}$$

where $I_j(u, v, R)$ is the location with distance R from centre pixel (u, v) and $j \in \{1, N\}$. We used $N = 8$, which pointed out that circles in each rings ($R_1 = R_3/3$, $R_2 = 2R_1$, and $R_3 = 17$ pixels), were separated by $2\pi/N$ degrees. Different size of circles represent different Gaussian smoothing values which are proportioned to the radii of the circles. The Gaussian kernels used have standard deviations of $\Sigma_1 = 2.5$, $\Sigma_2 = 2\Sigma_1$ and $\Sigma_3 = 3\Sigma_1$. To summarise, having extracted features from 25 locations with each containing 8 orientation bins, a descriptor of size $8 + (3 \times 8 \times 8) = 200$ values for each image patch was defined.

Given the set of cluster centre IDs of the visual tokens, classification involves a simple process of associating each pixel location in the unseen images with the closest visual tokens, $\boldsymbol{b} \in \boldsymbol{B}$ available from the visual word dictionary where $\boldsymbol{B}$ is the total number of learned visual tokens. A window of dimension 35×35 pixels was slid across the image to generate a histogram of visual words for each position. Given a query histogram, h_a, for any pixel of interest, the closest token IDs, $\widetilde{b} = \arg\min_{b \in B} E_{L2}(h_a || h_b)$; was chosen, *i.e.* by minimising the Euclidean distance metric between the two histograms of d-dimensional length,

$$E_{L2}(h_a, h_b) = \sum_{i=1}^{d} (h_a(i) - h_b(i))^2 \tag{5.9}$$

Figure 5.11 illustrates sample texton maps of initial visual tokens IDs, $\boldsymbol{b} \in \boldsymbol{B} = 600$ (*middle row*). Later, we reassigned this initial label into their respective 4– class labels (*bottom row*). This was done by simply converting the initial label within a certain range registered in the visual codebook into one of four available class labels as seen in Figure 5.10(b). For example, visual IDs, $\boldsymbol{b} \in \{\mathbf{1,150}\}$ represent a stomach (reassigned

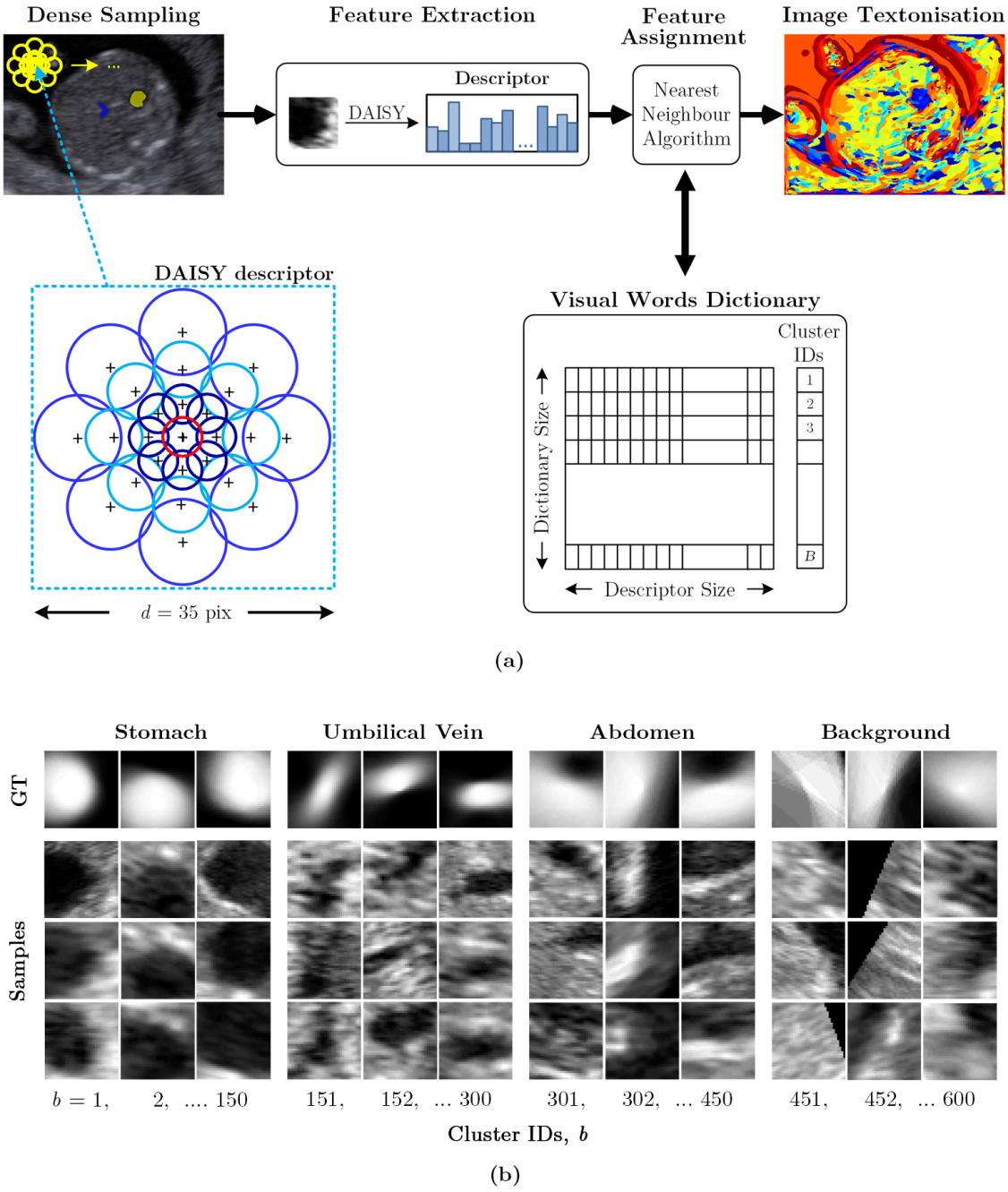


Figure 5.10: Graphical illustration of our image textonisation approach and sample patches of cluster centres. (a) Our approach involves taking an input image, densely sampling each pixel (*i.e.* pixels located centrally on a 35 × 35 pixels image patch), and translating it into a feature descriptor of size 200 words using DAISY algorithm. The extracted features are then compared with the constructed visual words dictionary and classified to one of the cluster IDs available using a nearest neighbour algorithm. (b) *Top row*: The corresponding mean of the ground truth (GT) for the respective visual tokens. *Bottom row*: Sample of clustered image patches for their respective visual token IDs.

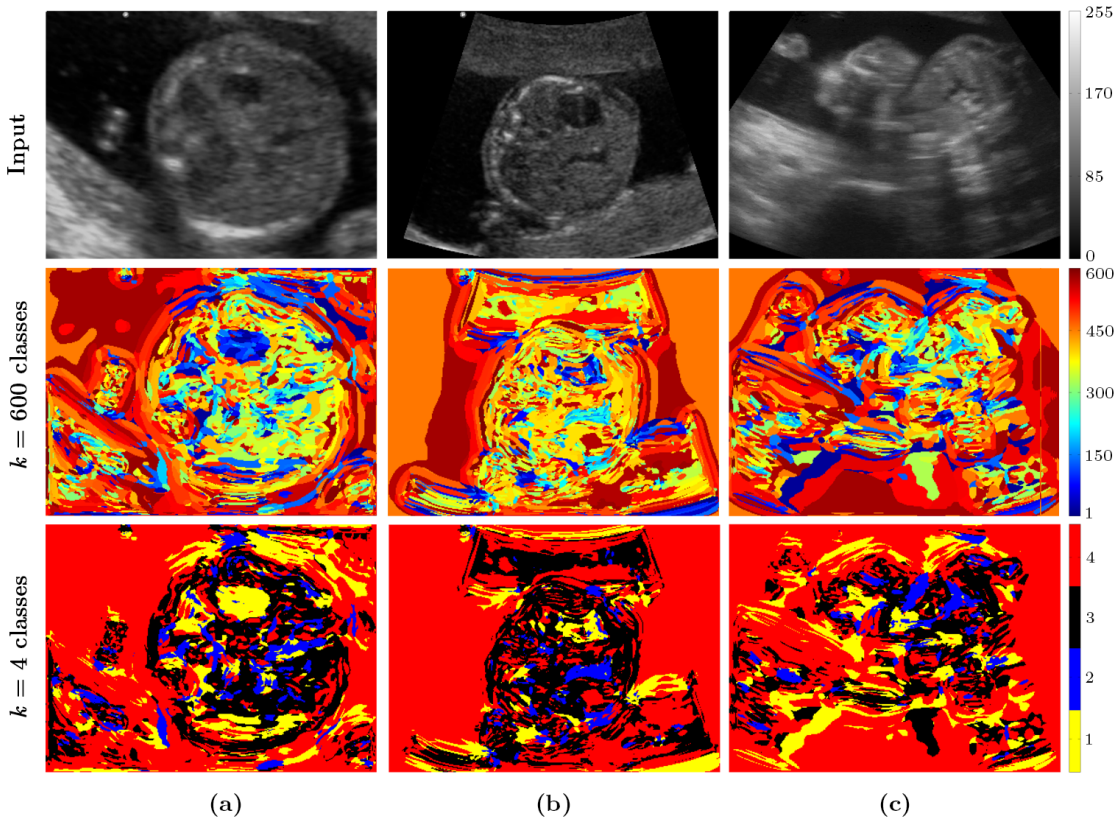


Figure 5.11: Illustration of texton maps. Different colours uniquely identify different textons (visual tokens). We demonstrated an initial 600 textons map (second row) which were later reassigned into its 4 main classes (last row). As demonstrated, abdominal images containing the stomach and the UV as in (a) and (b) demonstrate a visually good response especially to locate the possible abdominal (*in black*), the stomach (*in yellow*) and the UV region (*in blue*). Meanwhile, sample of negative image in (c) visually demonstrates a random response in which none of the structures could easily be recognised from the texton outputs. [Note: Best viewed in colour. Different colour represents different texton labels. Colour assignment is summarised by the colour bar legend on the right.

as label 1), visual IDs of $\boldsymbol{b} \in \{\mathbf{151,300}\}$ which represent the UV would be reassigned as label 2 while visual IDs of $\boldsymbol{b} \in \{\mathbf{301,450}\}$ were reassigned to the abdomen (label 3). The remaining visual IDs were reassigned to label 4 which represents a background class. The response map illustration visually demonstrates a favourable output for the positive images where structures of interest such as the abdominal region, the stomach, and the UV are visually distinct from the background. In contrast, the negative image displays an uncertain pattern response in which none of the structures could easily be recognised from the texton outputs.

This texton map alongside the probability map formed the basis of our region based mid-level cues when performing grouping at a superpixel level as explained next.

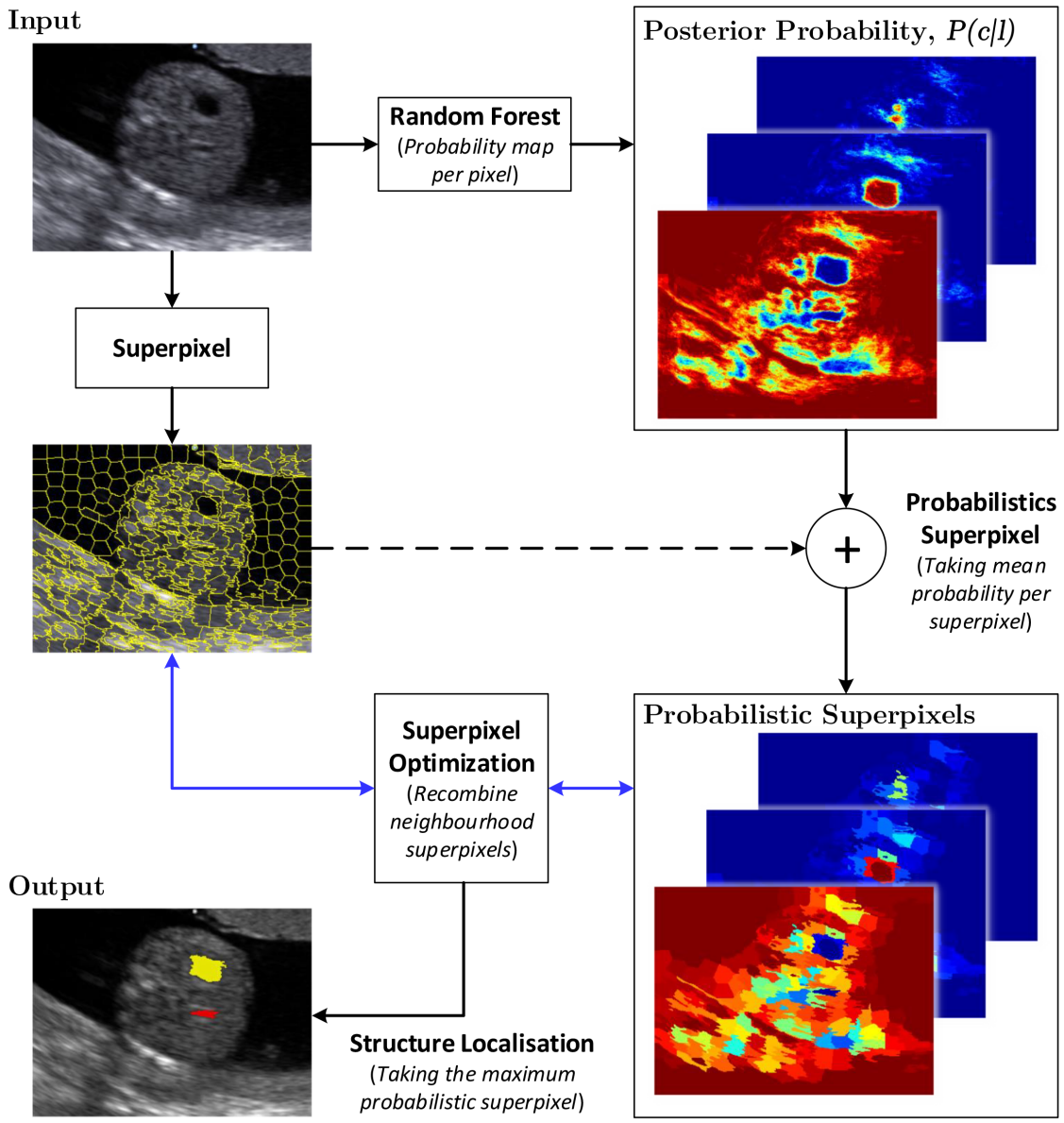


Figure 5.12: Superpixel Optimisation Framework. We obtained the posterior class distribution of each pixel through Random Forest classification and superpixel using VL SLIC [58]. We then assigned the superpixel with the mean probability before iteratively recombined the neighborhood superpixels which vary below certain thresholds. The structures with maximum mean probability are returned with their confidence level indicated by the mean probability.

5.2.2.4 Object Detection Framework via Refined Superpixels

While the posterior probabilistic output of the RF algorithm, as illustrated in Figure 5.8, and the texton maps, as in Figure 5.11 provide a delineate structure of interests to a certain extent, we still needed one more step to predict the location of each particular structure.

We went from bottom-up segmentation to structure localisation using the method summarised in Figure 5.12. We began by segmenting the input image into a single layer of superpixels. This superpixel map defines basic grouping primitives. Based on the generated superpixel lattices, object proposals were then obtained via iterative region grouping which is similar in principles to the greedy merging technique of van de Sande et al. (2011) and local search of object segmentation proposals of Rantalankila et al. (2014), except that we used different starting procedures and a diverse set of grouping cues.

Specifically, each adjacent superpixel pair is assigned a score that represents the visual similarity of the respective superpixels. Example of grouping cues that used to score the visual similarity between superpixel pair include the ***proximity*** and the ***similarity*** of visual tokens as explained next. For each structures of interest, the proposed region merging method picks the highest probabilistic superpixel as a seed to initialise the grouping process. Neighbourhood superpixels which possess a similarity score above threshold, T_{SIM} will be merged into a single superpixel. Similarity scores for the newly formed superpixels with its neighbours are then updated. Upon regrouping, the newly formed superpixel is reassigned with a new probability value according to the highest value. The merging process continues iteratively until there are no neighbourhood which observe the similarity score requirement. The regrouping process continues for the remaining region with $P(c = 0) \leqslant 0.4$ to collect a maximum 10 top proposal regions. This is done by restarting the same procedure on the next superpixel with the next highest probability score.

Seeding cues: Up to this stage, there are two possible ways to make an object localisation from the generated probabilistic superpixels:- 1) based on the highest scoring foreground probability of a superpixel belonging to either class, and 2) highlighting the probable locations based on background probability before deciding which object the region belongs to.

Our experimental observation highlighted the fact that the potential object regions mostly possess a minimum background probability score. This is strongly supported

by the background subtraction principles (Piccardi 2004) which discovered that a better object classification of the image foreground can be extracted effectively given a recognised background pattern. In contrast, given the context of our work, relying on the foreground classes alone in deciding the object localisation might produce a sub-optimal performance due to the fact that both structures of interest (*i.e.* the stomach and the UV) share quite a similar textural pattern.

This scenario will result in at least two potential candidates of comparable score and might lead to an erroneous decision. Hence, as considered here, we first used the probabilistic superpixel output of the background class ($P(c = 0)$) to select the initial location of the seed region and subsequently, to perform the regrouping process of neighbouring regions. More qualitative and quantitative comparison on either strategy will be further discussed in Section 5.4.1.3.

From the proposed region list, we then re-ranked the candidates according to their object class probability. This was done by creating a separate ranking list of the stomach and the UV based on their mean probabilistic superpixel score. The existence of the structures of interest was then finalised based on the candidates with the highest probability score, indicating the confidence level of localisation.

We will shortly describe how visual similarity between neighbourhood regions was measured, and which grouping cues were employed in providing the score.

Perceptual grouping cues: Given an image x, a compactness parameter, m and a superpixel size, S, we partitioned it into a set of Q superpixels (refer to Section 5.2.2.1 for formulation), from which the final objective was to seek a subset $R \subset Q$ that represents an object of interest. We represented R as a binary label, $y \in (0,1)^{|Q|}$, where $y_p = 1$ when superpixel p is in R, for $p = 1, ..., |Q|$, hence $R = \{p : y_p = 1\}$. By finding a labelling via optimisation of the similarity score of related pairwise potentials, we propose a region that exhibits strong grouping relations. We describe the notation and contribution of the corresponding cues as follows:

1) **Proximity:** We modelled grouping cues by optimising the pairwise potential defined over a superpixel graph. Each superpixel p acts as a node while edges represent a pairwise potential for every pair of adjacent superpixels. The basic concept of proximity is preserved by encoding group relations via these edge sets, $\mathcal{A}(Q)$.

2) **Similarity:** Regularity of the object can be represented by its similarity in appearance, where the more similar superpixels are to each other with regard to the tested elements (*e.g.* textural properties), the more likely they belong to the same object. In this approach, we defined similarity via two diverse feature sets which include a histogram of class probability, h_{PROBS} and histogram of visual token, h_{VT}.

   Firstly, a d-dimensional length histogram of background probability, h_{PROBS} is extracted. The prior background probabilities occupying the tested superpixels which implicitly encode the *'objectness'* of the region was quantised into d-bins. *Secondly*, following the visual token approach, as explained in Section 5.2.2.3, we constructed a histogram of visual token ID, h_{VT} of d-dimension, in which d represents the number of available texton classes *e.g.* $d = 4$.

   Hence, for every tested superpixel pair $(p, q) \in \mathcal{A}(Q)$, the similarity score, s^{pq} can be represented as follows:

$$s^{pq} = 1 - [\mathcal{d}(h^p_{PROBS}, h^q_{PROBS}) + \mathcal{d}(h^p_{VT}, h^q_{VT})] \tag{5.10}$$

   where $\mathcal{d}(a, b)$ is the Chi-square ($\tilde{\chi}^2$) distance between the histograms a and b.

$$\tilde{\chi}^2 = \frac{1}{2}\sum_{i=1}^{n}\frac{(a_i - b_i)^2}{(a_i + b_i + \epsilon)} \tag{5.11}$$

It should be noted that ϵ is a small positive constant to prevent division by zero and h represents number of histogram bins of the tested features. As a matter for comparison, we also provided a comparison with cues generated from intensity values. We use the same principles in generating the histogram of probabilities as described earlier. For each superpixel region, we constructed a grayscale histogram, h_{GRAY} by quantising the grayscale image into a bin of d-length, where d range from 8 to 64 bins.

5.2.3 Post-processing

***Diversification*:** In order to provide the final candidate, the superpixel optimisation algorithm as described in Section 5.2.2.4, reorganises and recombines adjacent superpixels based on their similarity in appearance. However, limiting the grouping process to only connected neighbourhood superpixels demonstrates sub-optimal performance, with under segmented structures becoming the main concern, as observed from cross-validation experiments.

UV localisation suffers the most, as the structure is disjointed by the presence of relatively small areas of different contrast (appearance) as highlighted by circle 1 of Figure 5.13 (b). As a result, the merging process was terminated before the whole part of the typically elongated structure was successfully detected (*i.e.* similarity score between adjacent superpixels drops below a threshold). This produces either a non-optimised detection (*e.g.* true positive (*TP*) cases where the detected region marginally passes the overlap threshold) or even worse, an under segmented structure (*e.g.* false negative (*FN*) cases).

In order to overcome this limitation, we propose a refinement stage where candidates acquired from the first stage are combined by utilising some object properties such as *location*, *eccentricity* and *circularity*. Pooling together proposals which observe the designated rules allows us to localise the intended structure with greater precision.

i) **Location:** Starting from the top-ranked candidate, the algorithm explores its extended neighbourhood regions that are distanced less than 20 pixels away (approximate length of a superpixel). Given two superpixel regions (p, q), we computed their similarity score s^{pq} similar to the formula given in Eq. (5.10) except that we replaced the histogram of probabilities with the difference of foreground probabilities, d^{pq}.

$$d^{pq} = \left|\left(P^{p}_{c=x} - P^{q}_{c=x}\right)\right| \qquad ; x = \max_{c\in\{1,2\}}\left(P^{p}_{c}\right) \tag{5.12}$$

At this stage, we only record the potential regions to be combined. The decision on whether to merge or not depends on the following object properties.

ii) **Eccentricity and Circularity:** For each (recorded) region pair, depending on whether the foreground likelihood is the stomach or the UV, a simple comparison test was performed with respect to the object properties as follows. For the UV, the combined structure should demonstrate higher eccentricity as compared to initial segmentation output, otherwise the grouping process will be rejected. For the stomach, we compared its circularity before and after the grouping process. Upon regrouping process, we computed the difference of a circularity score between an initial segmentation with the newly grouped object. Recombination is considered invalid if the measured circularity score exceeds a threshold, T_{CIRC}.

Figure 5.13 demonstrates how the algorithm regroups the disjointed UV structure using the stated properties.

Refined Segmentation using Random Walks: During the superpixelisation process, image pixels were grouped and partitioned, subject to the choice of the sampling size, S. Using an identical sampling size for both the stomach and the UV has remarkably offered better generality and simplicity. However, in return, the quality of

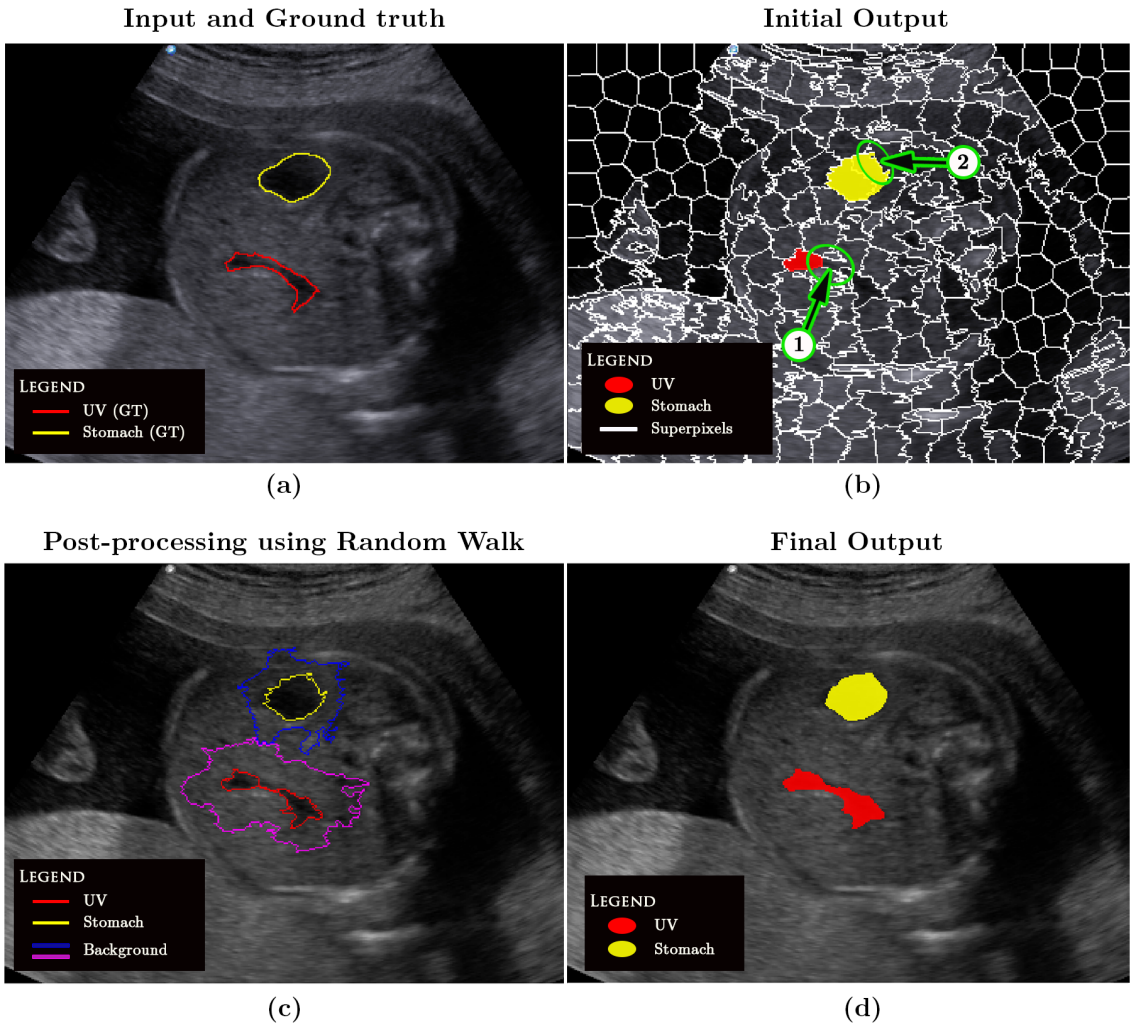


Figure 5.13: Post-processing formulation. (a) Sample of image with annotated ground truth illustrating the location of the stomach (in *yellow*) and the UV (in *red*). (b) Image was overlaid with superpixels mapping (in *white*). Initial segmentation output of the stomach (in *yellow*) and the UV (in *red*) which were obtained via bottom-up segmentation process is shown. Possible challenging scenario is highlighted in circle. *Circle 1* refers to a case where the structure is disjointed into two parts by a small area of slightly less-darker appearance while *Circle 2* refers to case where the structure is undersegmented (which mostly happen at the boundary of the structure as a result of limited superpixel sampling size, S). [Refer text for detail elaboration.] (c) Second-stage post-processing step employs a Random Walk approach (Grady 2006) to improve the regrouping process and optimise the structure's segmentation. Upon generating an initial segmentation via iterative regrouping process, the model will automatically place a respective label on the outer boundary of both structures. We then utilise the knowledge of neighbourhood regions of the localised structures (represented in *cyan* and *magenta* to signify the outer boundary of the neighbourhood regions of the stomach and the UV respectively) as a seed point of the background pixels. (d) Final segmentation output is obtained by assigning each foreground pixel to the labels. [Best viewed in colour].

initial segmentation output produced is sub-optimal. This is especially true to the pixels located close to the boundary. As a result of superpixel size limitation, these pixels are jointly grouped to the neighbouring superpixel (as illustrated by circle 2 of Figure 5.13 (b)).

We applied a post-processing step to the segmentation output by exploiting list of neighbourhood superpixels. We used the Random Walks approach (Grady 2006) to improve the segmentation quality of the detection automatically. Random Walks has been demonstrated to provide a superior segmentation output in a number of medical imaging applications (Chen et al. 2014c, Stevenson et al. 2015).

Figure 5.13(c) summarises how this algorithm could refine the segmentation quality. Upon generating an initial segmentation, the algorithm automatically assigns respective labels to the boundary pixels of the localised structures (*i.e.* class 1 (in *yellow*) and class 2 (in *red*) to the stomach and the UV respectively). This serve as a seed pixel for either structures. We then exploited the neighbourhood regions by giving the background label (class 3) to the outer boundary as demonstrated by the *blue* label (for stomach neighbouring regions) and the *magenta* label (for UV neighbouring regions).

Given a set of pre-labeled pixels as the initial seed for the respective classes, the Random Walks algorithm generates a final segmentation label for each pixel in the image. The algorithm functions by assigning each unseeded pixel to the respective seed point (*e.g.* class) depending on which label is most likely to arrive first at the unseeded pixel by a random walker.

5.3 Experimental Setup

In this section, we briefly describe the dataset and evaluation metrics used to measure the performance of the proposed algorithm.

5.3.1 Datasets

We ran 5-fold training-validation experiments on 500 full-frame training-validation images (*Dataset A1*) to parameterise the model. We then tested the model on a total of 360 positive unseen testing images (*Dataset D1*) to provide a qualitative and quantitative assessment on the performance of the proposed algorithm.

5.3.2 Evaluation

Given an unseen image, we ran our proposed algorithm to provide an object/non-object decision for both the stomach and the UV based on its confidence score. The evaluation was based on pixel-wise similarity using the Dice similarity measure, ***DSC*** as formulated in Eq. (3.10) (page 54). Structures for which $DSC \geqslant 50\%$ are considered a correct detection (True Positive, *TP*); otherwise considered a False Positive, *FP*.

Precision-Recall (*PR*) curve and 25-point interpolated Average Precision (*AP*) as described earlier in Eq. (3.7) and Eq. (3.8) (page 53) were used to report overall localisation performance. For model parameterisation, we report the mean Average Precision (*mAP*) which is equivalent to the mean of the average precision, *AP* over the *5*-fold validation results.

5.4 Results and Discussion

The results for model parametrisation are presented in Section 5.4.1. Qualitative and quantitative evaluation of the proposed model are discussed in Section 5.4.2 and Section 5.4.3, which includes the post-processing algorithm. We also assess the implementation of abdomen localisation (*Module B*) on the unseen test images to study how prior recognition of the abdomen could help bottom-up reorganisation of the multiple key structures.

5.4.1 Model parameterisation

The following experiments present and discuss results from cross-validation experiments. It should be noted that in generating the final object localisation, we did not use any knowledge of where the abdominal region location or post-processing in order to study the robustness of the proposed approach.

5.4.1.1 Superpixelisation parameters

Figure 5.14 evaluates the effect of the superpixel sampling size and compactness parameter on the localisation performance as presented in Figure 5.7. The result shows that the larger the compactness parameter (bigger m), the poorer the performance, especially for the UV. The stomach localisation attained the best *mAP* at 84.50% for $m = 0.002$ while the UV records 84.05% *mAP* for $m = 0.02$.

Due to the way a superpixel was defined, a smaller compactness parameter generates a more irregular-shaped structure. The algorithm in return will generate $2D$-lattices which are much closer to the image boundary, closely taking the structure shape. From this result, a shape's compactness can be seen to play a role in improving the localisation especially for an articulated object such as the UV. As for superpixel size, localisation of the UV shows the best performance at a smaller superpixel size ($mAP = 85.86\%$ for $S = 10$). On the other hand, the stomach localisation scores better for a slightly larger size (mAP $= 83.86\%$ for $S = 20$).

Since a general setting that works well for both structures in generating superpixel lattices was required, settings which had the best average performance were chosen for both parameters, which are $S = 20$ and $m = 0.02$. These settings were used for the rest of the experiments reported in this chapter.

5.4.1.2 Choice of Region Grouping Cues

In Section 5.2 , we described the formulation of our mid-level features such as a histogram of probability, h_{PROBS} and histogram of visual token ID, h_{VT} in addition

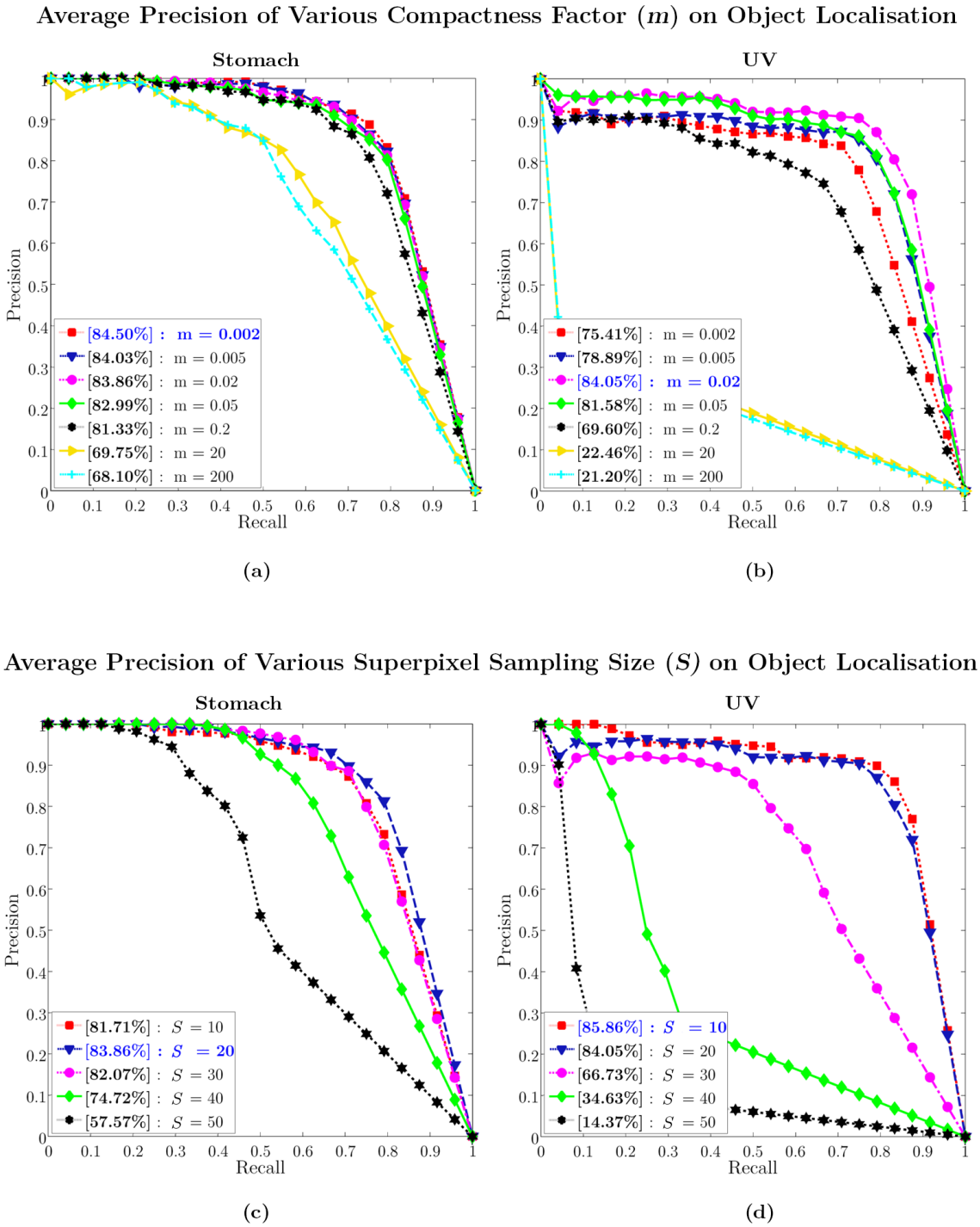


Figure 5.14: Evaluation of the effect of various sampling size *S* and compactness parameters *m*, in the key structures localisation. The average Precision-Recall curve performance in (a) and (b) demonstrates how compactness affects the stomach and the UV localisation respectively. The plots in (c) and (d) show the effect of various sampling size. Legend entries in bracket show the mean average precision (mAP).

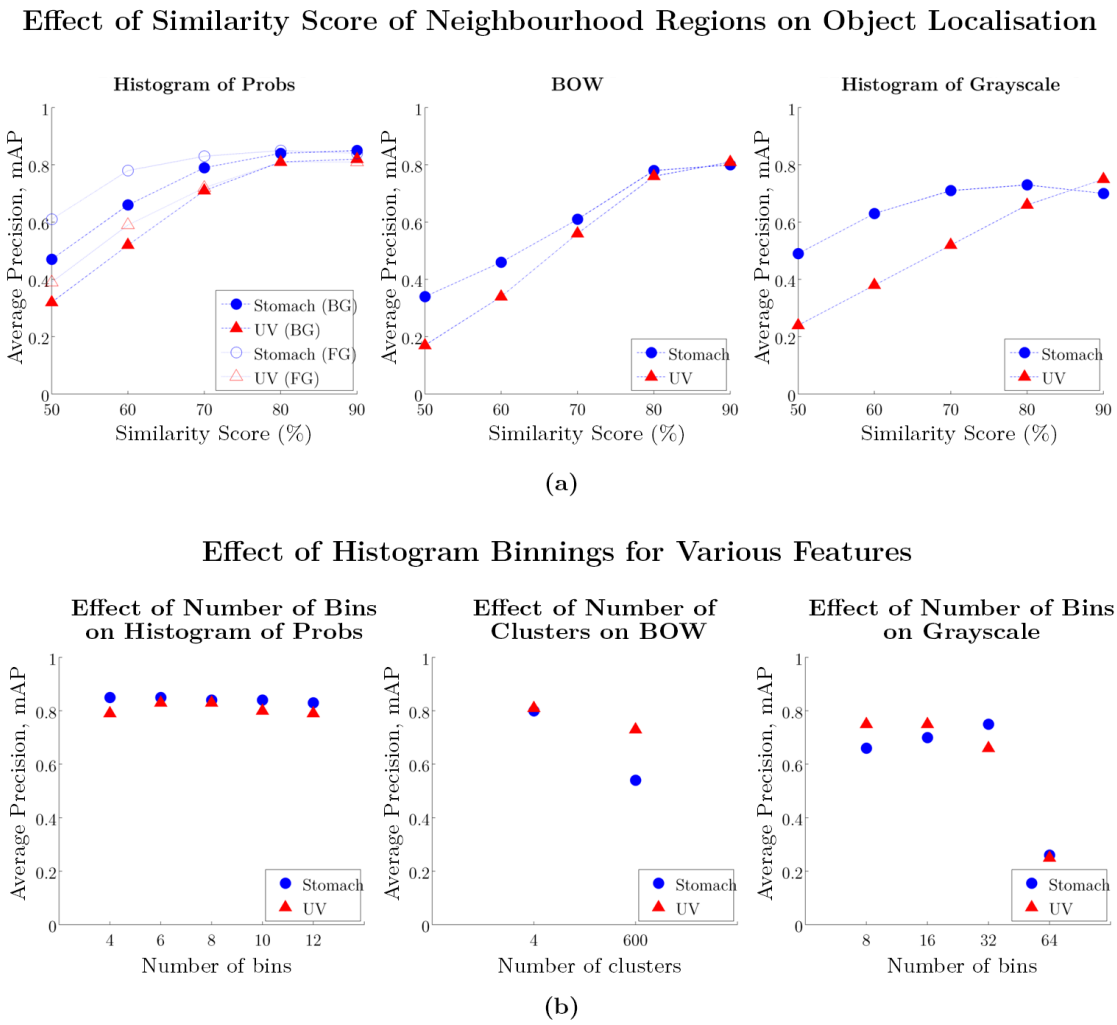


Figure 5.15: Evaluation of several parameters of various features use for generating object region. The mean average precision for three main features, histogram of mean probability h_{PROBS}, visual tokens (BoW) h_{VT}, and histogram of grayscale h_{GRAY} are compared for various criteria as shown on the x-axis.

to the low-level features such as histogram of grayscale features, h_{GRAY} as potential candidates of region grouping features to refine the superpixels. Here, we examined several factors with respect to those region grouping features using perceptual grouping cues as explained in Section 5.2.2.4. Hence, we called it region grouping cues.

Figure 5.15 shows mean average precision (*mAP*) plots for respective region grouping cues from which a similarity score threshold, T_{SIM} between two adjacent superpixels is derived. Specifically, we varied the similarity score threshold and observed the localisation performance of individual cues. We noticed that most

individual cues (except for h_{GRAY} in the stomach localisation) recorded the best score when the similarity score, $T_{SIM} \geq 90\%$. Better precision at higher visual similarity threshold indirectly indicates that cues such as h_{VT} and h_{PROBS} are highly discriminative in deciding whether the region belongs to the object or not. By contrast, the performance deteriorates when the threshold decreases due to segmentation leakage across surplus region(s).

In comparing background (BG) and foreground (FG) probabilities as the component of h_{PROBS}, both cues converge to a comparable average precision when the similarity score threshold, $T_{SIM} \geq 90\%$.

We also characterised several other options available within each cue considered for these experiments, such as number of histogram bins for h_{GRAY} and h_{PROBS}, and number of cluster centres (*e.g.* classes) for h_{VT}. As shown in Figure 5.15 (b), the best average performance for both structure localisation tasks is obtained for h_{PROBS} of $d = 6$ bins, h_{VT} of $d = 4$ classes and h_{GRAY} of $d = 16$ bins. We used these values in experiments, as reported in the next subsections.

5.4.1.3 Choice of Seeding Cues

Combination cues: Having experimented with the best choice for individual cues, we conducted experiments using those selected settings to observe whether a combination of any of them could further improve the algorithm's performance. As presented quantitatively in Figure 5.16, we showed that the h_{PROBS} cue alone scores the best with mAP = 84.93% for the stomach localisation. The next best score was recorded by a combination of h_{VT} and h_{PROBS} with 83.86% mAP. Nonetheless, a combination of h_{VT} and h_{PROBS} cues recorded the best score with 84.05% mAP in the UV localisation task, gaining almost 1.5% improvement as compared to using h_{PROBS} alone which recorded 82.58% mAP. The intensity histogram, h_{GRAY} performs worst for both tasks. Quantitatively, the combination of the histogram of probabilities h_{PROBS} and the visual token histogram h_{VT} scores the best average performance for both tasks.

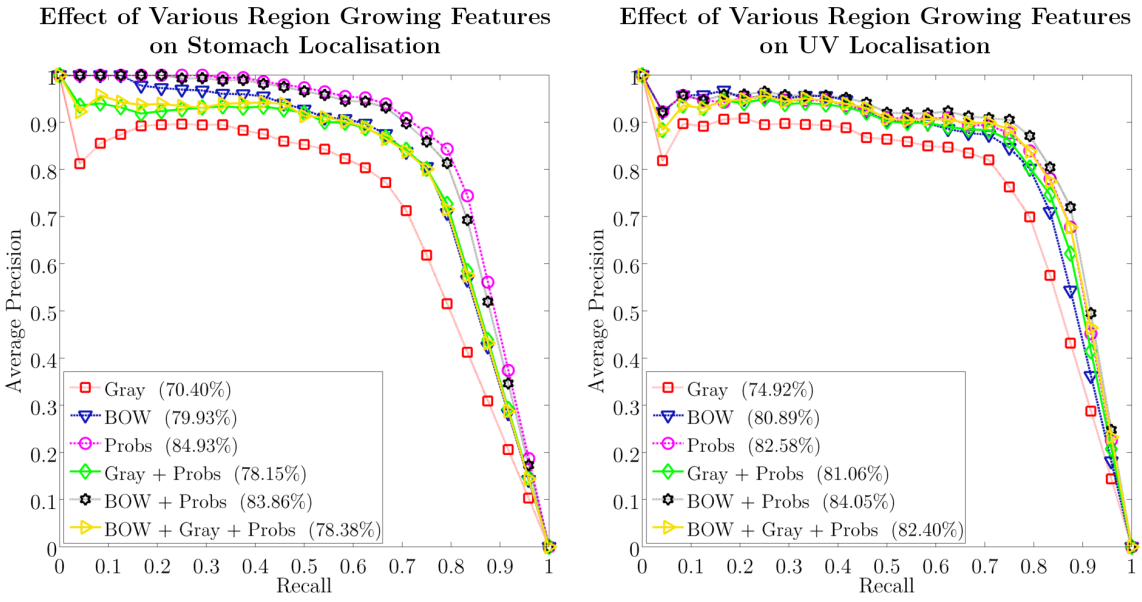


Figure 5.16: Comparison of various perceptual cues combination method. The mean average precision, mAP in bracket shows the cross-validation performance of different cues when used in isolation and when combined.

Our qualitative comparison also supports this finding. It can be clearly seen from Figure 5.17 that the combination of these two cues (*last column*) provides a higher segmentation-driven detection accuracy compared to individual implementation of either cue. The visual token histogram appears to improve the grouping decision as shown here for the stomach localisation (*row 2*) and for localising the UV (*row 3*).

It should also be noted that region grouping using intensity values records the poorest results. Adjacent regions with weaker boundaries, comparable contrast and similar textural patterns are among the factors that contribute to segmentation leakage, as shown in Figure 5.17 (*row 1, column 1*).

Background-Foreground Initialisation Seed: We experimented with the hypothesis that the probabilistic superpixel of the background class alone is powerful enough to initialise and propose the object location instead of using the foreground class probabilities (*i.e.* deriving the location using the respective stomach/UV probabilistic superpixel layers). Specifically, in this experiment, we studied the performance of the

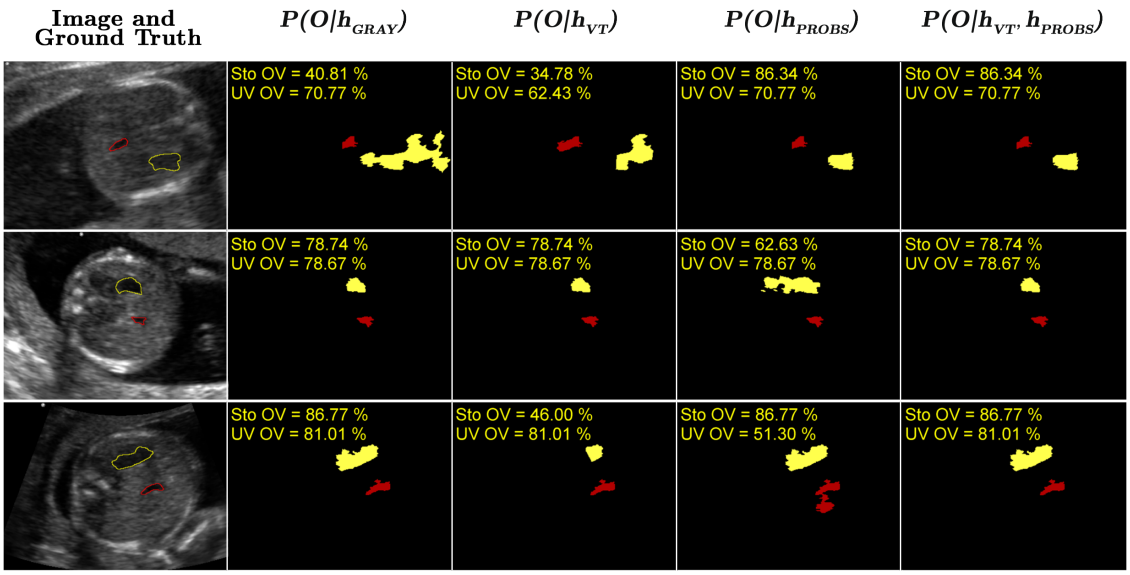


Figure 5.17: Qualitative comparison of various perceptual cues combination method. Three samples with their respective ground truth are shown in this comparison (*yellow* marker denotes stomach, *red* denotes UV). The remaining columns show localisation outputs of the structure of interests using different region grouping cues with the overlapping percentage given at top-left corner. Illustrations suggest that the combination of visual token histogram, h_{VT} into histogram of probabilities, h_{PROBS} could improve the detection accuracy. In contrast, intensity histogram, h_{GRAY} performs worst, suffering from segmentation leakage.

histogram of probability h_{PROBS} where its histogram elements are built upon either background or foreground probabilities.

Here are the differences between these two approaches:

1) **Background Probabilities Seeding:** The superpixel regions are ranked according to the mean background probabilistic score. The region grouping process starts from an initial seed with the lowest-score superpixel (*i.e.* referring to the region with highest probability of belonging to either structure). The histogram of probabilities consists of background probabilities as its histogram elements. Adjacent superpixels are merged according to their similarity in appearance, as explained in Section 5.2.2.4. The histogram of the merged superpixels is updated, summed and normalised after every successful merge. Upon completion of the merging process, each candidate region is assigned a probability score of belonging to either class (*i.e.* stomach or UV), whichever is maximum.

2) **Foreground Probabilities Seeding:** The process is similar to the concept explained above, except that the superpixel regions are ranked according to the respective mean foreground probabilistic score for the region (*i.e.* each structure maintains separate list). Again, for each structure, region merging starts from superpixel of the highest probability. A candidate takes the maximum probabilistic score of the newly formed superpixels as its confidence score.

We re-ranked all the candidates according to their confidence score and took the region with the maximum score of the respective foreground class as the final object detection. Figure 5.18 compares the performance of region grouping using different initialisation seeds. Using *DSC* threshold $= 50\%$ as an evaluation criterion, seeding using background probabilities provides had better performance, with both localisation tasks achieving approximately 1% improvement compared to when using foreground probabilities.

Also, we observed how accurate the object localisation is if *DSC* threshold $= 70\%$ (*i.e.* the detection should overlap more to the ground truth to be considered a true positive detection). The result shows that the background probabilities demonstrate approximately $3 - 5\%$ better localisation as compared to when using foreground probabilities as a seeding criterion.

A qualitative comparison is provided in Figure 5.19. As demonstrated, a seeding region grouping with a background probability shows superior performance and better object detection quality as compared to seeding with a foreground probability. The first two samples demonstrate a better segmentation quality of the stomach when using background probability as seeding criteria. The last two samples, as presented in column (b), show that region grouping using background probability seed (fourth row) contributes to better segmentation outputs.

The illustration also displays top potential region proposals which are presented in shaded grey (the lightest symbolises higher probability region). It should be noted

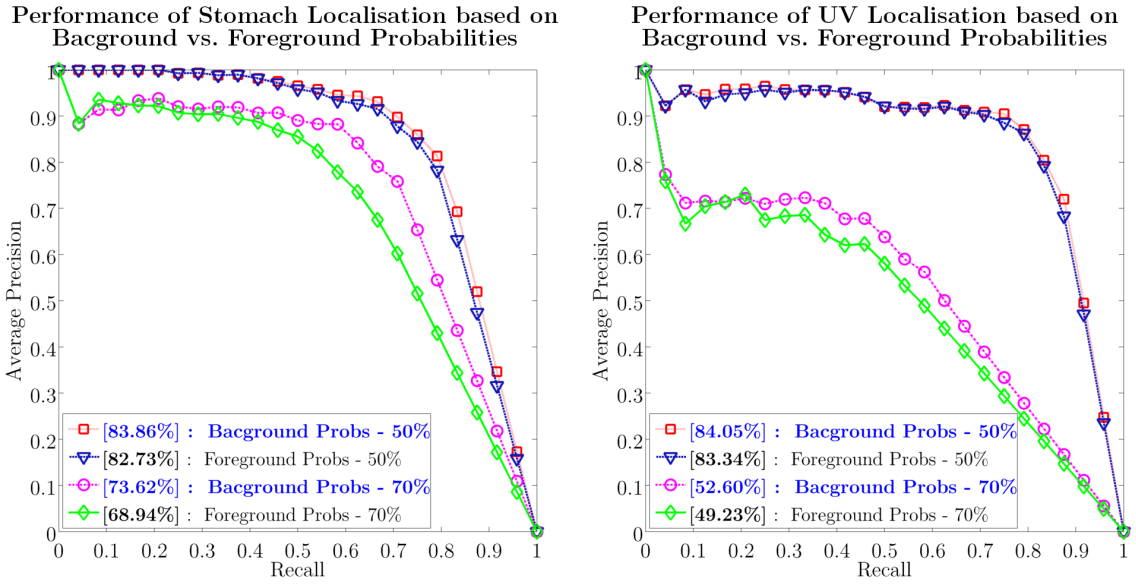


Figure 5.18: Comparison of object localisation performance given background vs foreground probabilities as seed grouping criteria. Given h_{PROBS} as one of the region grouping features, we tested the choice of background vs. foreground probabilities as the histogram element. Localisation is evaluated over two different dice coefficient values which are 50% and 70% overlap rules to show that object localisation is more precise given a background probability as seeding cues.

that the sample in the third column shows a falsely identified stomach. The second highest-ranked region (pointed out by yellow bounding-box) is correctly highlighted for the case of seeding with a background probability approach but is totally missed when using a foreground probability seeding (third row).

These results suggest several conclusions. *Firstly*, the localisation based on a procedure similar to the background rejection process works effectively to highlight the potential regions. Subsequent recognition on the highlighted regions, such as exploring the foreground probability score, will provide further assistance to identify the existence of the respective structures. *Secondly*, the technique also performs well in detecting and ranking the probable locations of the respective structures.

Given this list of candidates to a system which takes advantage of having an abdomen detector, the initially false negative region could be promoted as the right candidate by dismissing the detected regions that fall outside the detected abdomen area. As experimented later, the inclusion of an abdomen detector (*Module B*) on top of this detection (*Module A*) restricts the searching space within the abdominal region,

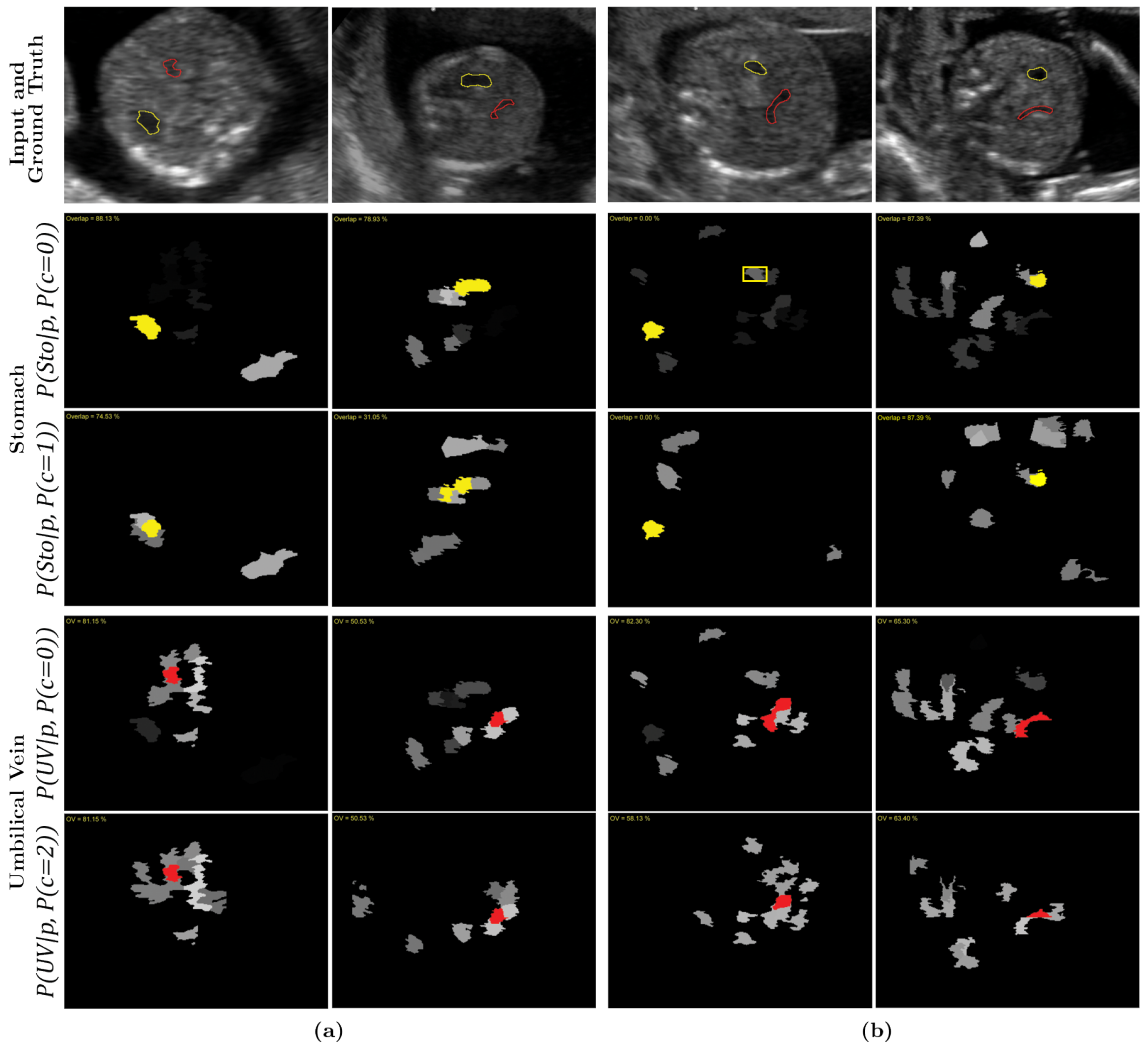


Figure 5.19: Qualitative analysis on background vs. foreground probabilities as seed grouping criteria. Given 4 samples with the respective ground truth (*first row*), we showed how the proposed algorithms combines different regions, given either background or foreground probability as its seed grouping criteria. Given background probability, [*row 2* and *row 4*] and foreground probability, $P(c = 0)$[*row 3* and *row 5*], listed regions belonging to the stomach (in *yellow*) and the UV (in *red*) are localised from a list of potential candidates (highlighted in grey). The first two samples in column (a) compare the quality of object localisation for the stomach, while the remaining samples in column (b) compare the quality of object localisation for the UV. These illustrations demonstrate that better object detection quality is obtained via background probability approach.

hence allowing the model to select a region highlighted by the bounding box instead, as demonstrated in Figure 5.19.

5.4.2 Quantitative Analysis

We evaluated the performance of the proposed segmentation-driven detection model on an unseen test set consisting of 360 abdominal images. In particular, we compared four different approaches, described as follows:

1) **Module A (only):** This is the original segmentation-driven model which optimises the output of probabilistic superpixel based on perceptual grouping algorithm.

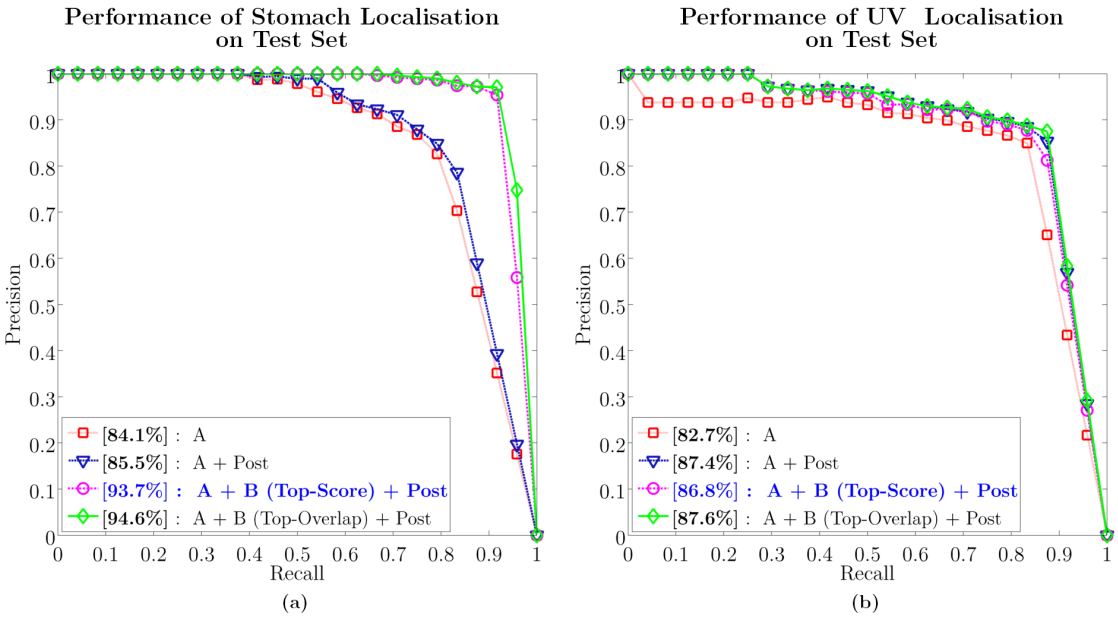
2) **Module A with Post-Processing:** This method presents the output after a post-processing steps which include the Random Walk algorithm, as explained in Section 5.2.3.

3) **Module A + Module B (with Top-Scoring Detection) + Post-Processing:** This approach refers to the implementation of *Module B* on top of the *Module A* (inclusive of post-processing). In this approach, the *Module B* will simply restrict the searching space to be within the potential abdomen, hence dismissing any regions outside the area marked as abdomen.

4) **Module A + Module B (with Top-Overlap Detection) + Post-Processing:** As summarised in Section 4.4.4 and illustrated in Figure 4.17, the *top-overlap* abdominal detection is evaluated instead of taking the *top-scoring* abdomen to foresee its impact in finding these key structures.

Figure 5.20(a) presents the performance of the proposed method on the stomach localisation while Figure 5.20 (b) compares the performance of the UV localisation task using four different approaches as described earlier.

Without post-processing, the localisation recorded a satisfactory performance by scoring 84.1% and 82.7% for the stomach and the UV respectively. After the implementation of post-processing technique, the performance of the stomach localisation recorded over 1% improvement to 85.5% while the UV localisation gained



Summary

Method	Description	Mean AP	
		Stomach	UV
A	Module A only (without post-processing)	84.1 %	82.7 %
A + Post	Module A with post-processing algorithm.	85.5 %	87.4 %
A + B (top- score) + Post	Module A + Top-scoring abdominal detection (Module B) + post-processing algorithm.	93.7 %	86.8 %
A + B (top-overlap) + Post	Module A + top-overlap abdominal detection (Module B) + post-processing algorithm.	94.6 %	87.6 %

(c)

Figure 5.20: Quantitative analysis of multiple key structure localisation on a test set. We evaluate the performance of the proposed method on an unseen test set. Four different approaches were tested and compared to measure the performance of the stomach localisation (a) and the UV localisation (b). Descriptions of the approaches are summarised in (c). Average precision of each method is given in the bracket.

massive 5% improvement, with $mAP = 87.4\%$. This reflects the quality improvement made on substantial failure cases of the UV localisation output where mostly suffer from disjointed parts of the structure.

Next, given the knowledge of where the abdomen is (Proposal *Module B*), we recorded and studied the performance of the algorithm. We provided side-by-side comparison for two different scenarios; 1) by taking the top-score abdominal detection, and 2) by using the top-overlap detection (*i.e.* to investigate the improvement achieved, given the best candidate). As summarised in Figure 5.20, stomach localisation recorded a significant improvement of 8.2% given the location

of the abdomen and gained another 0.9% improvement if the best candidate was selected. In contrast, the UV localisation performance dropped by 0.6% upon the implementation of the abdomen detector (*Module B*), but recorded 87.6%, a gain of 0.2%, given a much precise abdomen location.

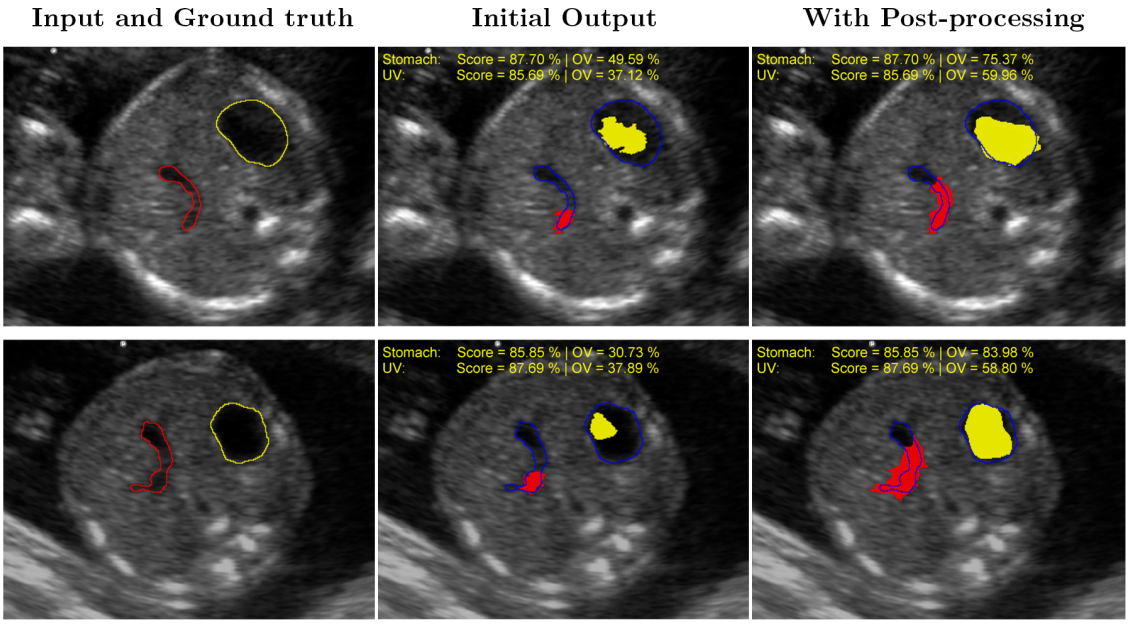
Two important conclusions can be drawn from this; 1) Recognition of precise abdomen location aids the process of bottom-up re-organisation process by restricting the searching space for the key landmarks. It also efficiently helps to reject any similar-looking structures outside the region, and 2) Any inaccurate abdominal detection will potentially cause the failure in detecting the intended structure(s).

Nevertheless, simple selection of the abdominal region based on top-score detection is insufficient. Results demonstrate that an accurate abdomen detection (using top-overlap abdominal candidates) could further improve the performance of UV localisation from 87.4% to 87.6%. Likewise, the stomach localisation could achieve another 1% improvement if given an accurate abdomen detection.

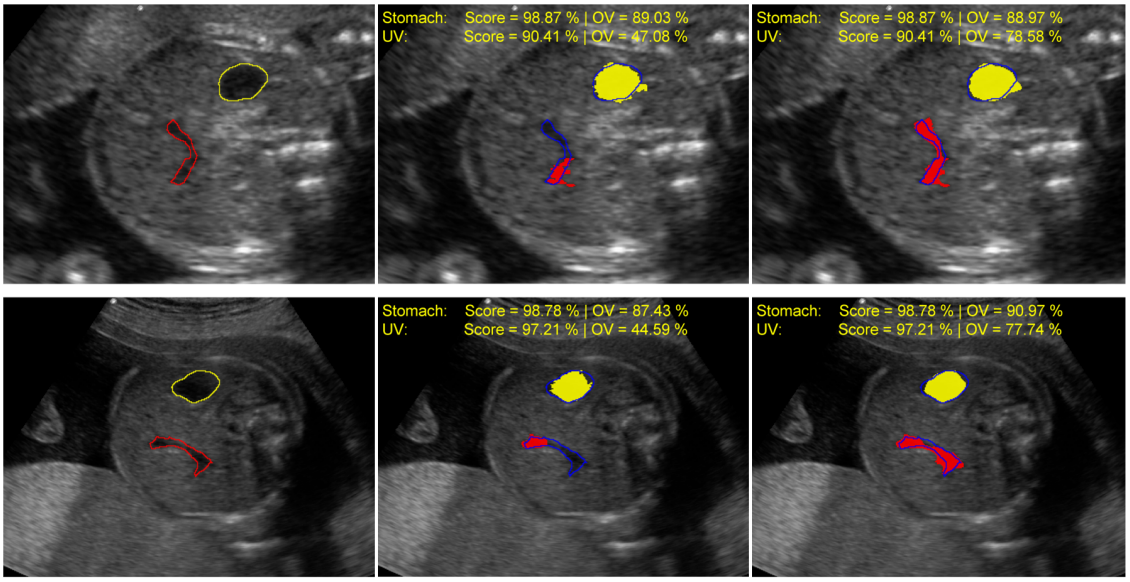
Overall, it was found that the post-processing technique improves the segmentation accuracy, especially for localisation of the UV. In general, it could be seen that the reorganisation process of grouping and segmentation of key landmarks fruitfully exploits the output of the recognition algorithm of an abdominal location. The precise location of the abdomen increases the probability that a particular landmark is inside the region, which then enables the algorithm to re-score the correct detections.

5.4.3 Qualitative Analysis

Next, the qualitative results in Figure 5.21 - Figure 5.24 are presented in order to provide a comparative analysis on the improvement attained due to the post-processing stage as well as to study the importance of attaining a precise abdomen detection prior to key anatomical structure localisation.



(a)



(b)

Figure 5.21: Qualitative Analysis I - Illustration of rectification of false positive cases into true positive detections using the proposed post-processing approach Given input images as in the *left* column, the proposed post-processing approach improves the localisation outputs from false positive (*FP*) detections (*middle* column) to true positive (*TP*) detections (*right* column). Samples in (a) show improvements for both structures which recombine missed neighbourhood regions due to reasons like differences in textural patterns and unoptimised superpixelisation grouping. Nonetheless, in most cases, the UV benefits much from this approach as illustrated by samples in (b). The UV structure of a region with weaker intensity in between typically breaks the anticipated structure into two separate regions. The algorithm combines the nearby region (separated by a superpixels) upon implementing the post-processing algorithm. Note that the input images (*first column*) display their respective ground truth for the stomach (*yellow*) and the UV (*red*) while the resulted outputs have their detections: the stomach in yellow and the UV in red with an overlaid ground truth in *blue*.

5.4.3.1 Post-processing vs None

Figure 5.21 shows a side-by-side comparison between the initial output without post-processing algorithm (*middle* column) against the output of the post-processing stage (*right* column). Other than under segmentation issues, as seen in Figure 5.21 (b), most misdetections of the UV occur where the structure is fragmented into two disjointed regions with the existence of a weaker region in between.

In the two cases involving the stomach (Figure 5.21 (a)), an initial grouping, based on a combination of visual tokens, h_{VT} and background probability, h_{PROBS} failed to group the regions due to a slight difference in the textural pattern histogram, which recorded a relatively lower score than the threshold provided. Consecutive refinement discovered that both proposed regions have a similar probabilistic score of belonging to the stomach, thus should be grouped together.

Figure 5.22 (a) and (b) respectively present examples where a better representation of the stomach and the UV can be obtained by incorporating a post-processing step. The selected highest-scoring region, as shown in the *middle* column, is already deemed as a true positive (*TP*) detection with an overlap ratio of over 50% with the ground truth. However, the detection can be clearly seen to not conform to the true shape of the actual structure. Post-processing has been found to improve the segmentation-detection quality by $10-40\%$, as shown by the examples (*last* column).

5.4.3.2 Abdomen Detection (Module B) Employment

In this experiment, we investigated and demonstrated the importance of having an accurate abdomen detection prior to multiple key structure localisation. We combined proposal *Module A* with the *Module B* learned in Chapter 4.

Figure 5.23 illustrates cases where abdomen detection assists in the key structure localisation. As shown, the stomach benefited most from the implementation of an abdomen detector. Frequently, the stomach is incorrectly identified when the fetal abdomen is attached to the homogeneous tissue of the uterus. Accurate abdomen

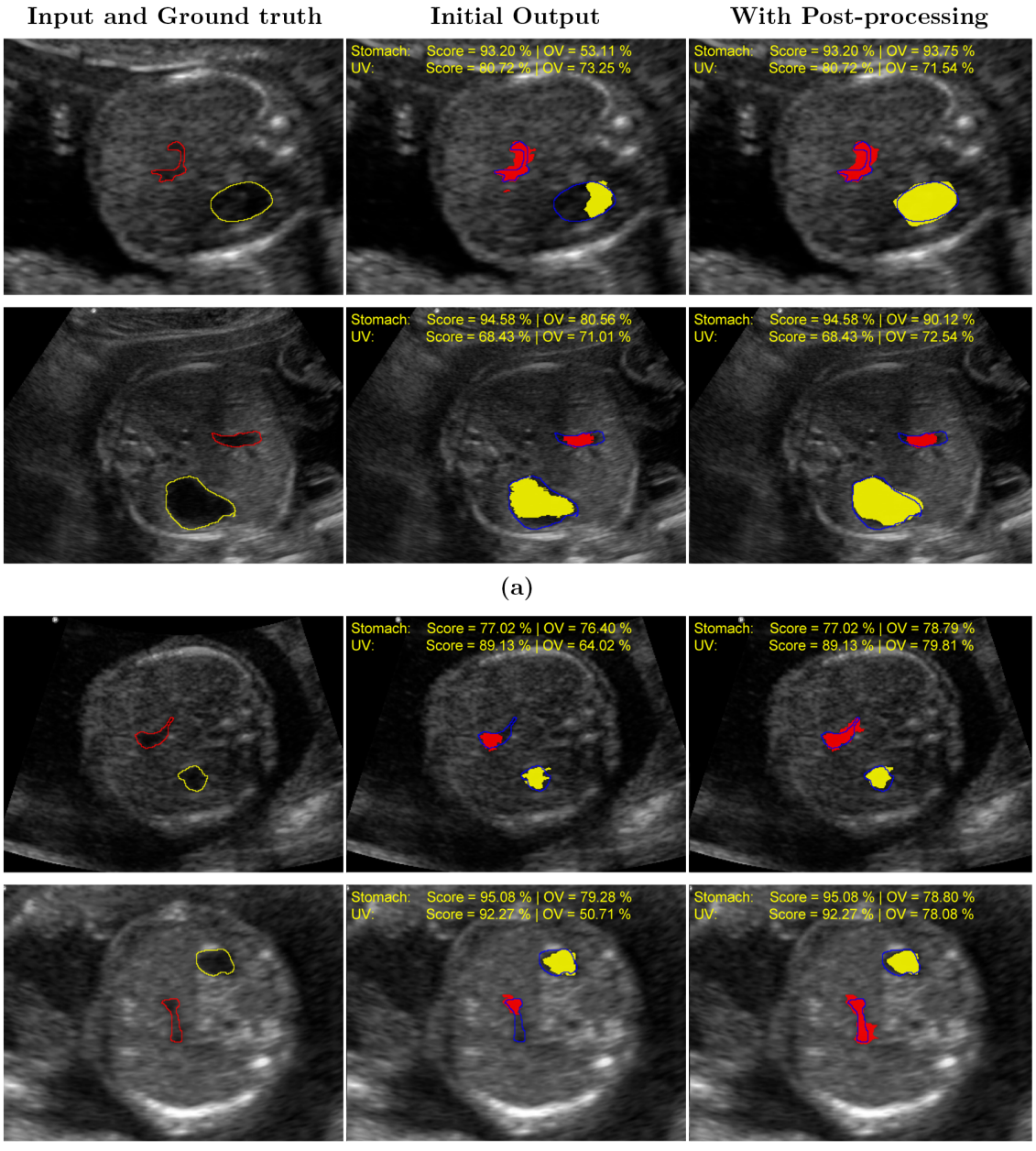


Figure 5.22: Qualitative Analysis II - Illustration of improvements on true positive detections using the proposed post-processing approach. In addition to the previous cases given in Figure 5.21, the post-processing algorithm does improve the segmentation quality of the already correctly identified structures. Samples in (a) show a better representation of the stomach (in *yellow*) with an improvement of 40% and 10% respectively. Samples in (b) show improvement of the UV outputs (in red) with a rise of 15% and 40% each in the overlap computation to the ground truth.

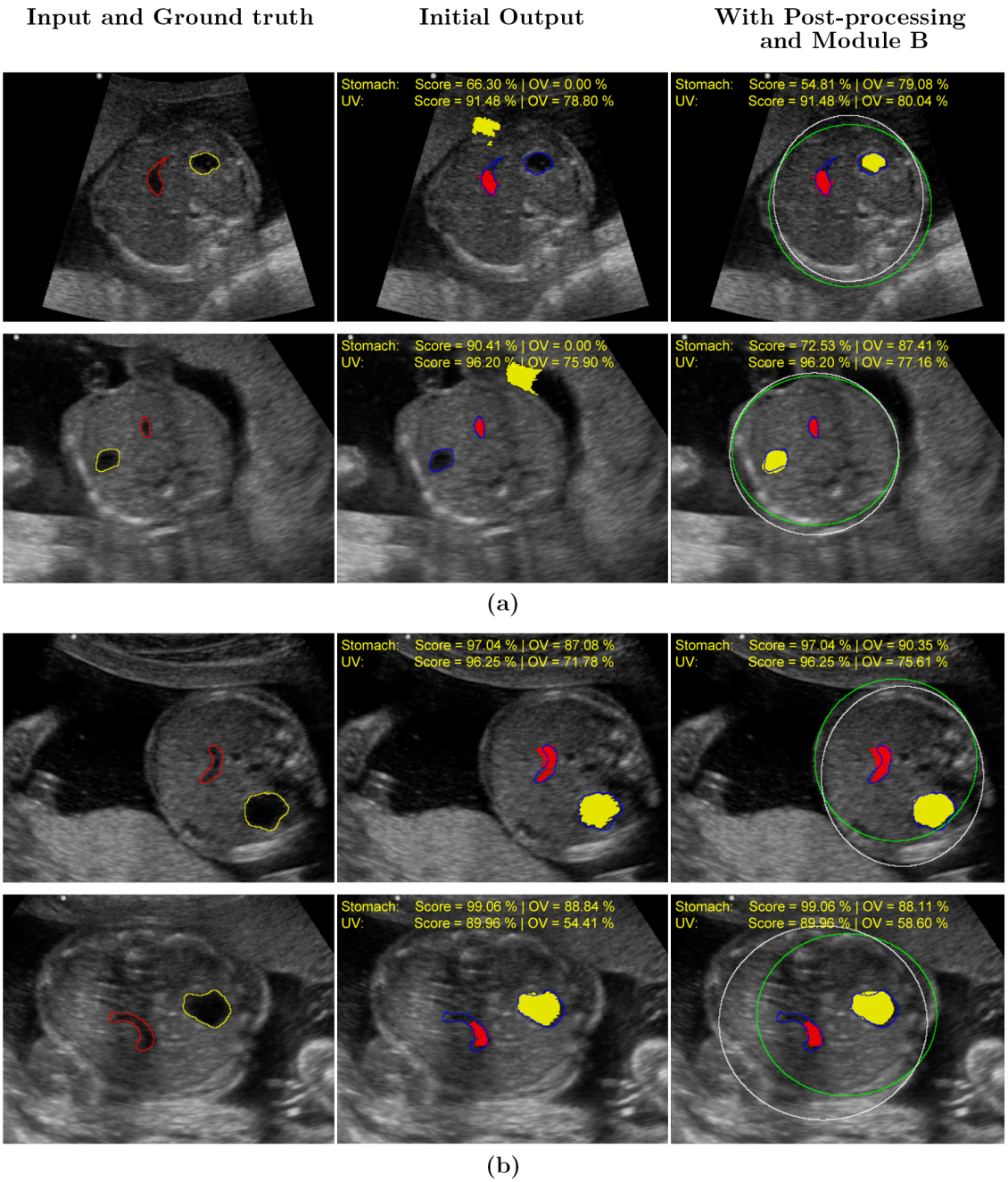


Figure 5.23: Qualitative Analysis III - Illustration of true positive detections by combining abdomen detection (Module B). (a) Major improvements are shown in the detection of the stomach when abdomen detection module is implemented beforehand as illustrated in these examples. Most incorrectly identified stomach localisations (annotated in *yellow* in the second column) refer to a small dark area just outside the stomach when the fetal abdominal wall is attached to the homogeneous tissue of the mother. This proposed candidate is then rejected, given an abdomen detection. Instead, the second-highest ranked proposal, which is located inside the abdomen, is selected. (b) We overlaid the highest-ranked abdomen ellipsoidal marker in yellow and the highest-overlap abdomen in *white*. These two samples show cases where the key structures (stomach in *yellow* and the UV in *red*) were still selected, although given a not so-accurate abdominal detection (yellow ellipsoidal marker).

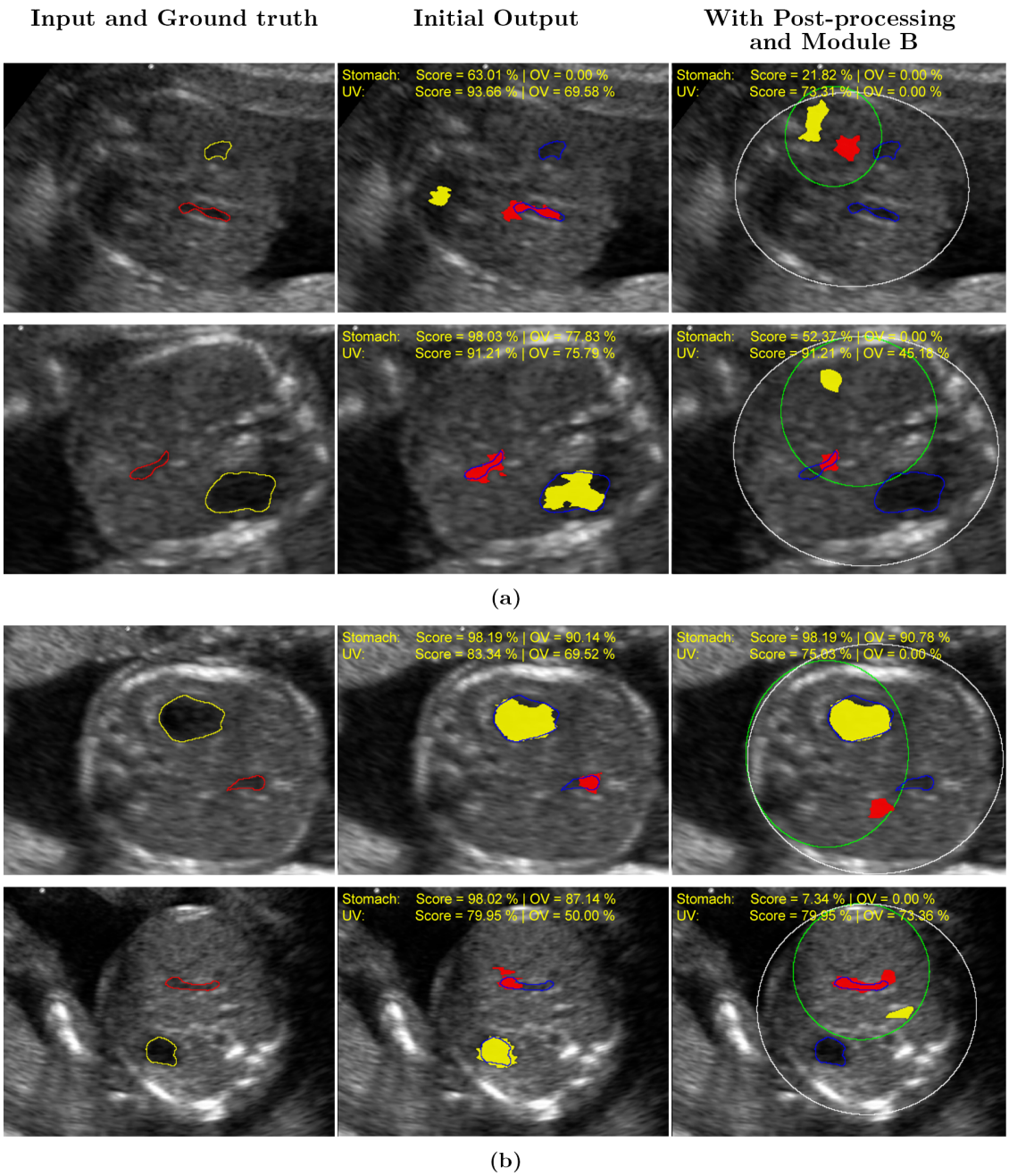


Figure 5.24: Qualitative Analysis IV - Illustration of false detections upon the implementation of the abdominal detection (Module B). These samples illustrate the effect of falsely localised abdomen. Cases in (a) shows how a falsely identified area of the abdomen region is missing both key structures. In the second sample (*second row*), it even cancels out the highest-ranked proposal of the key structures, which have been truly localised. Cases in (b) shows that either key structures is missed as a result of falsely identified abdomen. Note that in all samples provided here, the top-overlapped abdomen detection (*white* ellipsoidal marker) which is the true manifestation of the abdomen, still contains those structures. The importance of spatial relativity and interconnection between the locations of these key structures to the orientation of the abdomen in providing re-correction step will be further explored as part of future extension work.

detection allow the system to correctively locate the stomach as shown in Figure 5.23 (a). In contrast, samples provided in Figure 5.23 (b) illustrate cases where the key structures are correctly localised, although the abdomen detection (shown as yellow ellipsoidal marker) are marginally off from the ground truth provided.

Finally, a qualitative report presented in Figure 5.24 highlights some limitations of the technique where a falsely identified abdomen region leads to a wrong structure detection. In particular, a top-ranked structure can be cancelled out as a result of having a misclassified abdomen detection. It should be noted that in all samples provided here, the top-overlapped abdomen detection (in white ellipsoidal marker) which illustrates the true manifestation of the abdomen, still contains those key structures.

5.5 Summary

In this chapter, we have demonstrated that casting multiple object localisations via bottom-up grouping provides a good alternative to a conventional sliding window approach. The segmentation-driven detection model as proposed here, has a richer structure representation and achieves a remarkable performance with less candidate proposals compared to multi-scale sliding windows.

The ability to quickly detect and separate the object from the background given a real-time imaging modality such as US, is crucial. This work shows that figure-ground separation, largely inspired by work in perceptual grouping, is key in reducing search complexity and provides a solution to the problem of localising multiple key structures. Utilising a combination of low-level and mid-level information cues within a superpixel region framework, this approach resolves bottom-up ambiguity by transforming a pixel-wise probabilistic output into meaningful regions which makes up the real structure.

We then demonstrated the experimental results of the impact of the abdomen detector proposed in *Module B* as presented earlier in Chapter 4 to the detection framework. Generally, the introduction of an accurate abdomen detection is massive in improving the stomach localisation in particular, but not in the case of the UV due to the existence of multiple similar-looking UV regions within the abdomen. At this stage, we could conclude that finding the correct abdominal region is the first step towards recognising other key structures. Extension works to improve the detection accuracy, such as exploring the pose of the abdomen and examining the spatial relationship of key structures are proposed in Chapter 6.

6

Conclusions and Future Work

This thesis has investigated the roles of objective scoring criteria in characterising the standard diagnostic plane identification and verification for abdominal biometric measurement. We have presented a multiple proposals framework which incorporates two significant areas of computer vision for medical image analysis: 1) ***recognition*:** the development of coarse module for abdominal localisation, and 2) ***re-organisation*:** the localisation of multiple key structures via bottom-up grouping process of superpixels.

To conclude, we review the key contributions of this thesis and comment on the limitation of the proposed method. This will be followed by a discussion of a number of possible areas for future work

6.1 Summary of Contributions

We begin by summarising our work and contributions reported in this thesis, which spread out over Chapter 3 - Chapter 5. This include brief comments on what we have learned from our experimental works.

6.1.1 Clinical Dataset

We presented our collection of annotated clinical dataset of fetal ultrasound abdominal images in Chapter 3. The dataset contains a collection of cropped and full-size images for localisation and segmentation task. In addition, we also annotate the viewpoint and pose of the abdomen for future pose estimation task.

During data collection, the sonographers were free to perform their best measurement in order to avoid any possibility of biases. As a result of that, we observe that the images produced were mostly taken accustomed to personal preferences based on regular practice. Most of the data collection ending up having an image facing a certain direction as traditionally preferred by the sonographer. Although the images could later be manually transformed to a different view, scale, position *etc.*, it is interesting to study whether the proposed approach could learn a new pattern given the data collection made from an arbitrary position of the structures.

6.1.2 Coarse Recognition Module of Fetal Ultrasound Structures

In Chapter 4, we proposed a coarse localisation framework for abdominal images using a boosting tree and channel features. In designing *Module B*, we performed extensive experimental assessment of model parameters. These included experiments on window size, padding size, feature types, number of weak classifiers and several others as summarised in Table 4.4. We chose to seek to understand how to design a good feature encoding scheme and classifier settings for ultrasound image recognition.

Boosting Decision Tree Classifier: In order to give us a quick decision-making framework for coarse abdominal localisation which ensuring balance performance between speed and accuracy, we train a shallow decision tree with boosting mechanism. Given a specific target on permissible miss-rate performance within a real-time performance, the designed classification architecture cleverly demonstrated that several weak classifiers are sufficient with a proper combination of calibration parameter and soft-rejection threshold settings.

Channel Features: We found out that simple channel features based on low-level pixel values gave as meaningful results as richer one such as Haar features given a challenging US image modality. While a Haar feature, for example, involve a computation of multiple regions comparison, channel features are designed directly on single pixel values and its approximation has been demonstrated to work across scale-space without the need for feature re-computation. Hence, experimental works carried out in this thesis had shown that the channel features work well in US images and do offer a quick solution especially for a real-time solution.

Soft-Cascade Threshold: On top of that, we found that a soft-cascade rejection threshold and recalibration on the classifier settings provided a good combination to tolerate hard positive images. With these features, the proposed coarse recognition module is not only offering a fast technique to abdominal image localisation but also produces a high recall-rate.

6.1.3 A Fine Bottom-Up Re-Organisation Framework of Multiple Key Anatomical Structures

We presented in Chapter 5, a bottom-up segmentation-driven detection framework which takes the input images, generates a probabilistic superpixel and incorporates low-level and mid-level cues such as grayscale pixel values and visual tokens to transform and combine neighbourhood regions into meaningful regions which indirectly translated to the real structure of interests.

Probabilistic Superpixel: We demonstrated that transforming a pixel-wise probabilistic output into meaningful regions in a bottom-up fashion using a superpixel region framework offers a good alternative to a conventional sliding window approach in reducing the searching complexity and in providing a solution to the problem of localising multiple key structures. Also, working on a region level (*e.g.* superpixel) compared to individual pixel could help to improve object localisation accuracy especially in diminishing the speckle effect of US signal.

Background Rejection Concept: Our experimental study shows that the background rejection theory is implicitly important in revealing the probable location of foreground objects. Using background probability as seeding criterion for region merging demonstrates superior performance compared to when using foreground probability. The results suggest that the localisation works effectively by first highlighting the potential region via procedure similar to the background rejection process. Further recognition process on the highlighted regions such as exploring the foreground probability score will then help to identify the existence of the respective structures. Not only that, it also provides a better segmentation quality of the structure of interests.

Fine Segmentation via Random Walker algorithm: We also introduced an automatic post-processing algorithm via Random Walker technique. This improved the segmentation accuracy especially in the case of localisation of an articulated object such as the UV. Nevertheless, the qualitative comparison of the results suggests that UV localisation may suffer from an under segmentation problem. This scenario happens when the UV is fragmented into two disjointed regions by the existence of a weaker region in between.

6.1.4 Multiple Proposal Framework

Finally, we report the result of our 2-stage multiple proposal framework in Chapter 5 by presenting the experimental results of the impact of the abdomen detector (*object*

recognition) as proposed in *Module B*, when combining with the detection of multiple-key structures (*object re-organisation*) as designed in *Module A*. In general, the introduction of an accurate abdomen detection was found to improve the stomach localisation, but not in the case of the UV due to the existence of multiple similar-looking UV regions within the abdomen.

Two important conclusions can be drawn from this; 1) Recognition process of precise abdominal location aids the process of bottom-up re-organisation by restricting the searching space for the key landmarks. It also efficiently helps to reject any similar-looking structures outside the region, and 2) Any inaccurate abdominal detection will potentially cause failure in detecting the intended structure(s). This finding indirectly signifies the importance of finding the accurate abdominal location prior to localising the key structures. Exploring the pose of the abdomen and examining the spatial relationship of key structures are anticipated as part of extension works to improve the detection accuracy. Such extensions of the method are addressed in the following section.

6.2 Supplementary Future Work

Based on the experimental work presented here in this thesis, there are few potential areas which could deliver further improvement to the area of US medical image analysis using machine learning framework. This section provides and discusses some pointers to possible future research.

6.2.1 Joint Detection and Pose Estimation Framework

The main motivation of this thesis is based on the conceptual framework of 3R: *recognition*, *re-organisation* and *reconstruction* in the computer vision work. However, as considered here, we have only presented work on *recognition* and *re-organisation* areas, leaving the last part to be explored as future work. Our first recognition module (*Module B*) as proposed in Chapter 4 was targeted to perform a quick and accurate

detection of an abdomen. We then presented the second *re-organisation* module (*Module A*) as discussed in Chapter 5 to demonstrate the importance of abdomen detection as a pre-processing stage to the localisation of the key anatomical structures. As observed from the experiments, the introduction of an accurate abdomen detection is a massive step in improving the stomach localisation in particular, but not in the case of the UV due to the existence of multiple similar-looking UV regions within the abdomen.

In real-time 2D sonography, the operator is required to actively and continuously engages by performing several tasks simultaneously; both hands manipulating the transducer and the machine consoles while maintaining hand-eye coordination in searching for the best diagnostic plane. Despite the dynamic changes tha results from the manoeuvring process in addition to the fetal movement, the instantaneous response of a sonographer to combine their knowledge about fetal anatomy, reconstructing a 3D mental visualization from 2D images to guide the subsequent movement is quite amazing. Ahmed & Noble (2015) have conducted eye-tracking experiments to discover the searching strategy of a sonographer in defining the best standardised plane. By recording the eye-gazing pattern of the participants, their eye-tracking experiments demonstrated a few important findings; 1) confirming the importance of the stomach and UV landmarks in defining the plane, and; 2) repeated alternating eye movement back and forth on several landmarks suggesting that the spine in particular, was used as a cross-referencing landmark to verify the position of other anatomical landmarks.

We revisit the possible scenario in a multiple structural localization problem as shown in Figure 6.1. Looking from another perspective, the more permutations permissible of the diagnostic plane, the more challenging the interpretation would be, as the pattern and geometrical context of the structures will be hugely varied. In other words, the viewing angles (known as 'poses' or 'viewpoints') from where the diagnostic plane is taken from will further determine the location of specific structures. Figure 6.1 (*last row*) illustrates the important role of context (*e.g.* pose viewpoint) in

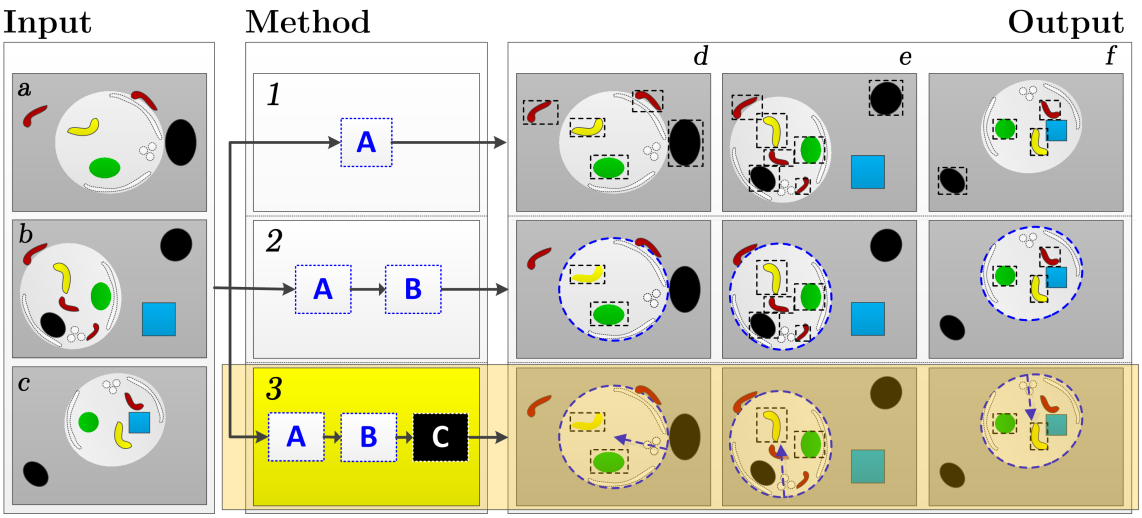


Figure 6.1: Coarse-to-fine multiple proposal framework. A precise location of the abdomen as illustrated by Method 2 (*middle* row) enable the elimination process of probable key anatomical structures outside an abdomen. However, there are still many cases in which similar-looking structures appear at multiple locations within the abdomen as illustrated by the last two columns (column *e* and *f*). A pose estimation framework (*Module C*) is proposed herein to solve this problem by considering the implicit role of the abdomen's pose in aiding the decision-making of selecting the best standard plane view.

understanding and interpreting the real scene. The last two columns pointed out that given an accurate viewpoint direction, the probability of finding another structure at certain relative positions is higher. In column *e*, having a knowledge that the spine pointing inwards from the 6 o'clock position will limit the probable locations of the stomach at either 3 o'clock or 9 o'clock position. Similar to the case highlighted in column *f*, the location of the UV could be easily predicted given the location of the stomach with respect to the viewpoint of the spine.

Hence, it is observed that the systematic manual navigation which produces the optimal imaging diagnostic plane for biometry measurement is not only limited to identifying the presence of the key anatomical landmarks within the field of view, but hugely subject to the ability to generate a geometrical reconstruction of the organ of interest. Indirectly, knowing the pose relative to the key landmarks allows their relationships to be easily recognised, thus reducing the chance of mis-detections or better, the possibility to select the exact location of the respective structures in the cases where there are multiple probable candidates. Rather than looking for the spine, the

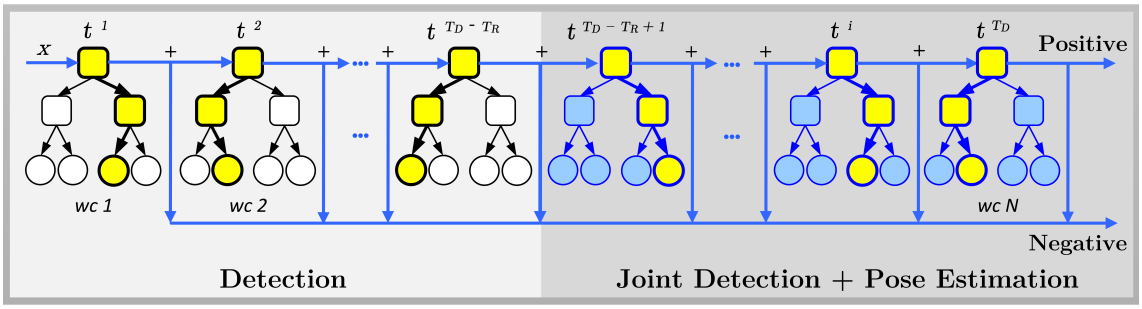


Figure 6.2: Joint detection and pose estimation module. We could implement a similar classifier and feature vector as described in Section 4.2. The pose estimation module can be introduced by stacking up a regression process in the last T_R trees, marked by blue colour. The viewpoint for a sample which successfully classified as positive detection can be retrieved by examining the accumulated partial pose response from the T_R regression maps.

introduction of an implicit viewpoint or pose, such as the one proposed herein, may simplify the examination of the standard abdominal imaging plane.

Based on this insightful observation, the ideas developed in this thesis, in particular the experimental works done earlier in Chapter 4 could be extended by proposing a joint detection and pose estimation framework. Juránek et al. (2015) had employed a single compact model based on similar boosting tree architecture to perform an object detection and pose estimation jointly in a cascade style for a car detection problem. The approach provides promising evidence that both tasks can be executed concurrently at the speed of real-time operation.

Following this method, an interesting extension to the research described in Chapter 4 and Chapter 5 would be to add a pose estimation stage, which would have only a small additional computational cost at test time as shown in Figure 6.2. By adding that, a comprehensive unifying multiple proposal framework, which jointly performed recognition, re-organisation and reconstruction tasks to classify the best diagnostic plane for biometric measurement can be accomplished as illustrated in Figure 6.3 .

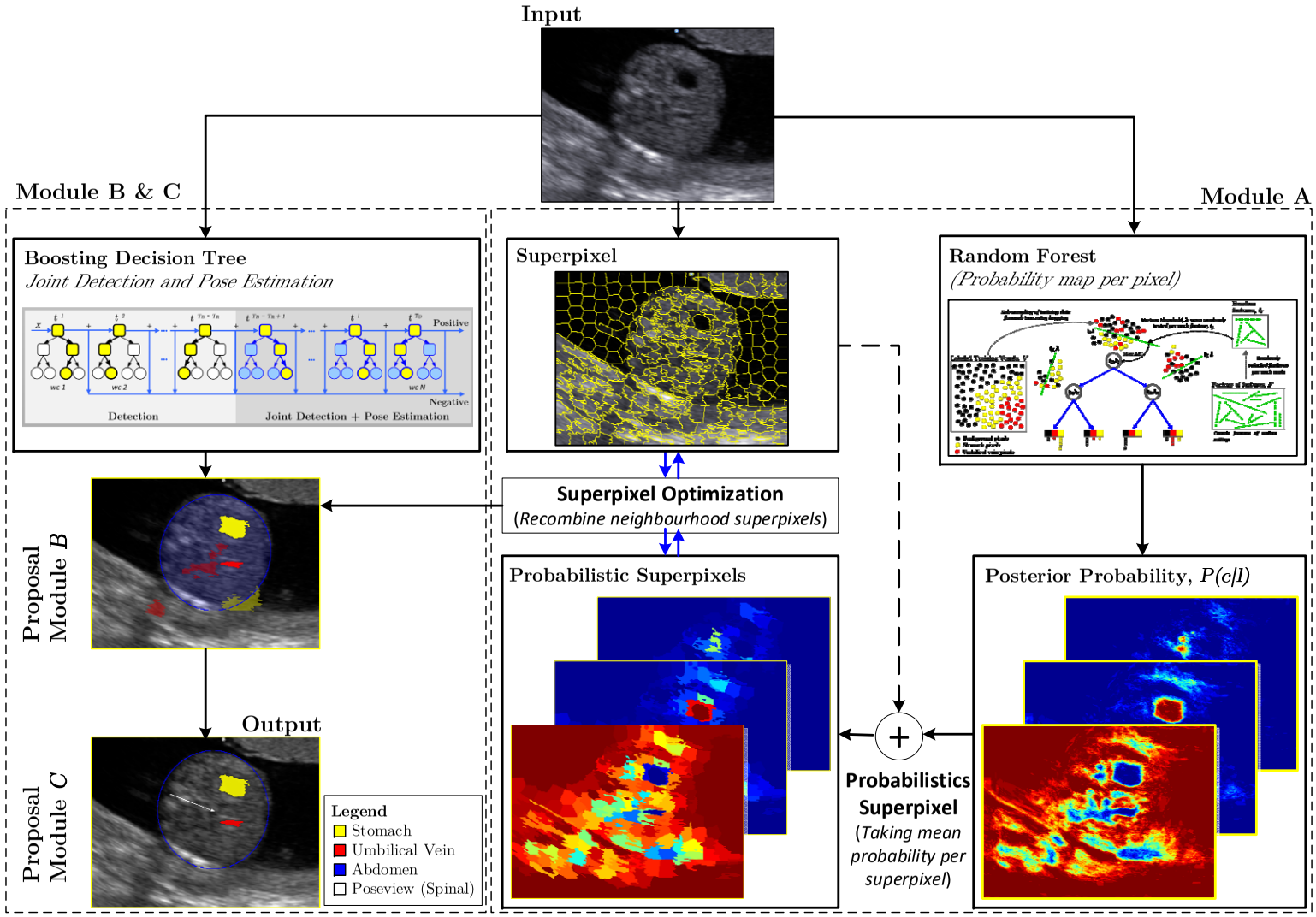


Figure 6.3: A unifying multiple proposal framework. The three proposals as proposed in this research work were learnt independently before jointly combined to classify the best diagnostic plane for biometric measurements.

6.2.2 Deep Learning Algorithm

Deep-leaning algorithms has recently become popular and has gained huge success in diverse fields of applications, not only among computer vision communities but also in medical image analysis. Moving forward, it will be interesting to investigate whether a deep learning architecture can be formulated from the insightful observation and recommendation of the work presented here to offer substantial improvement in US image analysis.

As argued by the work of HasanPour et al. (2018), deep learning methods such as Convolutional Neural Networks (CNNs) might suffer from ad-hoc designing procedure, and the tendency to introduce an unnecessary increase in depth and network complexity while working for a better accuracy. Hence, knowing the basic principles and how important these three aspects of computer vision: *recognition, re-organisation* and *reconstruction* are in achieving different aspects of a network might be the key to devising and improving the architecture of the deep learning algorithm and the performance of the system. In other words, a deep learning architecture might be designed to achieve customary goals by manoeuvring a different design space effectively. For example, one could investigate how interrelationships between these three aspects of vision may help the arrangement of convolutional layers or one can look into the usefulness of the feature encoding schemes in designing feature extraction, as recommended in this thesis.

Elsewhere, it is also interesting to blend the work of Ahmed & Noble (2015), which highlighted the importance of human visual system in identifying the salient region for further classification. The information obtained from eye-fixation experiment can be adapted into CNN architecture in building up convolutional layers that could resemble individual cells within the human visual system. Analogous to how a sonographer reconstruct a 3D mental visualization from 2D fetal images, the information generated from these fully connected layers could be harnessed to aid in building up the

geometrical relationship between key structures, which in turn could potentially improve the localisation result.

Other exciting exploratory works within deep learning paradigm might also include the work that employed a transfer learning and temporal features. Due to the limited availability of training data, a fully-supervised deep CNNs architecture might suffer from the problem of over-fitting. In solving this, some studies (Chen et al. 2015b; 2014b) have shown promising outcomes in exploiting the features learned and trained on natural images toward the domain of US imaging. Besides that, one could also explore and investigate the potential of temporal information on top of spatial information given a $2D + T$ video sweeps in regularising the classification results.

6.2.3 Generalisation to Other Imaging Plane

Experimental work presented in this thesis has been designed for analysis of the standardised fetal abdominal imaging plane. Work in fetal image analysis has been dominated by the work on fetal head imaging plane (Cuingnet et al. 2013, Maraci et al. 2013, Yaqub et al. 2017). Compared to the abdomen area, the anatomical landmarks used to describe a fetal head biometry such as cavum septum pellucidum and thalami are more visible and pose stronger geometrical relationships.

Hence, for this reason, the generalisation of our multiple proposal framework perhaps is more well suited and should be pursued for further investigations to other standardised imaging plane such as fetal head imaging plane.

6.2.4 Clinical Evaluation on 3D Ultrasound Dataset

There are two clear directions that would contribute to increasing the robustness of the proposed system. Firstly, with regards to the data collection, as commented earlier in Section 6.1.1, most of the collected data had the abdomen positioned in a certain direction; the preferred direction by a sonographer. Hence, a good number of clinical dataset containing multiple arbitrary view of the 2D images would be good to learn

the natural US pattern and to investigate the effect of different acquisition settings *e.g.* effect of the ultrasound beam pointing from a given direction.

Secondly, to test the robustness of our proposed framework, one would expect its successes in a natural scenario, which normally demonstrated in a video swipe of fetal ultrasound. Will the method be responsive enough to localise and score the anatomical structures while the sonographer spontaneously adjusts their transducer to numerous random direction? Hence, it would be interesting to extend the analysis of our proposed method to multiple sweeps US video to get an insight on how the information fusion process can assist in clinical decision-making. This would require experiments on newly acquired data using a freehand US sweep protocol (Maraci et al. 2014) as well as routine video scans from a hospital database (Yaqub et al. 2017).

Bibliography

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, (99), 1.

Ahmed, M., & Noble, J. A. (2015). Eye Tracking to Boost Recognition of Anatomical Features in Fetal Ultrasound. In *Medical Image Understanding and Analysis - MIUA 2015. 19th Annual Conference, University of Lincoln, Lincoln, UK, July 15-17, 2015.*, (pp. 112–117).

Ahmed, M., & Noble, J. A. (2016a). An Eye-Tracking Inspired Method for Standardised Plane Extraction From Fetal Abdominal Ultrasound Volumes. In *13th IEEE International Symposium on Biomedical Imaging, ISBI 2016, Prague, Czech Republic, April 13-16, 2016*, (pp. 1084–1087).

Ahmed, M., & Noble, J. A. (2016b). Fetal Ultrasound Image Classification Using a Bag-of-words Model Trained on Sonographers' Eye Movements. In *20th Conference on Medical Image Understanding and Analysis, MIUA 2016, Loughbrough University, Leicestershire, UK, July 6-8, 2016*, (pp. 157–162).

Alexe, B., Deselaers, T., & Ferrari, V. (2010). What is an object? In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, (pp. 73–80).

Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, *34*(11), 2189–2202.

Appel, R., Fuchs, T., Dollar, P., & Perona, P. (2013). Quickly Boosting Decision Trees – Pruning Underachieving Features Early. In S. Dasgupta, & D. McAllester (Eds.) *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, vol. 28, (pp. 594–602). JMLR Workshop and Conference Proceedings.

Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., & Malik, J. (2014). Multiscale Combinatorial Grouping. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 328–335).

Baumgartner, C. F., Kamnitsas, K., Matthew, J., Fletcher, T. P., Smith, S., Koch, L. M., Kainz, B., & Rueckert, D. (2017). SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Transactions on Medical Imaging*, *36*(11), 2204–2215.

Bernstein, I. M., Goran, M. I., Amini, S. B., & Catalano, P. M. (1997). Differential Growth of Fetal Tissues During The Second Half of Pregnancy. *American Journal of Obstetrics and Gynecology*, *176*(1, Part 1), 28 – 32.

Blum, L., & Kurtz, A. B. (1990). Gestational age: What to measure and when. *Seminars in Roentgenology*, *25*(4), 299–308.

Bourdev, L., & Brandt, J. (2005). Robust object detection via soft cascade. In *Computer Vision and Pattern Recognition (CVPR), 2005.*, vol. 2, (pp. 236–243 vol. 2).

Breiman, L. (1996). Bagging Predictors. In *Machine Learning*, vol. 24, (pp. 123–140).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Campbell, S., & Wilkin, D. (1975). Ultrasonic Measurement of Fetal Abdomen Circumference in the Estimation of Fetal Weight. *BJOG: An International Journal of Obstetrics and Gynaecology*, *82*(9), 689–697.

Carneiro, G., Amat, F., Georgescu, B., Good, S., & Comaniciu, D. (2008a). Semantic-Based Indexing of Fetal Anatomies From 3-D Ultrasound Data Using Global/Semi-Local Context and Sequential Sampling. In *CVPR*, (pp. 1 –8).

Carneiro, G., Georgescu, B., Good, S., & Comaniciu, D. (2007). Automatic Fetal Measurements in Ultrasound Using Constrained Probabilistic Boosting Tree. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, MICCAI'07, (pp. 571–579). Berlin, Heidelberg: Springer-Verlag.

Carneiro, G., Georgescu, B., Good, S., & Comaniciu, D. (2008b). Detection and Measurement of Fetal Anatomies from Ultrasound Images using a Constrained Probabilistic Boosting Tree. *IEEE Trans. on Medical Imaging*, *27*(9), 1342 –1355.

Carreira, J., & Sminchisescu, C. (2012). CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(7), 1312–1328.

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. In *ICML*, ICML '06, (pp. 161–168). New York, NY, USA: ACM.

Chalana, V., Thomas C., W. I., Dale R., C., Haynor, D. R., & Kim, Y. (1996). Automatic Fetal Head Measurements From Sonographic Images. *Academic Radiology*, *3*(8), 628 – 635.

Chang, T. C., Robson, S. C., Boys, R. J., & Spencer, J. A. D. (1992). Prediction of the small for gestational age infant: Which ultrasonic measurement is best? *Obstetrics and gynecology*, *80*(6), 1030–1038.

Chen, H., Dou, Q., Ni, D., Cheng, J.-Z., Qin, J., Li, S., & Heng, P.-A. (2015a). Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent

Neural Networks. In N. Navab, J. Hornegger, W. M. Wells, & A. Frangi (Eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, (pp. 507–514). Cham: Springer International Publishing.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., & Ann Heng, P. (2015b). *Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks*, vol. 19.

Chen, H., Ni, D., Wu, L., Qin, J., Li, S., & Ann Heng, P. (2014a). *Ultrasound Standard Plane Localization via Spatio-Temporal Feature Learning with Knowledge Transfer*.

Chen, H., Ni, D., Yang, X., Li, S., & Heng, P. A. (2014b). Fetal Abdominal Standard Plane Localization through Representation Learning with Knowledge Transfer. In G. Wu, D. Zhang, & L. Zhou (Eds.) *Machine Learning in Medical Imaging*, (pp. 125–132). Cham: Springer International Publishing.

Chen, Y. F., Huang, P. C., Lin, K. C., Lin, H. H., Wang, L. E., Cheng, C. C., Chen, T. P., Chan, Y. K., & Chiang, J. Y. (2014c). Semi-Automatic Segmentation and Classification of Pap Smear Cells. *IEEE Journal of Biomedical and Health Informatics*, *18*(1), 94–108.

Cheng, M. M., Zhang, Z., Lin, W. Y., & Torr, P. (2014). BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3286–3293).

Chitty, L. S., Altman, D. G., Henderson, A., & Campbell, S. (1994a). Charts of Fetal Size: 2. Head Measurements. *BJOG: An International Journal of Obstetrics and Gynaecology*, *101*(1), 35–43.

Chitty, L. S., Altman, D. G., Henderson, A., & Campbell, S. (1994b). Charts of Fetal Size: 3. Abdominal Measurements. *BJOG: An International Journal of Obstetrics and Gynaecology*, *101*(2), 125–131.

Chitty, L. S., Altman, D. G., Henderson, A., & Campbell, S. (1994c). Charts of Fetal Size: 4. Femur Length. *BJOG: An International Journal of Obstetrics and Gynaecology*, *101*(2), 132–135.

Ciurte, A., Rueda, S., Bresson, X., Nedevschi, S., T Papageorghiou, A., Noble, J., & Bach Cuadra, M. (2012). *Ultrasound Image Segmentation of the Fetal Abdomen: a Semi-Supervised Patch-Based Approach*.

Criminisi, A., Shotton, J., Robertson, D., & Konukoglu, E. (2011). Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In B. Menze, G. Langs, Z. Tu, & A. Criminisi (Eds.) *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, vol. 6533 of *Lecture Notes in Computer Science*, (pp. 106–117). Springer Berlin / Heidelberg.

Cui, X., Liu, Y., Shan, S., Chen, X., & Gao, W. (2007). 3D Haar-Like Features for Pedestrian Detection. In *IEEE International Conference on Multimedia and Expo*, (pp. 1263 –1266).

Cuingnet, R., Somphone, O., Mory, B., Prevost, R., Yaqub, M., Napolitano, R., Papageorghiou, A., Roundhill, D., Noble, J. A., & Ardon, R. (2013). Where is My Baby? A Fast Fetal Head Auto-Alignment in 3D-Ultrasound. In *2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, (pp. 768–771).

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR)*, vol. 1, (pp. 886 –893 vol. 1).

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Dollár, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast Feature Pyramids for Object Detection. *PAMI*.

Dollár, P., Appel, R., & Kienzle, W. (2012a). Crosstalk Cascades for Frame-Rate Pedestrian Detection. In *ECCV*.

Dollár, P., Belongie, S., & Perona, P. (2010). The Fastest Pedestrian Detector in the West. In *BMVC*.

Dollár, P., Tu, Z., Perona, P., & Belongie, S. (2009). Integral Channel Features. In *BMVC*.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2012b). Pedestrian Detection: An Evaluation of the State of the Art. *PAMI*, *34*.

Donner, R., Birngruber, E., Steiner, H., Bischof, H., & Langs, G. (2011). Localization of 3D Anatomical Structures Using Random Forests and Discrete Optimization. In B. Menze, G. Langs, Z. Tu, & A. Criminisi (Eds.) *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, vol. 6533 of *Lecture Notes in Computer Science*, (pp. 86–95). Springer Berlin / Heidelberg.

Dudley, N. J., & Chapman, E. (2002). The Importance of Quality Management in Fetal Measurement. *Ultrasound in Obstetrics and Gynecology*, *19*(2), 190–196.

Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J., & Zisserman, A. (2015). The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, *111*(1), 98–136.

Felsberg, M., & Sommer, G. (2001). The monogenic signal. *IEEE Transactions on Signal Processing*, *49*(12), 3136 –3144.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 1627–1645.

Feng, S., Zhou, S., Good, S., & Comaniciu, D. (2009). Automatic Fetal Face Detection From Ultrasound Volumes Via Learning 3D and 2D Information. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, (pp. 2488 –2495).

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119 – 139.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2008). Localizing Objects with Smart Dictionaries. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, (pp. 179–192). Berlin, Heidelberg: Springer-Verlag.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2009). Class Segmentation and Object Localization With Superpixel Neighborhoods. In *IEEE 12th International Conference on Computer Vision, (ICCV)*, (pp. 670 –677).

Gao, Y., Maraci, M. A., & Noble, J. A. (2016). Describing Ultrasound Video Content Using Deep Convolutional Neural Networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, (pp. 787–790).

Ghodrati, A., Pedersoli, M., & Tuytelaars, T. (2014). Is 2D Information Enough For Viewpoint Estimation? In *Proceedings of the British Machine Vision Conference*. BMVA Press.

Grady, L. (2006). Random Walks for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(11), 1768–1783.

Grossmann, E. (2004). AdaTree: Boosting a Weak Classifier into a Decision Tree. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, (pp. 105–105).

Guo, W., Xiao, Y., & Zhang, G. (2014). Multi-scale Pedestrian Detection by Use of AdaBoost Learning Algorithm. In *2014 International Conference on Virtual Reality and Visualization*, (pp. 266–271).

Hadlock, F., Deter, R., Harrist, R., & Park, S. (1982). Fetal abdominal circumference as a predictor of menstrual age. *American Journal of Roentgenology*, *139*(2), 367–370.

Hadlock, F. P., Harrist, R. B., Sharman, R. S., Deter, R. L., & Park, S. K. (1985). Estimation of Fetal Weight With the Use of Head, Body, and Femur Measurements–A Prospective Study. *Am J Obstet Gynecol*, *151*(3), 333–7.

Hanna, C., & Youssef, A. (1997). Automated Measurements in Obstetric Ultrasound Images. In *Proceedings. International Conference on Image Processing*, vol. 3, (pp. 504 –507 vol.3).

HasanPour, S. H., Rouhani, M., Fayyaz, M., Sabokrou, M., & Adeli, E. (2018). Towards Principled Design of Deep Convolutional Networks: Introducing SimpNet. *CoRR*.

Herman, A., Maymon, R., Dreazen, E., Caspi, E., Bukovsky, I., & Weinraub, Z. (1998). Nuchal Translucency Audit: A Novel Image-Scoring Method. *Ultrasound in Obstetrics and Gynecology*, *12*(6), 398–403.

Hogg, D. C. (1983). Model-based vision: a program to see a walking person. *Image Vision Comput.*, *1*, 5–20.

International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) (2013). ISUOG Practice Guidelines: performance of first-trimester fetal ultrasound scan. *Ultrasound in Obstetrics and Gynecology*, *41*(1), 102–113.

Jardim, S. M. G. V. B., & Figueiredo, M. A. T. (2005). Segmentation of fetal ultrasound images. *Ultrasound in Medicine & Biology*, *31*(2), 243–250.

Jin, M., Govindarajan, L. N., & Cheng, L. (2014). A Random-Forest Random Field Approach for Cellular Image Segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, (pp. 1251–1254).

Juránek, R., Herout, A., Dubská, M., & Zemcík, P. (2015). Real-Time Pose Estimation Piggybacked on Object Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*, (pp. 2381–2389).

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, *1*(4), 321–331.

Kovesi, P. (1997). Symmetry and Asymmetry from Local Phase. In *Tenth Australian Joint Converence on Artificial Intelligence*, (pp. 2–4).

Lampert, C. H., Blaschko, M. B., & Hofmann, T. (2009). Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2129–2142.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR)*, vol. 2, (pp. 2169 – 2178).

Lee, T., Fidler, S., & Dickinson, S. (2015). Learning to Combine Mid-Level Cues for Object Proposal Generation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, (pp. 1680–1688).

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust Object Detection with Interleaved Categorization and Segmentation. *Int. Journal Computer Vision*, *77*(1-3), 259–289.

Lempitsky, V., Verhoek, M., Noble, J. A., & Blake, A. (2009). Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3D Echocardiography. In *Proceedings of the 5th International Conference on Functional Imaging and Modeling of the Heart*, (pp. 447–456). Nice, France: Springer-Verlag.

Lepetit, V., Lagger, P., & Fua, P. (2005). Randomized Trees for Real-Time Keypoint Recognition. In *CVPR 2005*, vol. 2, (pp. 775 – 781).

Liebelt, J., & Schmid, C. (2010). Multi-View Object Class Detection With a 3D Geometric Model. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (pp. 1688–1695).

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision, (ICCV)*, vol. 2, (pp. 1150 –1157 vol.2).

Lu, W., Tan, J., & Floyd, R. (2005). Automated Fetal Head Detection and Measurement in Ultrasound Images by Iterative Randomized Hough Transform. *Ultrasound in Medicine & Biology*, *31*(7), 929 – 936.

Malik, J., Arbeláez, P., Carreira, J., Fragkiadaki, K., Girshick, R., Gkioxari, G., Gupta, S., Hariharan, B., Kar, A., & Tulsiani, S. (2016). The three R's of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, *72*, 4–14.

Maraci, A., Napolitano, R., T. Papageorghiou, A., & Noble, J. (2013). Automatic Fetal Head Detection On Video Clips From A Low-Cost Portable USB Ultrasound (US) Device. *Ultrasound in Obstetrics and Gynecology*, *42*.

Maraci, M. A., Napolitano, R., Papageorghiou, A., & Noble, J. A. (2014). Searching for Structures of Interest in an Ultrasound Video Sequence. In G. Wu, D. Zhang, & L. Zhou (Eds.) *Machine Learning in Medical Imaging*, (pp. 133–140). Cham: Springer International Publishing.

Marhaban, M. H., Kaid, R. S., & Mohd Noor, S. B. (2010). Automatic Estimation of Gestational Age in Ultrasound Images Based on Direct Least-Squares Fitting of Ellipse. *IEEJ Transactions on Electrical and Electronic Engineering*, *5*(5), 569–573.

Marszatek, M., & Schmid, C. (2007). Accurate Object Localization with Shape Masks. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, (pp. 1 –8).

Martin, A. F., Doddington, G. R., Kamm, T., Ordowski, M., & Przybocki, M. A. (1997). The DET curve in assessment of detection task performance. In *EUROSPEECH*.

Melamed, N., Yogev, Y., Meizner, I., Mashiach, R., Bardin, R., & Ben-Haroush, A. (2009). Sonographic Fetal Weight Estimation. *Journal of Ultrasound in Medicine*, *28*(5), 617–629.

Metz, C. E. (1978). Basic Principles of ROC Analysis. *Seminars in nuclear medicine*, *8 4*, 283–98.

Mulet-Parada, M., & Noble, J. (2000). 2D+T Acoustic Boundary Detection in Echocardiography. *Medical Image Analysis*, *4*(1), 21 – 30.

Murphy, K. P., Torralba, A., Eaton, D., & Freeman, W. T. (2006). Object Detection and Localization Using Local and Global Features. In *Toward Category-Level Object Recognition*, vol. 4170, (pp. 382–400). Springer.

Nahum, G. G., & Stanislaw, H. (2003). Ultrasonographic prediction of term birth weight: How accurate is it? *American Journal of Obstetrics and Gynecology*, *188*(2), 566 – 574.

Nakai, A., & Oya, A. (2002). Accuracy and Reproducibility of Ultrasound Measurements in Obstetric Management. *Gynecol Obstet Invest*, *54*(1), 31–36.

Namburete, A., Rahmatullah, B., & Noble, J. A. (2012). Nakagami-Based Choroid Plexus Detection in Fetal Ultrasound Images Using AdaBoost. In *Medical Image Understanding and Analysis (MIUA)*, (pp. 31–36).

Namburete, A. I. L., & Noble, J. A. (2013). Fetal Cranial Segmentation in 2D Ultrasound Images Using Shape Properties of Pixel Clusters. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, (pp. 720–723).

Neilson, J. P. (2008). Ultrasound for fetal assessment in early pregnancy. *Cochrane Database of Systematic Review*.

NHS Fetal Anomaly Screening Programme (NHS FASP) (2010). *NHS Fetal Anomaly Screening Programme - 18 to 20 Weeks Fetal Anomaly Scan National Standards and Guidance for England*.

NHS Fetal Anomaly Screening Programme (NHS FASP) (2015). *Fetal Anomaly Screening Programme Handbook for Ultrasound Practitioners*.

Ni, D., Li, T., Yang, X., Qin, J., Li, S., Chin, C.-T., Ouyang, S., Wang, T., & Chen, S. (2013). Selective Search and Sequential Detection for Standard Plane Localization in Ultrasound. In H. Yoshida, S. Warfield, & M. Vannier (Eds.) *Abdominal Imaging*.

Computation and Clinical Applications, vol. 8198 of *Lecture Notes in Computer Science*, (pp. 203–211). Springer Berlin Heidelberg.

Ni, D., Yang, X., Chen, X., Chin, C.-T., Chen, S., Heng, P. A., Li, S., Qin, J., & Wang, T. (2014). Standard Plane Localization in Ultrasound by Radial Component Model and Selective Search. *Ultrasound in Medicine and Biology*, *40*(11), 2728–2742.

Nithya, J., & Madheswaran, M. (2009). Detection of Intrauterine Growth Retardation Using Fetal Abdominal Circumference. In *2009 International Conference on Computer Technology and Development*, vol. 2, (pp. 371–375).

Noble, J. A. (2010). Ultrasound image segmentation and tissue characterization. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, *224*(2), 307–316.

Ozuysal, M., Lepetit, V., & Fua, P. (2009). Pose estimation for category specific multiview object localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 778–785).

Pantofaru, C., Dorko, G., Schmid, C., & Hebert, M. (2006). Combining Regions and Patches for Object Class Localization. In *CVPR*, (pp. 23–). Washington, DC, USA.

Pantofaru, C., Schmid, C., & Hebert, M. (2008). Object Recognition by Integrating Multiple Image Segmentations. In *ECCV*, (pp. 481–494). Berlin, Heidelberg: Springer-Verlag.

Papageorghiou, A., Sarris, I., Ioannou, C., Todros, T., Carvalho, M., Pilu, G., Salomon, L., for the International Fetal, & for the 21st Century (INTERGROWTH-21st), N. G. C. (2013). Ultrasound Methodology Used To Construct the Fetal Growth Standards in The INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics and Gynaecology*, *120*, 27–32.

Papageorghiou, A. T., Ohuma, E. O., Altman, D. G., Todros, T., Ismail, L. C., Lambert, A., Jaffer, Y. A., Bertino, E., Gravett, M. G., Purwar, M., Noble, J. A., Pang, R.,

Victora, C. G., Barros, F. C., Carvalho, M., Salomon, L. J., Bhutta, Z. A., Kennedy, S. H., & Villar, J. (2014). International Standards for Fetal Growth based on Serial Ultrasound Measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*, *384*(9946), 869 – 879.

Papageorgiou, C., Oren, M., & Poggio, T. (1998). A General Framework for Object Detection. In *Sixth International Conference on Computer Vision, (ICCV)*, (pp. 555 –562).

Pathak, S. D., Chalana, V., & Kim, Y. (1997). Interactive Automatic Fetal Head Measurements From Ultrasound Images Using Multimedia Computer Technology. *Ultrasound in Medicine & Biology*, *23*(5), 665–673.

Peng, X., Setlur, S., Govindaraju, V., & Sitaram, R. (2010). Text Separation from Mixed Documents Using a Tree-Structured Classifier. In *2010 20th International Conference on Pattern Recognition*, (pp. 241–244).

Perronnin, F., Liu, Y., Sánchez, J., & Poirier, H. (2010). Large-Scale Image Retrieval With Compressed Fisher Vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.*, (pp. 3384–3391).

Piccardi, M. (2004). Background Subtraction Techniques: A Review. In *2004 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, (pp. 3099–3104 vol.4).

Rahmatullah, B. (2012). *Assessment of obstetric ultrasound images using machine learning*. Theses, University of Oxford.

Rahmatullah, B., Papageorghiou, A., & Noble, J. (2011a). Automated Selection of Standardized Planes from Ultrasound Volume. In *Machine Learning in Medical Imaging*, vol. 7009, (pp. 35–42). Springer Berlin / Heidelberg.

Rahmatullah, B., Papageorghiou, A., & Noble, J. A. (2012a). Image Analysis Using Machine Learning: Anatomical Landmark Detection in Fetal Ultrasound Images.

In *IEEE Proceedings of the Annual Computer Software and Applications Conference (COMPSAC)*. Izmir, Turkey.

Rahmatullah, B., Papageorghiou, A., & Noble, J. A. (2012b). Integration of local and global features for anatomical object detection in ultrasound. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, vol. To appear. Nice, France.

Rahmatullah, B., Papageorghiou, A., & Noble, J. A. (2012c). Multi-Scale Local Phase Features for Anatomical Object Detection in Fetal Ultrasound Images. In *Proceedings of the Medical Image Understanding and Analysis (MIUA)*, (pp. 199–204). Swansea, UK.

Rahmatullah, B., Sarris, I., Papageorghiou, A., & Noble, J. A. (2011b). Quality Control of Fetal Ultrasound Images: Detection of Abdomen Anatomical Landmarks Using AdaBoost. In *International Symposium on Biomedical Imaging (ISBI)*, (pp. 6–9).

Rajpoot, K., Grau, V., & Noble, J. (2009). Local-Phase Based 3D Boundary Detection Using Monogenic Signal and Its Application To Real-Time 3-D Echocardiography Images. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, (ISBI)*, (pp. 783 –786).

Rantalankila, P., Kannala, J., & Rahtu, E. (2014). Generating Object Segmentation Proposals Using Global and Local Search. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 2417–2424).

Rawat, V., Shrimali, V., & Jain, A. (2018). *Automated Techniques for the Interpretation of Fetal Abnormalities: A Review*, vol. 2018.

Redondo Cabrera, C., Lopez Sastre, R., & Tuytelaars, T. (2014). All together now: Simultaneous Detection and Continuous Pose Estimation using a Hough Forest with Probabilistic Locally Enhanced Voting. In *Proceedings of the British Machine Vision Conference*. BMVA Press.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 2nd ed.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.*, *73*, 814–817.

Rueda, S., Fathima, S., Knight, C., Yaqub, M., Papageorghiou, A., Rahmatullah, B., Foi, A., Maggioni, M., Pepe, A., Tohka, J., Stebbing, R., McManigle, J., Ciurte, A., Bresson, X., Cuadra, M., Sun, C., Ponomarev, G., Gelfand, M., Kazanov, M., wei Wang, C., Chen, H.-C., Peng, C.-W., Hung, C.-M., & Noble, J. (2014). Evaluation and Comparison of Current Fetal Ultrasound Image Segmentation Methods for Biometric Measurements: A Grand Challenge. vol. 33, (pp. 797–813).

Rueda, S., L Knight, C., T Papageorghiou, A., & Noble, J. (2015). *Feature-Based Fuzzy Connectedness Segmentation of Ultrasound Images with an Object Completion Step*, vol. 26.

Sairam, S., Awadh, A. M. A., Cook, K., Papageorghiou, A. T., & Carvalho, J. S. (2009). Impact of Audit of Routine Second-Trimester Cardiac Images Using a Novel Image-Scoring Method. *Ultrasound in Obstetrics and Gynecology*, *33*(5), 545–551.

Salomon, L. J., Bernard, J. P., Duyme, M., Doris, B., Mas, N., & Ville, Y. (2006). Feasibility and Reproducibility of An Image-Scoring Method For Quality Control of Fetal Biometry in The Second Trimester. *Ultrasound in O&G*, *27*(1), 34–40.

Salomon, L. J., Nassar, M., Bernard, J. P., Ville, Y., & Fauconnier, A. (2009). A Score-Based Method To Improve The Quality of Emergency Gynaecological Ultrasound Examination. *European journal of obstetrics, gynecology, and reproductive biology*, *143*(2), 116–20.

Salomon, L. J., & Ville, Y. (2009). The science and art of quality in obstetric ultrasound. *Current Opinion in O&G*, *21*(2), 153–160.

Salomon, L. J., Winer, N., Bernard, J. P., & Ville, Y. (2008). A Score-Based Method for Quality Control of Fetal Images at Routine Second-Trimester Ultrasound Examination. *Prenatal Diagnosis*, *28*(9), 822–827.

Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D. G., Papageorghiou, A. T., for the International Fetal, & for the 21st Century (INTERGROWTH-21st), N. G. C. (2012). Intra- and Interobserver Variability in Fetal Ultrasound Measurements. *Ultrasound in Obstetrics and Gynecology*, *39*(3), 266–273.

Sarris, I., Ioannou, C., Dighe, M., Mitidieri, A., Oberto, M., Qingqing, W., Shah, J., Sohoni, S., Al Zidjali, W., Hoch, L., Altman, D. G., Papageorghiou, A. T., for the International Fetal, & for the 21st Century (INTERGROWTH-21st), N. G. C. (2011). Standardization of Fetal Ultrasound Biometry Measurements: Improving the Quality and Consistency of Measurements. *Ultrasound in Obstetrics and Gynecology*, *38*(6), 681–687.

Sarris, I., Ioannou, C., Ohuma, E., Altman, D., Hoch, L., Cosgrove, C., Fathima, S., Salomon, L., Papageorghiou, A., for the International Fetal, & for the 21st Century (INTERGROWTH-21st), N. G. C. (2013). Standardisation and Quality Control of Ultrasound Measurements Taken in the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics and Gynaecology*, *120*, 33–37.

Savarese, S., & Fei-Fei, L. (2007). 3D generic object categorization, localization and pose estimation. In *2007 IEEE 11th International Conference on Computer Vision*, (pp. 1–8).

Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine Learning*, *5*(2), 197–227.

Schick, A., Bäuml, M., & Stiefelhagen, R. (2012). Improving Foreground Segmentations with Probabilistic Superpixel Markov Random Fields. In *IEEE Workshop on Change Detection 2012, CVPR*.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248.

Shotton, J., Johnson, M., & Cipolla, R. (2008). Semantic Texton Forests for Image Categorization and Segmentation. In *CVPR*, (pp. 1 –8).

Sivic, J., & Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. In *Proceedings. Ninth IEEE International Conference on Computer Vision, (ICCV)*, (pp. 1470 –1477 vol.2).

Smith, B. J., & Arabshahi, P. (1996). A fuzzy decision system for ultrasonic prenatal examination enhancement. In *Proceedings of IEEE 5th International Fuzzy Systems*, vol. 3, (pp. 1712–1717 vol.3).

Sofka, M., Zhang, J., Zhou, S., & Comaniciu, D. (2010). Multiple Object Detection by Sequential Monte Carlo and Hierarchical Detection Network. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, (pp. 1735 –1742).

Stevenson, G. N., Collins, S. L., Ding, J., Impey, L., & Noble, J. A. (2015). 3-D Ultrasound Segmentation of the Placenta Using the Random Walker Algorithm: Reliability and Agreement. *Ultrasound in Medicine and Biology*, *41*(12), 3182–3193.

Teuber, H. L. (1955). Physiological Psychology. *Annual Review of Psychology*, *6*(1), 267–296.

The American Institute of Ultrasound in Medicine (AIUM) (2010). AIUM Practice Guideline for the Performance of Obstetric Ultrasound Examinations. *Journal of Ultrasound in Medicine*, *29*(1), 157–166.

Thomas, J., Peters, I., R.A., & Jeanty, P. (1991). Automatic Segmentation of Ultrasound Images Using Morphological Operators. *IEEE Transactions on Medical Imaging*, *10*(2), 180 –186.

Tin Kam, H. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1, (pp. 278–282 vol.1).

Tin Kam, H. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.

Tola, E., Lepetit, V., & Fua, P. (2008). A Fast Local Descriptor for Dense Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.*, (pp. 1–8).

Tola, E., Lepetit, V., & Fua, P. (2010). DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(5), 815–830.

Tolay, P., Vajinepalli, P., Bhattacharya, P., Firtion, C., & Sisodia, R. S. (2009). Automated Fetal Spine Detection in Ultrasound Images. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 7260.

Tu, Z. (2005). Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, (pp. 1589–1596). Washington, DC, USA: IEEE Computer Society.

Tulsiani, S., & Malik, J. (2015). Viewpoints and Keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tuytelaars, T., & Mikolajczyk, K. (2008). *Local Invariant Feature Detectors: A Survey*. Now Publishers Inc.

van de Sande, K., Uijlings, J., Gevers, T., & Smeulders, A. (2011). Segmentation as selective search for object recognition. In *IEEE International Conference on Computer Vision, (ICCV)*, (pp. 1879 –1886).

Vedaldi, A., & Fulkerson, B. (2008). VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/.

Villar, J., Altman, D., Purwar, M., Noble, J., Knight, H., Ruyan, P., Cheikh Ismail, L., Barros, F., Lambert, A., Papageorghiou, A., Carvalho, M., Jaffer, Y., Bertino, E.,

Gravett, M., Bhutta, Z., Kennedy, S., for the International Fetal, & for the 21st Century (INTERGROWTH-21st), N. G. C. (2013). The objectives, design and implementation of the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics and Gynaecology*, *120*, 9–26.

Villar, J., Ismail, L. C., Victora, C. G., Ohuma, E. O., Bertino, E., Altman, D. G., Lambert, A., Papageorghiou, A. T., Carvalho, M., Jaffer, Y. A., Gravett, M. G., Purwar, M., Frederick, I. O., Noble, A. J., Pang, R., Barros, F. C., Chumlea, C., Bhutta, Z. A., & Kennedy, S. H. (2014a). International Standards For Newborn Weight, Length, and Head Circumference By Gestational Age And Sex: The Newborn Cross-Sectional Study Of The INTERGROWTH-21st Project. *The Lancet*, *384*(9946), 857 – 868.

Villar, J., Papageorghiou, A. T., Pang, R., Ohuma, E. O., Ismail, L. C., Barros, F. C., Lambert, A., Carvalho, M., Jaffer, Y. A., Bertino, E., Gravett, M. G., Altman, D. G., Purwar, M., Frederick, I. O., Noble, J. A., Victora, C. G., Bhutta, Z. A., & Kennedy, S. H. (2014b). The Likeness of Fetal Growth and Newborn Size Across Non-Isolated Populations in the INTERGROWTH-21st Project: the Fetal Growth Longitudinal Study and Newborn Cross-Sectional Study. *The Lancet Diabetes & Endocrinology*, *2*(10), 781 – 792.

Ville, Y. (2008). ´Ceci n´est pas une échographie´: A Plea for Quality Assessment in Prenatal Ultrasound. *Ultrasound in Obstetrics and Gynecology*, *31*(1), 1–5.

Viola, P., & Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR)*, vol. 1, (pp. 511–518).

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*, 137–154.

Wertheimer, M. (1938). *Laws of Organization in Perceptual Forms*. London: Harcourt, Brace & Jovanovitch.

Xiang, Y., Mottaghi, R., & Savarese, S. (2014). Beyond PASCAL: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, (pp. 75–82).

Yang, B., Yan, J., Lei, Z., & Li, S. Z. (2014a). Aggregate Channel Features for Multi-View Face Detection. In *IEEE International Joint Conference on Biometrics*, (pp. 1–8).

Yang, X., Ni, D., Qin, J., Li, S., Wang, T., Chen, S., & Heng, P. A. (2014b). Standard Plane Localization in Ultrasound By Radial Component. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, (pp. 1180–1183).

Yang, Y., & Ramanan, D. (2011). Articulated Pose Estimation With Flexible Mixtures-of-Parts. In *CVPR 2011*, (pp. 1385–1392).

Yaqub, M., Javaid, M., Cooper, C., & Noble, J. (2011). Improving the Classification Accuracy of the Classic RF Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation. In *Machine Learning in Medical Imaging*, vol. 7009, (pp. 184–192). Springer Berlin / Heidelberg.

Yaqub, M., Javaid, M. K., Cooper, C., & Noble, J. A. (2014). Investigation of the Role of Feature Selection and Weighted Voting in Random Forests for 3-D Volumetric Segmentation. *IEEE Transactions on Medical Imaging*, *33*(2), 258–271.

Yaqub, M., Kelly, B., Papageorghiou, A. T., & Noble, J. A. (2017). A Deep Learning Solution for Automatic Fetal Neurosonographic Diagnostic Plane Verification Using Clinical Standard Constraints. *Ultrasound in Medicine & Biology*, *43*(12), 2925 – 2933.

Yaqub, M., Napolitano, R., Ioannou, C., Papageorghiou, A. T., & Noble, J. A. (2012). Automatic Detection of Local Fetal Brain Structures in Ultrasound Images. In *International Symposium on Biomedical Imaging (ISBI)*, (pp. 1555 –1558).

Yu, J., Wang, Y., & Chen, P. (2008a). Fetal Ultrasound Image Segmentation System And Its Use in Fetal Weight Estimation. *Medical and Biological Engineering and Computing*, *46*, 1227–1237. 10.1007/s11517-008-0407-y.

Yu, J., Wang, Y., Chen, P., & Shen, Y. (2008b). Fetal Abdominal Contour Extraction and Measurement in Ultrasound Images. *Ultrasound in Medicine & Biology*, *34*(2), 169 – 182.

Zheng, Y., Georgescu, B., & Comaniciu, D. (2009a). Marginal Space Learning for Efficient Detection of 2D/3D Anatomical Structures in Medical Images. In *Proceedings of the 21st International Conference on Information Processing in Medical Imaging*, IPMI '09, (pp. 411–422). Berlin, Heidelberg: Springer-Verlag.

Zheng, Y., Georgescu, B., Ling, H., Zhou, S., Scheuering, M., & Comaniciu, D. (2009b). Constrained Marginal Space Learning For Efficient 3D Anatomical Structure Detection in Medical Images. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, (pp. 194 –201).

Zhu, X., & Ramanan, D. (2012). Face Detection, Pose Estimation, and Landmark Localization in the Wild. In *Computer Vision and Pattern Recognition, (CVPR)*. Providence, Rhode Island.

Zitnick, C. L., & Dollár, P. (2014). Edge Boxes: Locating Object Proposals from Edges. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.) *Computer Vision – ECCV 2014*, (pp. 391–405). Cham: Springer International Publishing.