Eye Tracking to Aid Fetal Ultrasound Image Analysis

Maryam Ahmed

Wolfson College
University of Oxford

Submitted in partial completion of the Doctor of Philosophy

Trinity 2017

Abstract

Current automated fetal ultrasound (US) analysis methods employ local descriptors and machine learning frameworks to identify salient image regions. This 'bottom-up' approach has limitations, as structures identified by local descriptors are not necessarily anatomically salient. In contrast, the human visual system employs a 'top-down' approach to image analysis guided primarily by image context and prior knowledge. This thesis attempts to bridge the gap between top-down and bottom-up approaches to US image analysis. We conduct eye tracking experiments to determine which local descriptors and global constraints guide the visual attention of human observers interpreting fetal US images. We then implement machine learning frameworks which mimic observers' visual search strategies for anatomical landmark localisation, standardised image plane selection, and video classification. We first developed a framework for landmark localisation in 2-D fetal abdominal US images. Informed by the eye movements of observers searching for anatomical landmarks in images, we derived a pictorial structures model which achieved mean detection accuracies of 87.2% and 83.2% for the stomach bubble and umbilical vein. We extended this framework to automate standardised imaging plane detection in 3-D fetal abdominal US volumes, achieving a mean standardised plane detection accuracy of 92.5%. We then implemented a bag-of-visual-words model for 2-D+t fetal US video clip classification. We recorded the eye movements of observers tasked with classifying videos, and trained a feed-forward neural network directly on eye tracking data to predict visually salient regions in unseen videos. This perception inspired spatiotemporal interest point operator was used within a framework for the classification of fetal US video clips, achieving 80.0% mean accuracy. This work constitutes the first demonstration that high-level constraints and visual saliency models obtained through eye tracking experiments can improve the accuracy of machine learning frameworks for US image analysis.

Eye Tracking to Aid Fetal Ultrasound Image Analysis

Maryam Ahmed

Wolfson College

University of Oxford

Submitted in partial completion of the

Doctor of Philosophy

Trinity 2017

Abstract

Current automated fetal ultrasound (US) analysis methods employ local descriptors and machine learning frameworks to identify salient image regions. This 'bottom-up' approach has limitations, as structures identified by local descriptors are not necessarily anatomically salient. In contrast, the human visual system employs a 'top-down' approach to image analysis guided primarily by image context and prior knowledge. This thesis attempts to bridge the gap between top-down and bottom-up approaches to US image analysis. We conduct eye tracking experiments to determine which local descriptors and global constraints guide the visual attention of human observers interpreting fetal US images. We then implement machine learning frameworks which mimic observers' visual search strategies for anatomical landmark localisation, standardised image plane selection, and video classification. We first developed a framework for landmark localisation in 2-D fetal abdominal US images. Informed by the eye movements of observers searching for anatomical landmarks in images, we derived a pictorial structures model which achieved mean detection accuracies of 87.2% and 83.2% for the stomach bubble and umbilical vein. We extended this framework to automate standardised imaging plane detection in 3-D fetal abdominal US volumes, achieving a mean standardised plane detection accuracy of 92.5%. We then implemented a bag-of-visual-words model for 2-D+t fetal US video clip classification. We recorded the eye movements of observers tasked with classifying videos, and trained a feed-forward neural network directly on eye tracking data to predict visually salient regions in unseen videos. This perception inspired spatiotemporal interest point operator was used within a framework for the classification of fetal US video clips, achieving 80.0% mean accuracy. This work constitutes the first demonstration that high-level constraints and visual saliency models obtained through eye tracking experiments can improve the accuracy of machine learning frameworks for US image analysis.

Contents

List of Figures

List of Tables

Glossary of Key Terms

US . . . . . B-Mode fetal ultrasound imaging.

2-D US . . . . . An ultrasound image with two spatial dimensions.

3-D US . . . . . An ultrasound image with three spatial dimensions, otherwise referred to as an ultrasound volume.

2D+t US . . . . . A temporal sequence of 2-D ultrasound images, or frames, otherwise referred to as an ultrasound video.

AC . . . . . Abdominal circumference; a fetal biometric measurement obtained via ultrasound imaging.

Static consistency . . . A measure of the spatial similarity between the eye movements of one or more observers.

Dynamic consistency . A measure of the temporal similarity between the eye movements of one or more observers.

Global-focal search . . A proposed model of human visual search in images consisting of a rapid image search, followed by a detailed inspection of potential targets, and finally a cross-referencing stage.

BoVW . . . . . Bag-of-visual-words, a classification framework frequently used for image and video classification.

SVM . . . . . Support vector machine, a form of binary classifier trained by finding the optimal separating hyperplane between two sets of data.

Thesis: A story all about how my life got flipped, turned upside down.

— The Fresh Prince of Bel-Air (1990-1996)

1 Introduction

Contents

1.1 Clinical Motivation

Routine clinical ultrasound (US) scans are crucial for fetal growth monitoring. Biometric measurements are obtained from standardised imaging planes (Figure 1.1) as defined, for instance, by the International Society for Obstetrics and Gynecology (ISUOG)[1] and compared against standardised growth charts for the diagnosis of small-for-gestational-age (SGA) fetuses and the identification of risk in pregnancy[2–4].

The manual acquisition of standardised imaging planes poses several challenges. The quality of US images is inherently variable due to regions of poor contrast, signal attenuation, acoustic shadows, noise and artefacts. The position, orientation and shape of the key anatomical landmarks, such as the umbilical vein and stomach bubble, is highly variable. Image quality reduces with increased gestational age

(GA) due to increased maternal adipose tissue and fetal size, and the manipulation of the US probe to identify anatomical features and obtain the correct plane is a highly skilled and complex task, with sonographers' scanning techniques varying significantly both on an inter-operator and intra-operator basis[5].

Automated fetal US analysis methods are not currently used within clinical practice but are being investigated in research settings[6,7]. Some of the proposed methods employ local descriptors and machine learning frameworks to localise anatomical landmarks in 2-D US images, detect standardised imaging planes in 3-D US volumes, and identify structures of interest in 2-D+t US videos. However this 'bottom-up' approach has limitations, as structures identified by local descriptors are not necessarily anatomically salient and may lead to the misclassification of artefacts as anatomical landmarks. In contrast, the human visual system employs a 'top-down' approach to image analysis guided primarily by image context and prior knowledge[8]. There remains, therefore, a significant disparity between human and machine learning driven strategies for US image analysis.

This thesis attempts to bridge the gap between top-down and bottom-up approaches to fetal US image analysis. Insights into the human visual system are obtained via eye tracking to address two primary challenges associated with existing US analysis methods, namely the need for more anatomically meaningful and discriminative interest point operators, and the need to incorporate high level constraints and prior knowledge of fetal anatomy into classifier design. Specifically, eye tracking experiments are conducted to determine which local descriptors and global constraints guide the visual attention of human observers interpreting fetal US images, volumes and videos. Machine learning frameworks are then implemented to mimic observers' visual search strategies for anatomical landmark localisation, standardised image plane selection, and video classification.

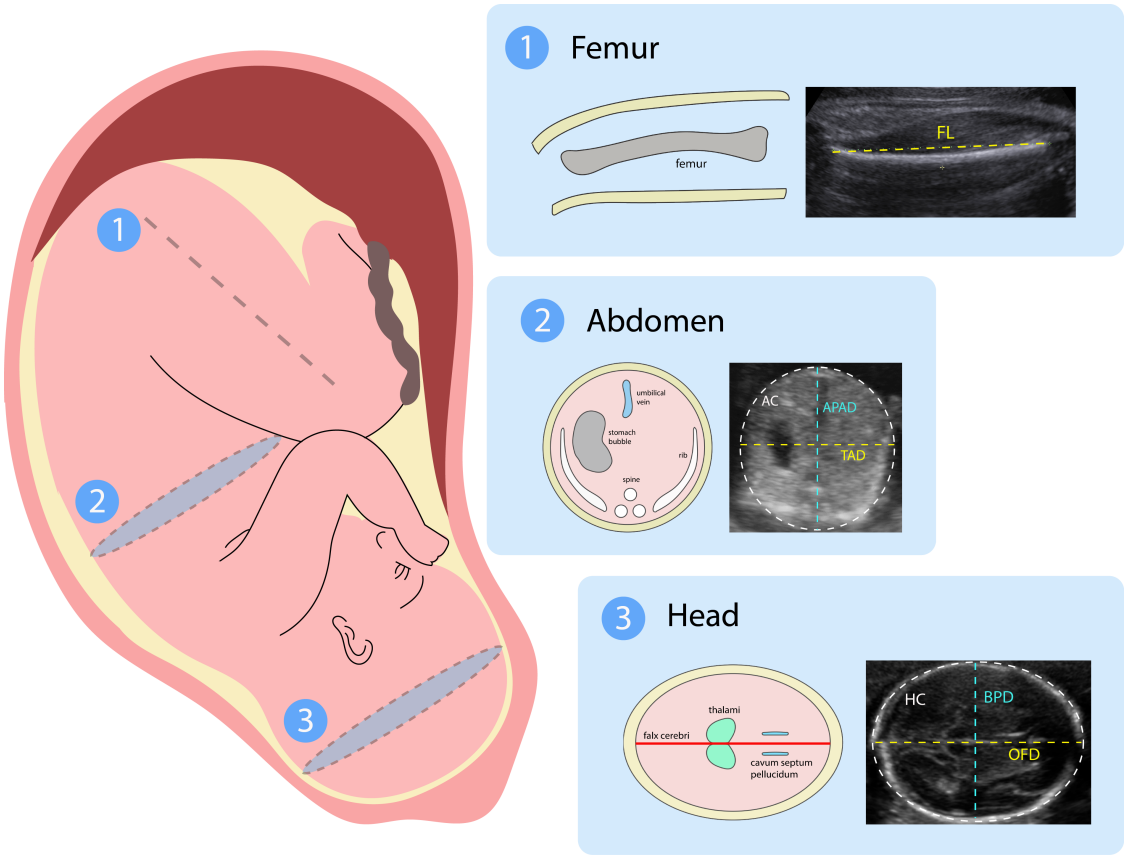


Figure 1.1: Schematic diagram showing key fetal biometric measurements taken from standardised planes through the fetal head, including head circumference (HC), bi-parietal diameter (BPD) and occipito-frontal diameter (OFD), the fetal abdomen, including abdominal circumference (AC), transverse abdominal diameter (TAD) and anterio-posterior abdominal diameter (APAD), and the femur, namely femur length (FL).

1.2 Contributions

This thesis contributes a series of eye tracking inspired methods for anatomical landmark localisation in 2-D US images, standardised plane detection in 3-D US volumes, and 2-D+t US video classification.

The first contribution consists of an eye tracking inspired method for the localisation of the stomach bubble and umbilical vein in 2-D fetal abdominal US images (Figure 1.2). Eye movements were recorded for ten expert and novice observers searching for the stomach bubble and umbilical vein in 150 2-D fetal abdominal US images. Relatively high spatial (73.8%) and temporal (37.8%) similarities[9] were established between the eye movements of observers, and the 'global-focal' model[10] of visual search was validated on the US eye tracking

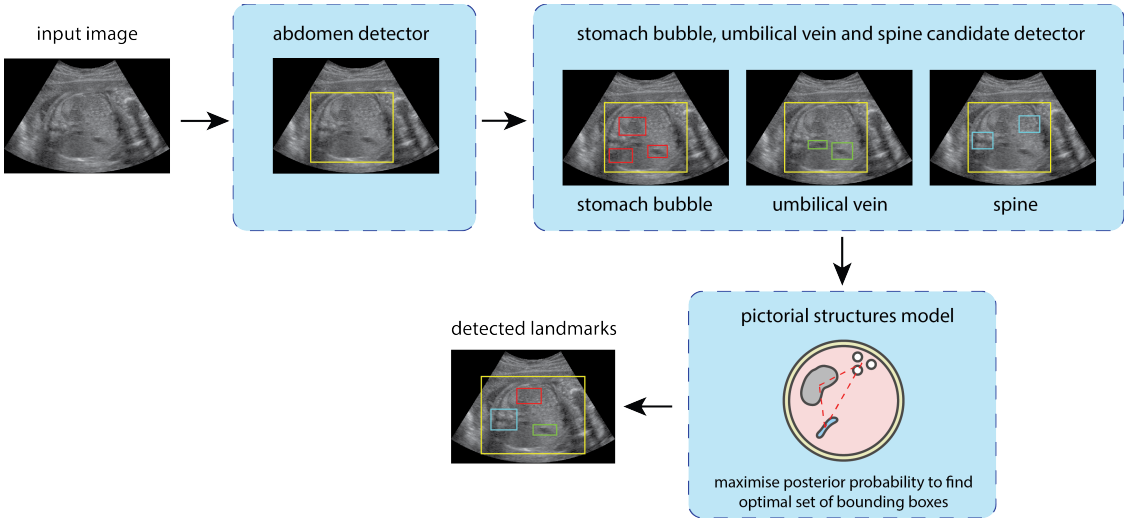


Figure 1.2: A schematic overview of the 2-D pictorial structures model developed for the automated localisation of the fetal stomach bubble and umbilical vein in 2-D US images

dataset. A pictorial structures model[11] was then derived to mimic the visual search behavior of observers: candidate anatomical landmarks were identified by a sliding window detector trained on a boosted set of decision stumps, and their optimal configuration was determined by a probabilistic model trained on the relative positions of anatomical landmarks. When trained on 1000 images and tested on a further independent set of 250 images, the framework achieved detection accuracies of 87.2% and 83.2% for the stomach bubble and umbilical vein respectively, an improvement on existing methods[6].

The pictorial structures model was then extended to develop an eye tracking inspired method for standardised plane selection in 3-D fetal abdominal US volumes (Figure 1.3). Eye movements were recorded for ten observers searching for the standardised abdominal plane in 150 volumes and, based on analysis of the visual search strategies of observers, optical flow and length descriptors were employed to act as 3-D constraints on the positions of anatomical landmarks between consecutive volume frames. A dynamic programming algorithm was then implemented for efficient standardised plane selection. When trained on 200 volumes and tested on a further independent set of 80 volumes, the framework achieved standardised plane selection accuracies of 92.5%, again an improvement on existing methods[7].

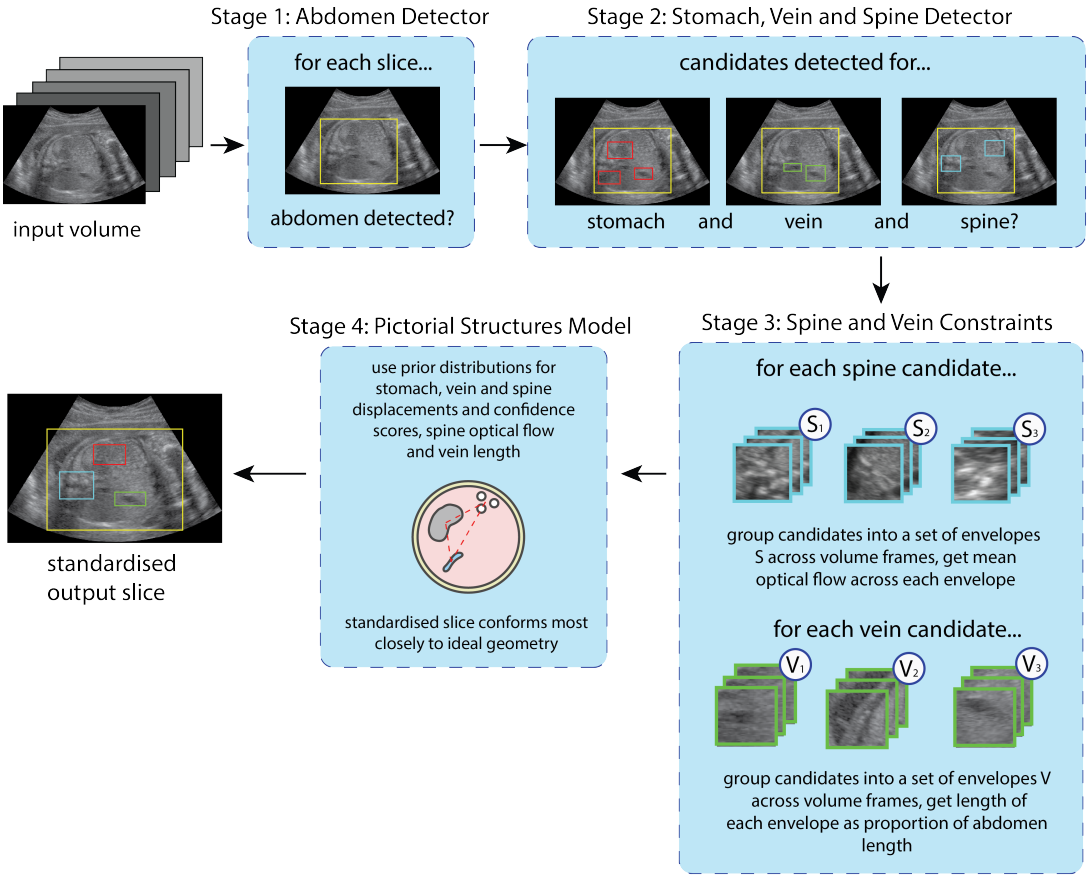


Figure 1.3: A schematic overview of the pictorial structures model developed for the automated detection of standardised abdominal planes in 3-D US volumes, incorporating optical flow and length constraints on the fetal spine and umbilical vein.

Finally a bag-of-visual-words (BoVW) model was implemented for the classification of 2-D+t fetal US video clips, with a spatio-temporal interest point operator learned directly from eye movements[12] (Figure 1.4). Eye movements were recorded for ten observers tasked with classifying 60 fetal US video clips. A low correspondence was found between the fixated locations of observers and salient video regions identified by the spatio-temporal Harris and Periodic (respectively $AUC = 0.55$ and $AUC = 0.57$) operators. Subsequently a feed-forward neural network was trained directly on fixated locations to classify visually salient regions in unseen videos ($AUC = 0.66$), and the learned operator was used within a BoVW pipeline for the classification of 240 fetal US video clips with 80.0% mean accuracy, compared with 57.50% and 71.67% attained using the Harris and Periodic operators respectively. This is the first demonstration that an interest point operator

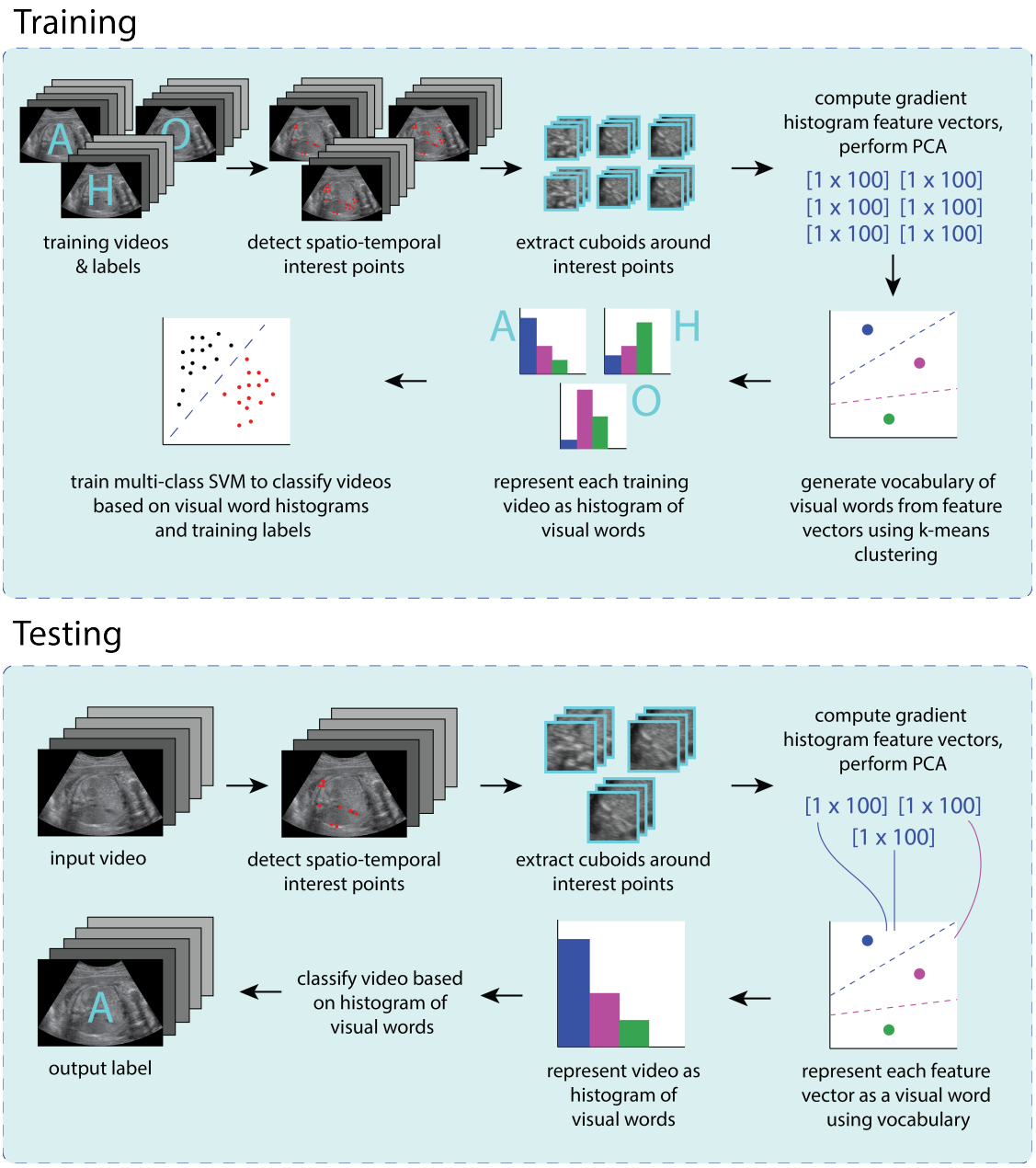


Figure 1.4: A schematic overview of the bag-of-visual-words model incorporating a perception inspired spatio-temporal interest point operator for the automated classification of 2-D+t video clips showing various portions of the fetal anatomy.

inspired by human perception can better identify visually salient US video regions than conventional interest point operators, and the first use of eye tracking to inform US video classification.

1.3 Thesis Structure

Chapter 1 outlines the clinical motivation for the research to follow, the structure of this thesis, and the original contributions and publications arising from the thesis at the time of submission.

Chapters 2 & 3 describe key motivations and theory underpinning this thesis. Specifically:

Chapter 2 details the importance of US imaging in fetal growth monitoring, and difficulties associated with standardised image plane acquisition and the identification of key anatomical structures. This is followed by a review of machine learning methods for automated fetal US analysis, their current limitations, and the need for high level constraints to reduce misclassification of artefacts and increase classification, segmentation and localisation accuracies.

Chapter 3 describes the potential for insights obtained through eye tracking experiments to improve automated US analysis methods. It begins with a review of the uses of eye tracking in natural image analysis to discover which factors guide visual search in human observers, and how these insights have been incorporated into automated image analysis frameworks. Existing applications of eye tracking within medical image analysis are discussed, concluding with how eye tracking inspired high level constraints and interest point operators could be applied to the challenge of automated US image analysis.

Chapter 4 describes the various US image, volume and video datasets employed throughout this thesis.

Chapters 5-7 describe the original contributions of this thesis. Specifically:

Chapter 5 presents an eye tracking inspired framework for automated anatomical landmark localisation in 2-D fetal abdominal US images. Eye tracking experiments are described to determine which factors guide visual search in observers analysing fetal abdominal US images. Based on these findings, a pictorial structures model is implemented for fetal stomach bubble and umbilical vein detection, producing higher localisation accuracies than existing methods[7].

Chapter 6 presents an eye tracking inspired framework for automated standardised abdominal plane selection in 3-D fetal abdominal US volumes. The findings of further eye tracking experiments are presented, establishing which factors guide visual search in observers scrolling through abdominal US volumes. The previously derived pictorial structures model is then extended by incorporating 3-D optical flow and length based constraints, producing higher standardised abdominal plane selection accuracies than existing methods[7].

Chapter 7 presents an eye tracking inspired spatio-temporal interest point operator, and a framework for automated 2-D+t US video clip classification. The chapter describes the design of a spatio-temporal interest point operator learned directly from the eye movements of observers viewing 2-D+t US video clips showing different parts of the fetal anatomy. It is then demonstrated that, when incorporated into a BoVW framework, this eye tracking inspired operator results in higher US video classification accuracies than equivalent frameworks using the Harris[13] and Periodic[14] spatio-temporal operators.

Chapter 8 summarises the main contributions of this thesis and discusses potential areas for future work.

1.4 Peer Reviewed Publications

The peer-reviewed publications resulting directly from the chapters presented in this thesis are, at the time of submission, as follows.

Chapter 5

M. Ahmed, J.A. Noble, 'Eye Tracking to Boost Recognition of Anatomical Features in Fetal Ultrasound' *Proceedings of the 19th IEEE Medical Imaging Understanding and Analysis Conference - MIUA 2015*, 15-17 July 2015, University of Lincoln, UK.

M. Ahmed, C. Knight, S. Rueda, A.T. Papageorghiou, J.A. Noble, 'Eye Tracking as a Tool to Assess the Behaviour of Sonographers when Quality-Scoring Fetal Ultrasound Images' *Proceedings of the 24th World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2014*, 14-17 September 2014, Barcelona, Spain.

M. Ahmed, C. Knight, J.A. Noble, 'Which Local Image Features Attract the Human Gaze in Fetal Ultrasound Images?' *Proceedings of the 5th Medical Engineering Centres Annual Meeting- MEC and Bioengineering 2014*, 10-11 September 2014, Imperial College London, UK.

Chapter 6

M. Ahmed, J.A. Noble, 'An Eye Tracking Inspired Method for Standardised Plane Extraction from Fetal Ultrasound Volumes' *Proceedings of the 19th International Symposium on Biomedical Imaging - ISBI 2016*, 13-16 April 2016, Prague, Czech Republic.

Chapter 7

M. Ahmed, J. A. Noble 'Fetal Ultrasound Image Classification using a Bag-of-Words Model Trained on Sonographers' Eye Movements', *Proceedings of the 20th Medical Imaging Understanding and Analysis Conference - MIUA 2016*, 6-8 July 2016, University of Loughborough, UK.

An original idea. That can't be too hard. The library must be full of them.

— Stephen Fry, British Actor and Author (1957-Present)

2

Advances in Ultrasound Image Analysis

Contents

2.1 Introduction

Fetal biometric measurements obtained from 2-D B-Mode US scans (Figure 2.1) are crucial for fetal growth monitoring[2]. For instance an abdominal circumference (AC) below the 10^{th} centile for a given gestational age (GA) is the most sensitive biometric marker for the diagnosis of intrauterine growth restriction (IUGR), a condition whereby a fetus does not reach a healthy size for its GA, and a contributing

factor in approximately 3.5 million neonatal deaths per year[15], the majority of which occur in the developing world.

In order for serial biometric measurements to be comparable to standardised growth curves (Figure 2.2) measurements must be taken from standardised image planes as defined, for example, by clinical bodies including the International Society for Ultrasound in Obstetrics and Gynaecology (ISUOG), British Medical Ultrasound Society (BMUS), and Royal College of Obstetricians and Gynaecologists (RCOG)[4,16,17]. While each clinical body varies in the precise details of standardised plane definition, there are similarities between the guidelines, for example they generally agree that a standardised abdominal plane must show the stomach bubble, umbilical vein and a circular abdominal wall occupying more than 50% of the image[1,16].

However, standardised plane acquisition can be challenging due to the deterioration of image contrast with GA, the variable orientation and shape of key anatomical landmarks, and variations in sonographers' skill level and scanning techniques[5,18–20].

Automated methods to identify anatomical landmarks and detect standardised abdominal planes would help increase clinical workflow speeds and potentially allow less experienced sonographers to perform fetal biometry. Early work aimed at automation has employed both semi and fully-automated machine learning based approaches including local gradient driven active contours[21], random forests[22], and boosting[6]. The latter methods were trained on local intensity, phase and Haar-like features to localise the abdominal wall, stomach bubble and umbilical vein in 2-D US images[23,24]. A limitation of these 'bottom-up' approaches lies in their lack of high-level constraints, such as the geometric relationship between anatomical structures, and the fact that hand-crafted image features may not necessarily identify anatomically salient regions[25–29]. This can lead to artefacts and shadows being misclassified as anatomical landmarks despite lying in anatomically implausible image regions, and misclassifications between the stomach bubble and umbilical vein which frequently present with a similar size and shape[23,24].

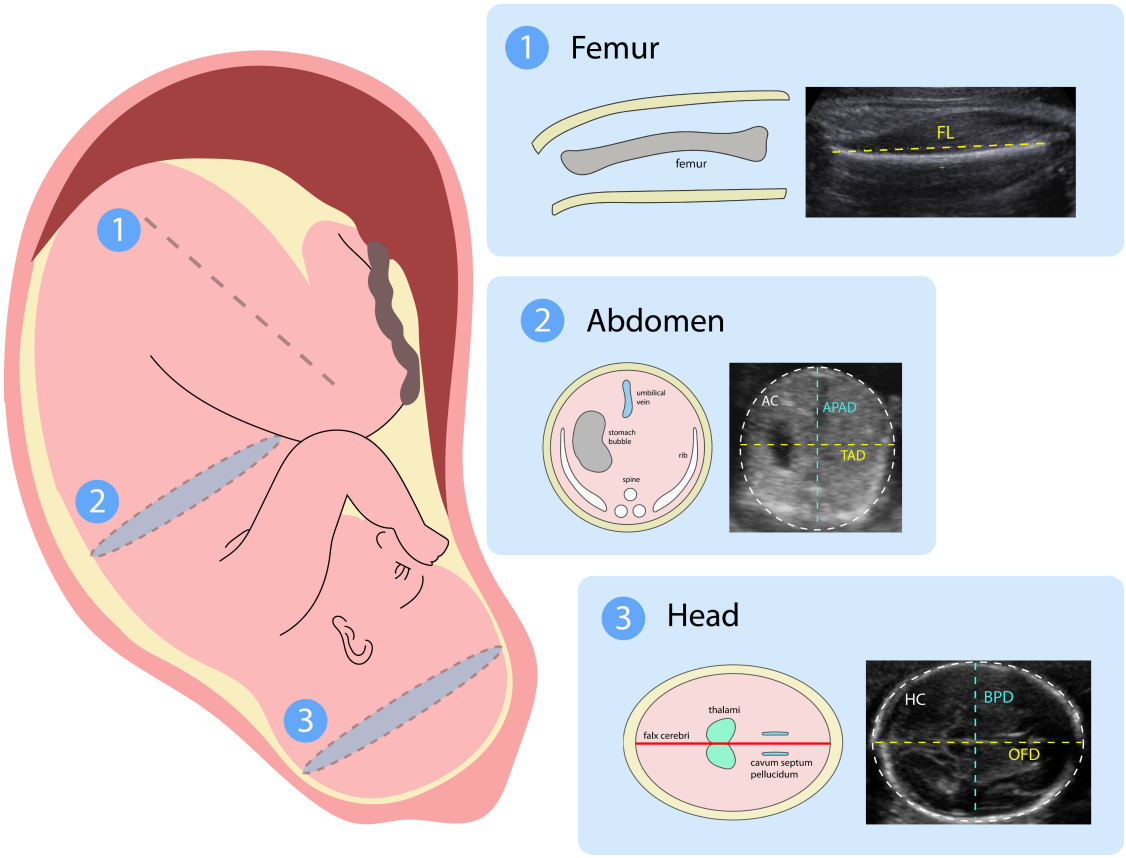


Figure 2.1: Standard fetal biometry, showing the three standard US planes and associated biometric measurements, namely femur length (FL), abdominal circumference (AC), and head circumference (HC)

.

2.2 Fetal Ultrasonography

2-D B-Mode US is the most widely used imaging modality for fetal growth monitoring and pregnancy risk assessment due to its safety, portability, low cost and the large existing body of guidelines on best clinical practice[16]. Within the United Kingdom, pregnant women attend two routine antenatal US scans[30]. A 'dating scan' is performed at $12-14$ weeks GA to establish the GA of the fetus; this is crucial for tracking fetal growth during subsequent scans. The crown rump length (CRL) is used in conjunction with standardised growth curves[31] to accurately date the pregnancy. This is based on the reasonable assumption that the CRL of the fetus is consistent with its GA, as there is very little variation in size between fetuses of the same GA during the first trimester of pregnancy. An 'anomaly scan' is performed at $18-20$

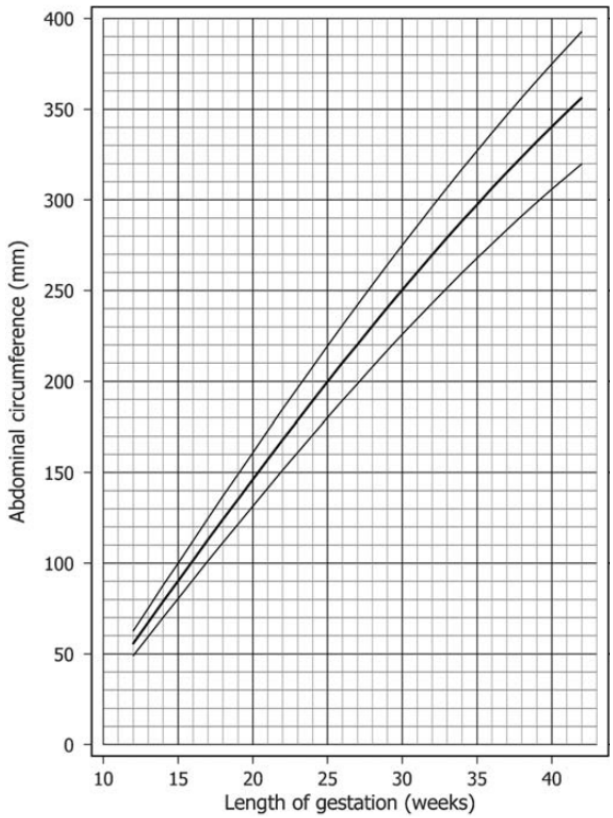


Figure 2.2: Standardised growth curve, as recommended by the British Medical Ultrasound Society, for abdominal circumference measurement, showing the 50^{th} centile (midline), 90^{th} centile (upper bound) and 10^{th} centile (lower bound) abdominal circumferences at different GAs. Image adapted from Loughna et al. [17]

.

weeks' GA to screen for major structural abnormalities in the fetus such as spina bifida and heart defects [17,32] and to detect non-viable or multiple pregnancies [33].

2.2.1 Standard Fetal Biometry

During both antenatal scans, standard biometric measurements are taken from standardised US imaging planes and compared against standardised growth curves to determine whether the size of the fetus falls within the 10^{th} and 90^{th} centiles for its GA. Key measurements obtained from the fetal head, abdomen and femur are shown in Figure 2.3 and Table 2.1.

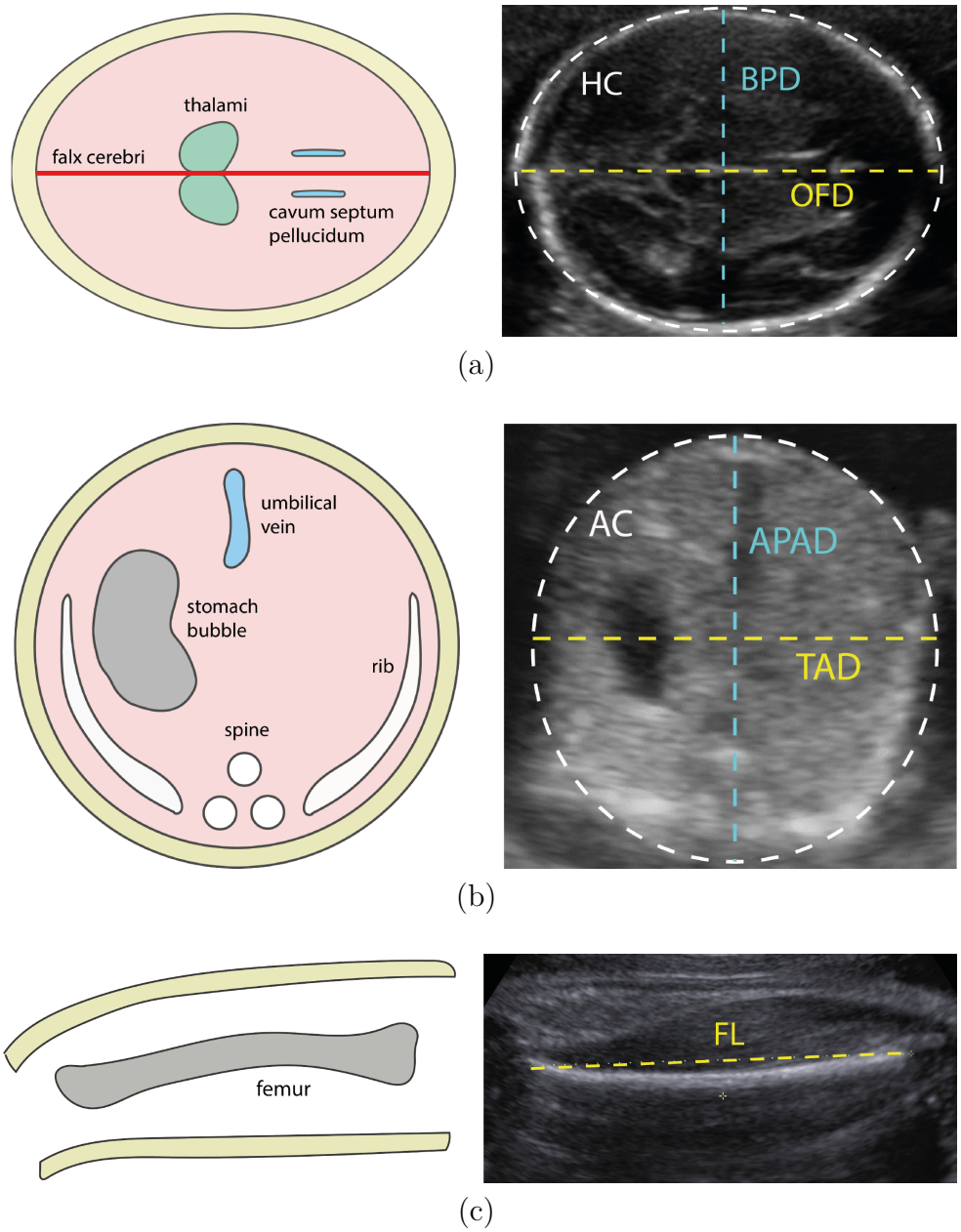


Figure 2.3: Key fetal biometric measurements obtained from standardised views of (a) The fetal head, specifically the biparietal diameter (BPD), occipito-frontal diameter (OPD), and the head circumference (HC) (b) The fetal abdomen, specifically the anterio-posterior abdominal diameter (APAD), the transverse abdominal diameter (TAD), and the abdominal circumference (AC) (c) The fetal femur, specifically femur length (FL).

Head	Abdomen	Femur
Occipito-frontal diameter (OFD), the distance between the outer borders of occipital and frontal bones along the longest axis of the skull	Anterio-posterior abdominal diameter (APAD), the distance between the outer borders of the anterior and posterior abdominal walls	Femur length (FL), the distance between the two endpoints of the fetal femur
Biparietal diameter (BPD), the distance between the outer borders of the parietal bones along the shortest axis of the skull	Transverse abdominal diameter (TAD), the distance between the outer borders of the abdominal wall perpendicular to the APAD	
Head circumference (HC), defined as $2\pi\sqrt{\frac{OFD^2+BPD^2}{2}}$	Abdominal circumference (AC), defined as $2\pi\sqrt{\frac{TAD^2+APAD^2}{2}}$	

Table 2.1: A summary of key fetal biometric measurements obtained from standardised views of the fetal head, abdomen and femur [1].

The use of standardised planes ensures that biometric measurements are comparable, both to measurements of the same fetus over a period of time and to growth curves. These planes, as defined by ISUOG, are shown in Figures 2.1 and 2.3, with Table 2.2 showing which anatomical features should be visible for standardised plane acquisition according to ISUOG guidelines [1]. Scoring schemes for standardised plane acquisition have also been proposed by ISUOG [1], where images are scored according to the number of criteria shown in Table 2.2 that are met, and one point is awarded for each criterion.

In growth restricted fetuses, glycogen stores in the fetal liver are depleted, leading to reduced liver size and hence a reduced AC measurement. AC therefore holds the greatest clinical significance for the diagnosis of IUGR [2,3,34]. An AC measurement below the 10^{th} centile or a HC/AC ratio above the 95^{th} centile will detect 62% and 82% of IUGR fetuses respectively, and an AC or HC/AC ratio within the normal range will correctly rule out 94% and 91% of non-IUGR fetuses respectively.

Head	Abdomen	Femur
Thalami visible	Stomach bubble visible	Both ends of the femur clearly visible
Cavum septi pellucidi visible	Umbilical vein visible	Femur is at an angle less than 45° to the horizontal
Head occupies more than 50% of image area	Kidneys not visible	Femur occupies more than 50% of image area
	Abdomen occupies more than 50% of image area	

Table 2.2: ISUOG guidelines for the acquisition of standardised planes showing the fetal head, abdomen and femur for biometry [1].

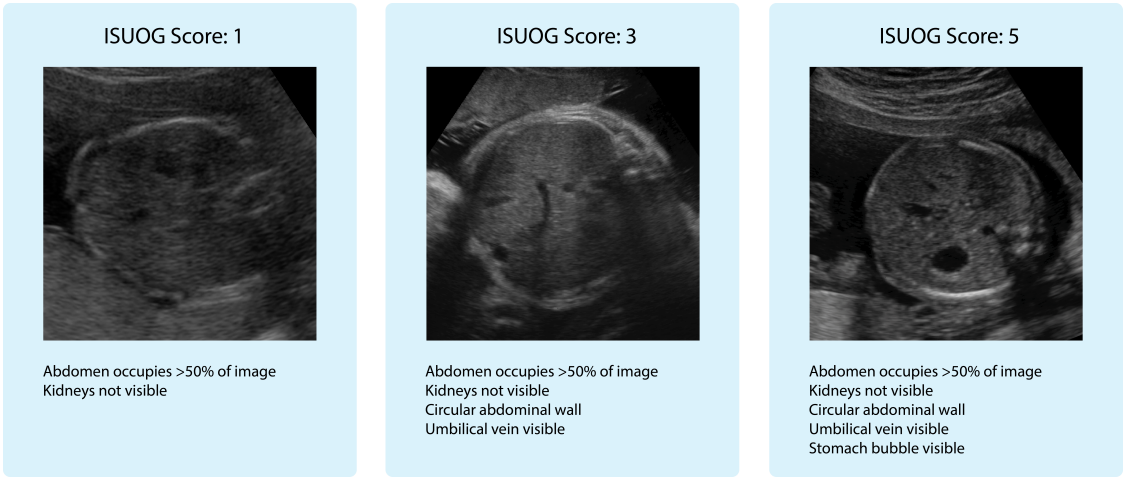


Figure 2.4: An illustration of US images of varying quality, as ranked by the ISUOG scoring method for standardised abdominal planes [1].

2.2.2 Challenges in Fetal Ultrasonography

The appearance of standardised planes in US images is highly variable due to the inherent physical properties of US imaging, sonographer skill, and human anatomy with the acquisition of standardised abdominal planes posing a particular challenge (Figure 2.5).

Acoustic scattering, which occurs at irregular tissue interfaces, leads to diffuse wave reflection and attenuation. Similarly speckle, caused by multiple scattering objects in close proximity, causes constructive and destructive wave superposition and increased image granularity. Despite the fact that soft tissue properties and

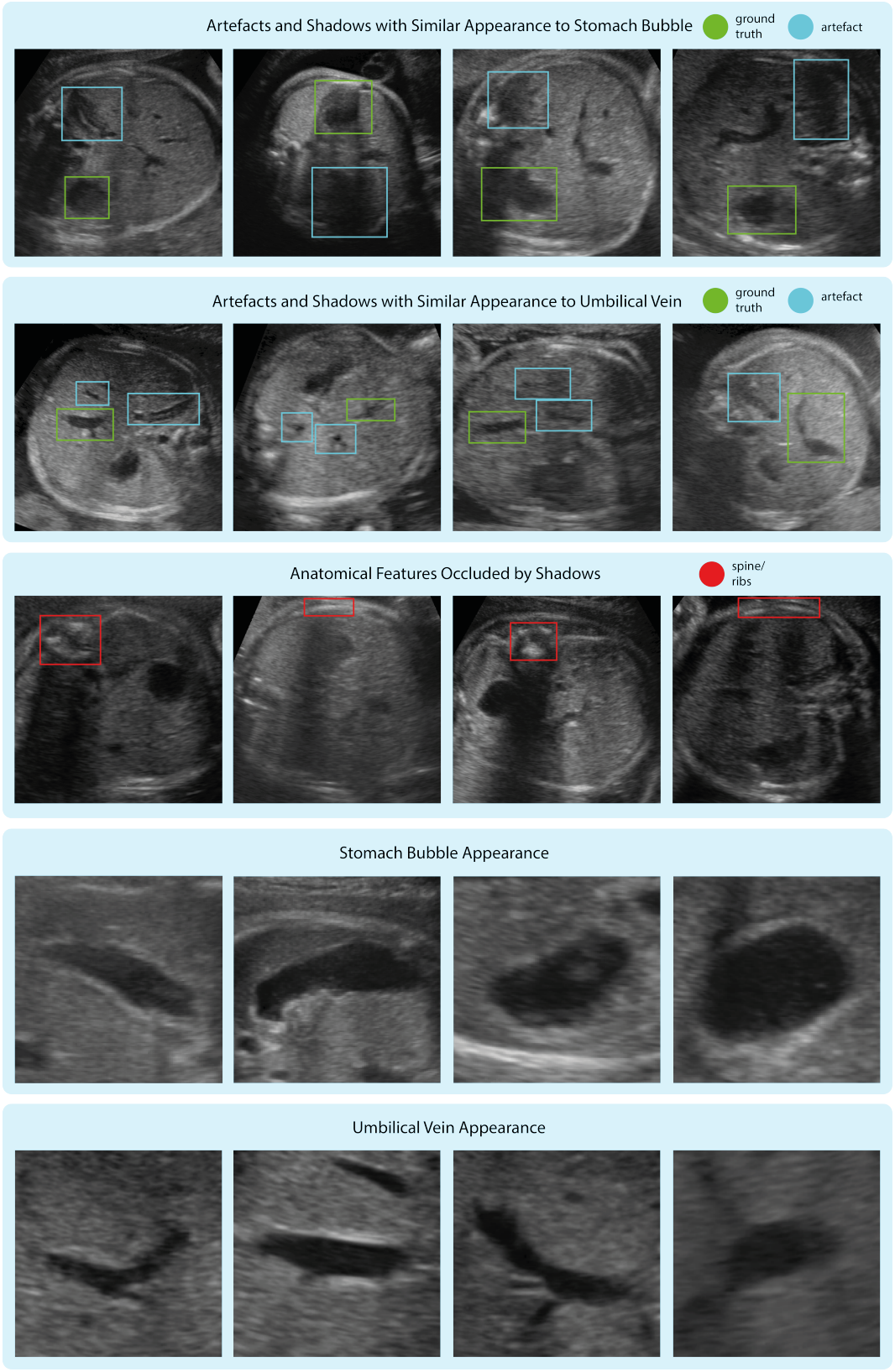


Figure 2.5: An illustration of the challenges in fetal US; namely artefacts, shadows, and the ambiguous shape and orientation of anatomical landmarks.

motion can be derived from image speckle[35,36], speckle tends to adversely affect image quality within the context of fetal biometry[5].

Acoustic shadows occlude soft tissue and key anatomical landmarks underlying the ribs and spine, and artefacts and shadows are frequently challenging to distinguish from the stomach bubble and umbilical vein. The highly deformable nature of the abdominal wall and anatomical landmarks can also be problematic. The position, orientation and shape of the stomach bubble and umbilical vein are highly variable and these two landmarks often present with a very similar appearance. Furthermore the abdominal wall is challenging to identify due to its lack of a bony perimeter and easily distorted shape. General image quality also deteriorates with GA, as increased maternal body mass index and adipose tissue leads to greater signal attenuation.

Additionally, manipulating an US probe to identify anatomical features and obtain the standardised plane is a highly skilled task. This complexity arises from the large number of US access windows and possible viewing planes[37], coupled with the need to optimise instrumentation parameters such as gain. Sonographers' scanning techniques vary greatly on an inter- and intra-operator basis[5,19] with some reported inter-operator agreements as low as 37%[20].

2.3 Automated Ultrasound Analysis Methods

Given the challenges associated with fetal US imaging and the limitations of US hardware, advances in automated US image analysis (typically driven by machine learning methods) are of particular importance. Existing work, summarised in Table 2.3, falls broadly into four categories; the acquisition of biometric measurements via semi- or fully-automated segmentation of 2-D US images, the localization of key anatomical landmarks for standardised plane acquisition in 2-D US images and 3-D US volumes, the classification of US images depicting different portions of fetal anatomy in 2-D US images, 3-D US volumes and 2-D+t US videos, and the discovery of new biomarkers for characterising healthy fetal growth or diagnosing fetal conditions.

Authors	Year	Fetal Anatomy	Method	Features	Validation

Chalana, Pathak et al. [38,39]	1996	Head	Active contours	Local gradient	35 2-D images
Jardim et al. [40,41]	2003	Head, Femur	Active contours	Probabilistic image model	None
Carneiro et al. [42]	2008	Head, Abdomen, Body, Humerus, Femur	AdaBoost	Haar-like features	3466 2-D images
Yu et al. [43]	2008	Abdomen	Active contours	Fuzzy C-Means clustering, local gradient, Iterative Randomised Hough Transform	150 2-D images
Yu et al. [44]	2008	Femur, Head	Active contours	Fuzzy C-Means clustering, local gradient, Iterative Randomised Hough Transform	150 2-D images
Nithya et al. [45]	2009	Abdomen	Active contours	Local gradient, Iterative Randomised Hough Transform	None
Yaqub et al. [22,46]	2010	Head, Abdomen, Body, Humerus, Femur, Face	Guided Random Forests	Novel scale and rotation invariant features	30000 2-D images
Rahmatullah et al. [47]	2012	Abdomen	AdaBoost	Intensity, gradient, Haar-like features, feature asymmetry	80 3-D volumes
Rueda et al. [48]	2012	Arm	Fuzzy connectedness	Phase, feature asymmetry, gradient	20 2-D images
Rackham et al. [21]	2013	Arm	Livewire	Local gradient, feature asymmetry, shape constraints	48 2-D images
Foi et al. [49]	2014	Head	Parametric model	Skull shape, intensity	90 2-D images
Maraci et al. [50]	2014	US video frame classification	Bag of visual words	SIFT	60 2D+t US videos
Maraci et al. [51]	2014	US video sequences of interest	Linear dynamical system model	SIFT	70 2D+t US videos
Namburete et al. [52]	2015	Head	Regression forests	3-D Haar-like features, morphological features	187 3-D volumes
Bridge et al. [53]	2015	Heart	Random forests	Haar-like features and gradient features	91 2-D+t US videos
Ryou et al. [54]	2016	Fetal body, head & abdomen	Random forests and convolutional neural network	-	-

Yaqub et al.[55]	2016	Head	Random forests	Intensity features and geometric distance based constraints	161 3-D US volumes showing the fetal brain
Gao et al.[56]	2016	-	Convolutional neural network	Transfer learning	-
Maraci et al.[57]	2017	Head, abdomen, heart	Linear dynamical system model	SIFT	323 2D+t US videos
Huang et al.[58]	2017	Heart	Recurrent convolutional neural network	-	91 2D+t US videos
Baumgartner et al.[59,60]	2017	Multiple	Convolutional neural network	-	-

Table 2.3: A summary of published semi-automated and fully automated US analysis methods.

2.3.1 Segmentation

Many semi-automated US segmentation methods employ deformable parametric curves (or active contours) defined by an initial curve which moves iteratively according to its internal energy (which constrains the shape and smoothness of the curve, specifying its elasticity and rigidity) and potential energy (computed from local image features around the curve's boundaries, attracting the curve to step changes in image gradient). The final segmentation is found by minimising an energy function comprised of these internal and potential energy terms[61].

Chalana et al.[38] and Pathak et al.[39] presented active contour based methods to segment the fetal head, using local image gradients to iteratively fit a contour around the HC, starting from a user specified initialisation point. This method achieved high HC segmentation accuracies on a testing set of 35 images, with a mean difference of 2.9% between computed and ground truth HC measurements and a mean run-time of $32s$ per image. However, this solely gradient based method would not be adaptable to AC segmentation due to the fetal abdomen's poorly defined boundaries.

Rackham et al.[21] presented a method for segmenting fetal arm adipose tissue in 2-D US images based on the Live Wire framework, which uses Dikstra's shortest path algorithm to compute the lowest cost contour between a user specified initialisation point and the user's subsequent manual segmentation points. Here local gradient, feature asymmetry and shape constraints were used to inform the cost function. When tested on a set of 48 2-D US arm images the algorithm showed improved repeatability and speed compared to solely intensity based Live Wire frameworks. Whilst accurately segmenting the fetal arm and head, gradient driven active contours would again not be adaptable to AC segmentation due to the fetal abdomen's poorly defined boundaries.

Nithya et al.[45] developed an adapted method for AC measurement using gradient based active contours, with an initial AC contour estimated using the Iterative Randomised Hough Transform, but quantitative validation of this method was limited. This approach was further extended by Yu et al.[43], using fuzzy C-Means clustering to detect strong edges before applying the Iterative Randomised Hough Transform and subsequently gradient based active contours, achieving a correlation coefficient of 98.78% when comparing gold standard manual AC measurements against AC measurements computed by the algorithm. The authors applied this same methodology[44] to the segmentation of the BPD, HC, and FL, resulting in a decrease in estimated fetal weight (EFW) calculation errors from 6.71% using manual segmentation to 4.66% using the automated method.

Other contour based methods include those of Jardim et al.[40,41]. The authors derived parametric curves and probabilistic models to describe the shape of the fetal femur and head, and implemented deformable contours to segment the FL and HC. However as the authors did not provide any quantitative validation for this method or make comparisons to ground truth measurements, it is not possible to gauge its accuracy.

Further work on the segmentation of fetal arm adipose tissue was conducted by Rueda et al.[48]. The authors used local phase and feature asymmetry to inform a fuzzy connectedness segmentation algorithm[62], which assessed the heterogeneity

between pairs of adjacent pixels, followed by a shape-based segmentation completion step. When tested on 20 2-D US arm images, the algorithm produced a mean Dice overlap coefficient of $86.62 \pm 2.73\%$ with the manually segmented ground truth.

Foi et al.[49] constructed a parametric model, or template, to describe the 2-D fetal skull shape and pixel intensities. The authors then matched this template to observed images by minimising a novel cost function using the Nelder-Mead algorithm. When tested on 90 2-D cranial US images, this method produced skull segmentations with a mean Dice overlap coefficient of $97.73 \pm 0.89\%$. and a mean run-time of $5.43s$ per image.

These approaches are of limited robustness, as active contours may erroneously settle at local minima and a degree of human input is still required for the initialisation of semi-automated segmentation methods.

Most recently, a segmentation challenge led by Rueda et al.[63] evaluated a number of methods for the segmentation of the fetal femur and head on a testing set of 180 2-D US images. Proposed approaches included the method of Ciurte et al.[64] which represented each US image as a graph of image patches, and segmented the fetal skull by solving this graph as a minimum cuts problem, and the method of Stebbing et al.[65] which segmented the fetal skull using a boundary fragment model.

However, no methods for abdominal circumference segmentation were proposed as part of this challenge, partly due to the inherent difficulties associated with fetal abdominal anatomy discussed previously. Furthermore, segmentation methods do not tackle the fundamental problem of finding standardised biometric planes, within a set of 2-D US images or a 3-D US volume, from which to obtain measurements.

2.3.2 Localisation

Initial advances in the fully automated detection of standardised abdominal planes via the localisation of the stomach bubble and umbilical vein were made by Rahmatullah et al.[23,24]. This approach employed local intensity, gradient magnitude, phase and Haar-like features to train an AdaBoost framework achieving detection accuracies of 78.94% and 62.80% for the stomach bubble and umbilical

vein respectively in 2-D abdominal US images. The limitations of this local feature based approach included the misclassification of shadows and artefacts as anatomical landmarks, due to the lack of high level contextual information.

Later work by Rahmatullah et al.[47] improved on this method, using feature asymmetry as a global constraint to identify candidate locations for anatomical landmarks, achieving improved detection rates of 82.75% and 72.55% for the stomach bubble and umbilical vein respectively in 2-D abdominal US images. This approach was extended to the extraction of standardised abdominal planes from 3-D abdominal US volumes[66], employing the same AdaBoost framework described above to select 3-D US volume slices which maximised a normalised classifier score and achieving a mean standardised plane selection accuracy of 91.29%.

Ryou et al.[54] presented a hybrid method for the localisation of the fetus and the extraction of standardised planes from 3-D US volumes showing fetuses between $11-13$ weeks GA. Structured and classical random forests were employed to localise the whole fetus in sagittal images, and a convolutional neural network (CNN) classified axial views of the fetus as belonging to the fetal head or body.

Gao et al.[56] also harnessed CNNs in fetal US image analysis, investigating the performance of a CNN pre-trained on a large set of natural images in localising anatomical structures in fetal US images. The transfer-learning based CNN trained on the larger dataset identified anatomical structures with a greater mean accuracy (91.5%) than an equivalent CNN trained on a smaller set of US images (87.9%), demonstrating that features learned from natural images can be applicable to US image analysis.

Baumgartner et al.[59,60] have developed a CNN based framework for the real time identification of 13 standardised views of the fetus during freehand fetal US scanning. The same framework also achieved a frame classification accuracy of 90.09% and anatomical landmark localisation accuracies of 77.8%.

2.3.3 Classification

Carneiro et al.[42] trained probabilistic boosting trees on Haar-like features to classify US images as displaying the fetal head, abdomen, body, humerus, or femur, and automatically acquire BPD, HC, AC, and FL measurements. This method was subsequently adopted by Siemens[67], providing automated BPD, HC, AC, FL, and CRL measurements given a standardised biometric plane leading to a 75% reduction in the number of keystrokes performed by sonographers compared to manual biometric measurement acquisition. However this work did not address the underlying challenge of standardised abdominal plane acquisition, as not all images showing the fetal head, abdomen and femur will meet the criteria for a standardised abdominal plane.

Yaqub et al.[46] improved these classification accuracies for US images using the weighted voting of trees within random forests (RF), and subsequent[22] guided RFs were trained on novel, scale and rotation invariant features extracted from anatomical regions of interest. This method achieved a mean classification accuracy of 75% for US images showing the fetal head, heart, abdomen, face, and femur.

Further work by Yaqub et al.[55] employed random forests to localise diagnostic imaging planes from 3-D US volumes showing the fetal brain, allowing the tracking of image-based biomarkers throughout gestation. Voxels displaying higher feature asymmetry were weighted more strongly during the training process, and novel geometric features were proposed, encoding the perpendicular distance between voxels and diagnostic planes. The final framework extracted diagnostic imaging planes with higher levels of reproducibility than expert clinicians performing the same task.

Maraci et al.[51] proposed a method for classifying image frames in 2-D+t US videos, constructing a bag-of-visual-words (BoVW) model using SIFT descriptors to model the temporal evolution of anatomical structures of interest in US videos as a linear dynamical system and establishing a general framework for US video analysis. Maraci et al. extended this work[57] to identify the fetal skull, abdomen, and heart with a mean classification accuracy of 83.4% across a set of 323 2-D+t US videos.

Bridge et al.[53] trained random forests with intensity and gradient features to devise a model for the automated interpretation of US video frames showing the fetal heart. The resulting framework predicted heart position, orientation and cardiac phase with similar accuracies to expert observers when tested on a set of 91 short US video clips of the fetal heart.

Huang et al.[58] proposed an alternative, CNN based method for the automated analysis of US videos showing the fetal heart. The authors designed a recurrent CNN architecture to classify the viewing plane, location and orientation of the fetal heart on a frame by frame basis, achieving classification accuracies comparable to manual expert annotations.

2.3.4 Biomarker Discovery

Namburete et al.[52] developed an automated framework for GA prediction and fetal neurodevelopment using novel biomarkers based on the appearance of structures in the fetal brain at different developmental stages. 3-D US volumes of the fetal head were used to learn hand-crafted features for discriminating each stage of fetal development, which were in turn used to train a regression forest for GA prediction. When tested on 187 3-D US volumes of the fetal head this method achieved a mean predictive accuracy of ± 6.10 days, outperforming current clinical methods during the third trimester.

2.4 Machine Learning in Ultrasound Analysis

This thesis makes use of several of the local image features and machine learning methods described in the above literature and applies them to the tasks of automated anatomical landmark localisation in 2-D US images, standardised abdominal plane selection 3-D US volumes and the classification of 2-D+t US video clips. Selecting appropriate local image features and classifiers for US analysis applications is crucial, as the straightforward application of machine learning methods designed for natural images does not necessarily result in high classification, segmentation or localisation accuracies with US images.

2.4.1 Local Image Descriptors

As demonstrated in Section 2.3, selecting discriminative local image features can play a key role in the design of automated US analysis algorithms.

Intensity Based Descriptors

Raw intensity and texture features[68,69] are not highly discriminative descriptors for US images given the variable contrast, signal-to-noise ratio and speckle associated with this imaging modality. However intensity based descriptors have been successfully used by many authors, for instance Rahmatullah et al.[47], Carneiro et al.[42] and Huang et al.[70] for anatomical segmentation and localisation in US images. Haar-like features, which compute differences between summed pixel intensities in rectangular windows, are computationally efficient to calculate from integral images[71]. Maximally stable extremal region (MSER) detectors iteratively grow connected image regions based on pixel intensities, effectively segmenting blob-like structures within US images[72]. To further mitigate the difficulties associated with poor image contrast and speckle, intensity invariant filters such as local feature symmetry and signed feature symmetry, derived from the monogenic signal, have been used to pre-process US images before further feature extraction[48].

Gradient Based Descriptors

Gradient based features have been used extensively to drive active contour energy functions for US image segmentation[38,39,43–45]. In US images with poorly defined edges and contrast, raw intensity gradients are unlikely to be robust discriminators between anatomical objects of interest and artefacts. Histograms of gradients (HoG) describe the distribution of intensity gradient directions in localised image patches, and were initially employed by Dalal and Triggs et al.[73] in pedestrian detection prior to use by Maraci et al.[51] in a promising BoVW framework for US video frame classification.

Scale and Rotation Invariant Interest Point Detectors

The primary disadvantage of extracting descriptors from dense rather than sparse sampling grids is the lack of discrimination between anatomically significant regions and artefacts which may display the same local image features. Interest point detection, via scale and rotation invariant descriptors such as SIFT[74] and SURF[75], partially mitigates this by detecting interest points at different levels of Gaussian scale-space based on local gradients. SIFT and SURF are commonly used within action recognition frameworks[9] and have also been employed, for instance, by Maraci et al.[50,51,76] due to their robustness to the variable orientations of anatomical landmarks[77] and their ability to discern blob-like structures.

Despite the advantages of these combined interest point detection and feature extraction methods, image regions which are identified as salient by conventional hand-crafted features and interest point operators may not correspond to anatomically meaningful regions. Therefore there remains a discrepancy between current feature extraction methods and image regions that are most meaningful for US image analysis. This notion of hand-crafted features failing to identify salient image regions is not unique to US image analysis, and is well acknowledged in the fields of natural and medical image analysis more broadly. This has partly driven the advancement of unsupervised feature selection via deep convolutional neutral networks[25–29].

2.4.2 Classifiers and Frameworks

Factors affecting the choice of machine learning classifier for image analysis include the dimensionality of the images (2-D, 3-D, or 2-D+t), the number of classes being distinguished, and available quantities of training data.

Support Vector Machines

Support vector machines (SVMs) are a classic form of supervised binary classifier, and have been used extensively as part of BoVW frameworks for action recognition[9] and more recently for US video frame classification[76].

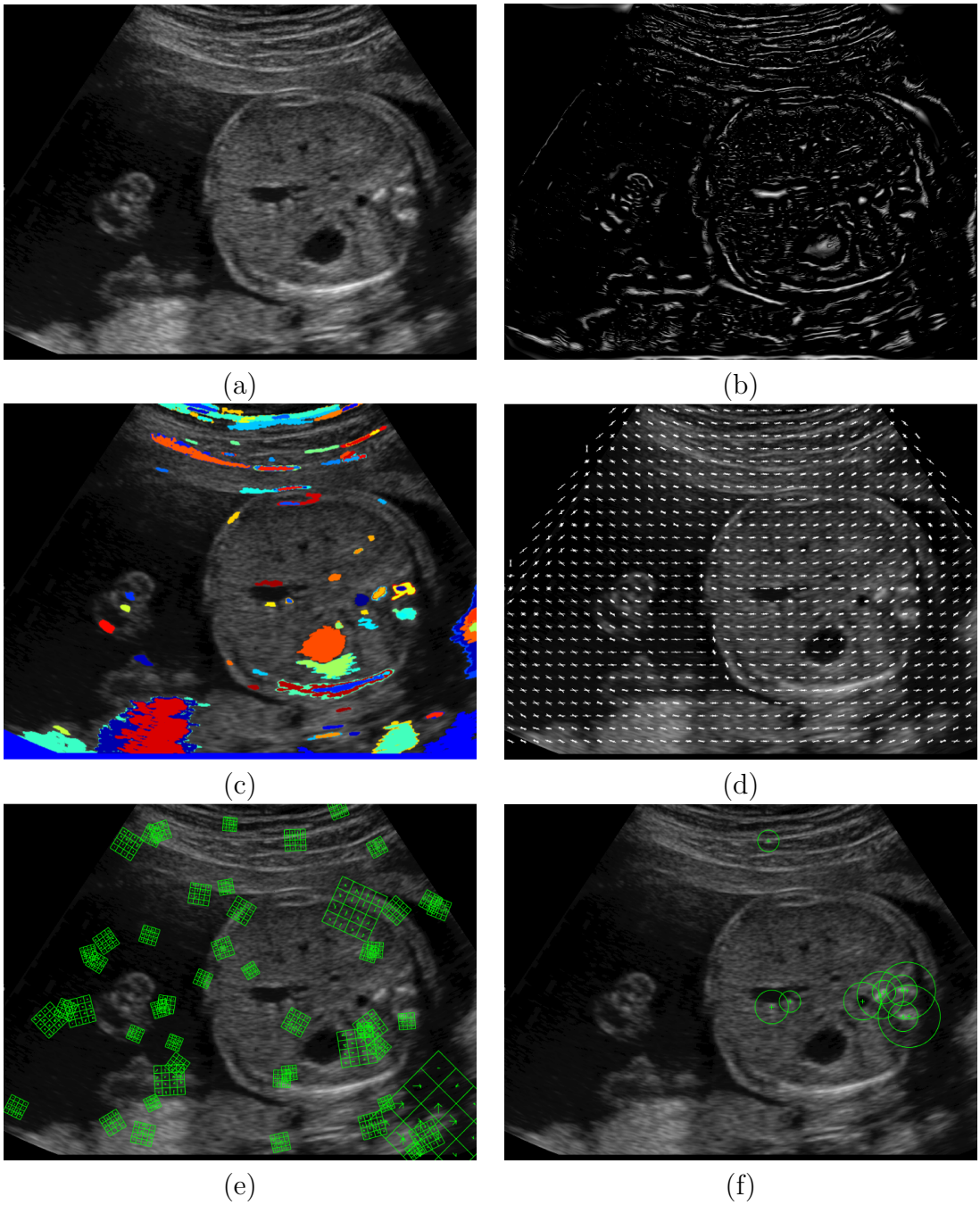


Figure 2.6: Frequently used local image features in US analysis, namely (a) An unprocessed grayscale image of the fetal abdomen (b) An unsigned feature symmetry map derived from the monogenic signal[78] (c) Maximally stable extremal regions[79](d) Histograms of gradients[73] (e) SIFT interest points at varying scales and orientations[74] (f) SURF interest points at varying scales and orientations[75]

.

Linear SVMs are trained by finding a separating hyperplane between two classes of training data in feature space. The optimal solution is one that gives the greatest margins between the separating hyperplane and its closest training data points, known as support vectors. Training sets that are not highly separable by linear hyperplanes may be mapped into higher dimensional feature space through kernel functions, producing greater classification accuracies.

In US analysis, image classification problems often involve several classes rather than just two. Multi-class SVMs can be constructed via a combination of binary SVMs which either distinguish between every possible pair of classes, or one class against all other classes. In Chapter 7, a multi-class SVM is employed within a BoVW framework for US video frame classification.

Adaptive Boosting

Adaptive boosting (AdaBoost) is an ensemble method initially employed by Viola and Jones et al.[80] for object detection, widely employed in computer vision. Within automated US analysis research it has been used for the localisation of anatomical features in abdominal US images by Rahmatullah et al.[23].

Here a number of weak classifiers are combined to form a strong classifier which performs significantly better than any individual weak classifier. The voting weights of each weak classifier are learned via an iterative process, where training data that is misclassified by one weak classifier is given a greater weight when training the next weak classifier (Figure 2.7).

In the case where each weak classifier consists of a single decision stump, or a single local image descriptor at a fixed scale with a learned classification threshold, the process of learning classifier voting weights also serves as a feature selection step. This makes AdaBoost a particularly efficient machine learning method, coupled with the fact that each weak classifier need only perform slightly better than random chance in order to achieve a strong overall classifier.

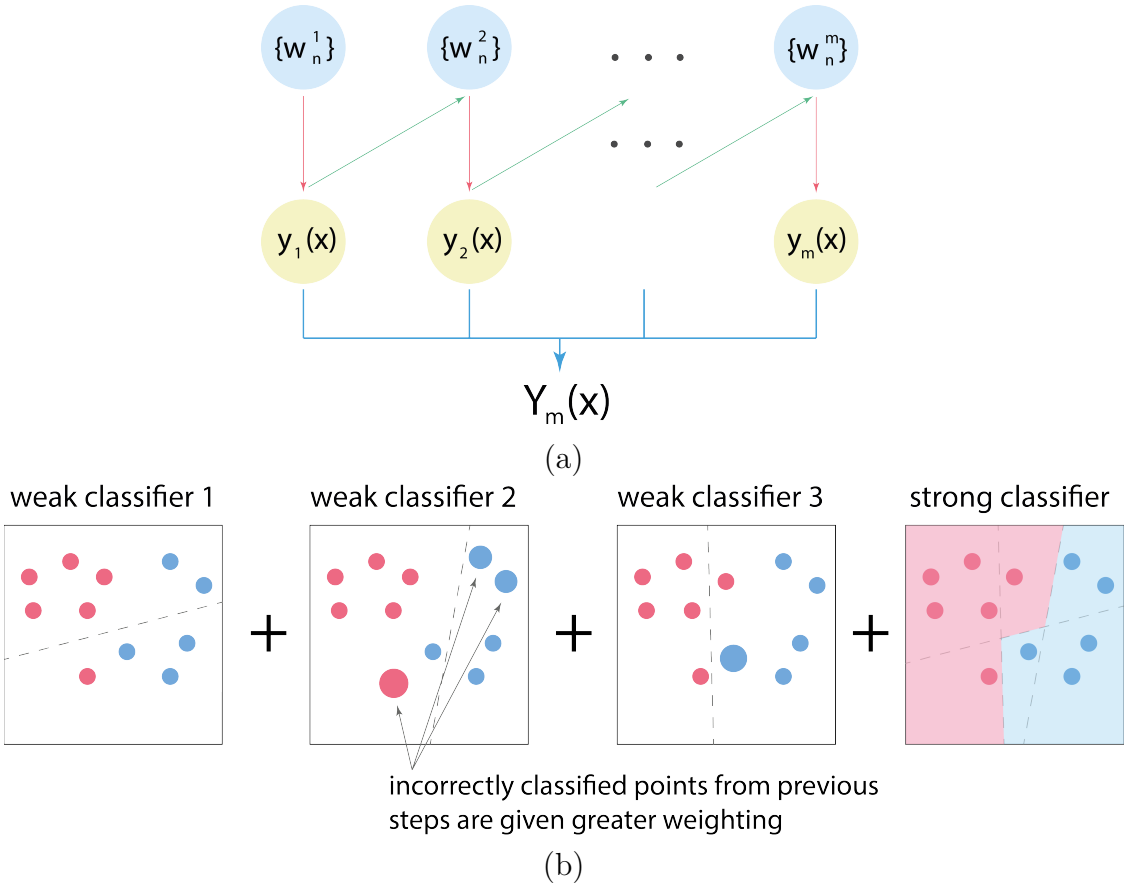


Figure 2.7: An illustration of the training process for AdaBoost, where (a) Each one of m weak classifiers (yellow $y_1(x), y_2(x), ..., y_m(x)$) is trained (red arrows) on a weighted variation of the training set (blue where w_n^1 is a vector giving initial weights for all n training datapoints) with training errors (green arrows) used to inform the subsequent set of weights to train the next weak classifier (b) The final strong classifier is a linear combination of all weak classifiers and their associated decision boundaries.

This thesis employs an AdaBoost framework comprised of decision stumps, followed by a pictorial structures model to localise the stomach bubble, umbilical vein and spine in 2-D and 3-D abdominal US images. This is discussed further in Chapter 5.

Bag of Visual Words

A bag of visual words (BoVW) model is a classification framework which often incorporates multi-class SVMs, and within the context of US image analysis has been used by Maraci et al.[51] and Kwitt et al.[81] for classifying sequences of interest in 2-D+t US video sequences. Local descriptors, typically HoG, SIFT or SURF, are

extracted densely or via an interest point operator from a set of testing images, and clustering is performed in feature space to produce a vocabulary of visual words.

Each image is then representable as a histogram of visual words, and a multi-class SVM is trained to classify images using the frequencies of visual words rather than underlying feature vectors themselves. This approach has been extended to US video sequences through the use of spatio-temporal features including 3-D HoG operators. As discussed, interest point operators and densely sampled features may not necessarily correspond to anatomically significant image regions or prove to be highly discriminative.

This thesis develops an interest point operator designed to mimic human visual attention on images, thus ensuring the image regions used to train a BoVW model for US video classification are anatomically significant. This is discussed further in Chapter 7.

2.5 Conclusions

Whilst US is the most widely used imaging modality for fetal growth monitoring, standardised abdominal image planes remain challenging to acquire well due to variations in sonographer skill, the inherent physical properties of US imaging, and changes in fetal anatomy and soft tissue. Many machine learning based methods for US analysis, anatomical landmark localisation and standardised abdominal plane recognition from 2-D, 3-D and 2-D+t US images are limited by their lack of high level constraints, and their dense approach to feature extraction[42,47,50,52,53]. Although the use of CNNs for automated US analysis has shown promising early results[53,54,56,58], these architectures do not necessarily make efficient use of training data, and the challenge of incorporating high-level contextual knowledge of human image interpretation strategies into machine learning frameworks remains.

Your eyes can deceive you. Don't trust them.

— Obi Wan Kenobi, Star Wars: A New Hope (1977)

3 Eye Tracking in Image Analysis

Contents

3.1 Introduction

Despite the advances in semi and fully automated US image analysis methods outlined in Chapter 2, there remains a significant discrepancy between the way in which human observers analyse US images and locate anatomical features, and the way in which the automated methods discussed previously perform the same task. The limitations of conventional object localisation frameworks lie in their purely 'bottom-up' approach and reliance on low-level local image descriptors[9]. In particular, densely sampled local features or interest points will not always correspond to anatomically meaningful image regions despite highlighting points which are statistically salient. When used to train machine learning frameworks, these features

may not result in highly discriminative classifiers. Classifiers themselves are limited by a lack of high level constraints, with the AdaBoost framework presented by Rahmatullah et al. showing a significant degree of confusion between stomach bubble and umbilical vein localisations, and misclassifications of artefacts and shadows as anatomical features despite lying in anatomically implausible positions[24].

In contrast, the human visual system employs a 'top-down' approach to visual search, using prior knowledge (such as the expected relative positions of objects) and visual cues from a given image to search for targets. Torralba et al.[82], for example, found that when searching for pedestrians in a street scene, observers use prior knowledge to deduce that targets are more likely to be found on the pavement against a grey background, than in the sky against a blue background, emphasising the importance of context and the relative positions of objects in images.

Eye tracking has been an area of interest within the computer vision and cognitive science communities for some years[8,83–85]. More recently eye tracking has been harnessed to obtain priors, develop perception inspired interest point operators, and establish high level constraints within human action recognition, natural image processing[9,12,82], and radiology contexts[86,87].

However, prior to this thesis, there have been no eye tracking studies involving US images and sonographers, and no assessment of how knowledge gained via these studies can inform the design of automated US analysis frameworks.

3.2 Recording Eye Movements

Eye Tracking Hardware

Advances in eye tracking hardware, primarily driven by the gaming, computer graphics, user interface design, and simulation industries[88–94], have enabled the development of accurate and portable eye tracking devices. Eye trackers operate on the principle of pupil centre corneal reflection, and consist of four key stages: the illumination of the pupils and cornea with near infra-red light, the imaging of the eyes with infra-red cameras to identify corneal reflections at sampling rates between $30Hz$ and $120Hz$, and a final step to calculate the direction of gaze and

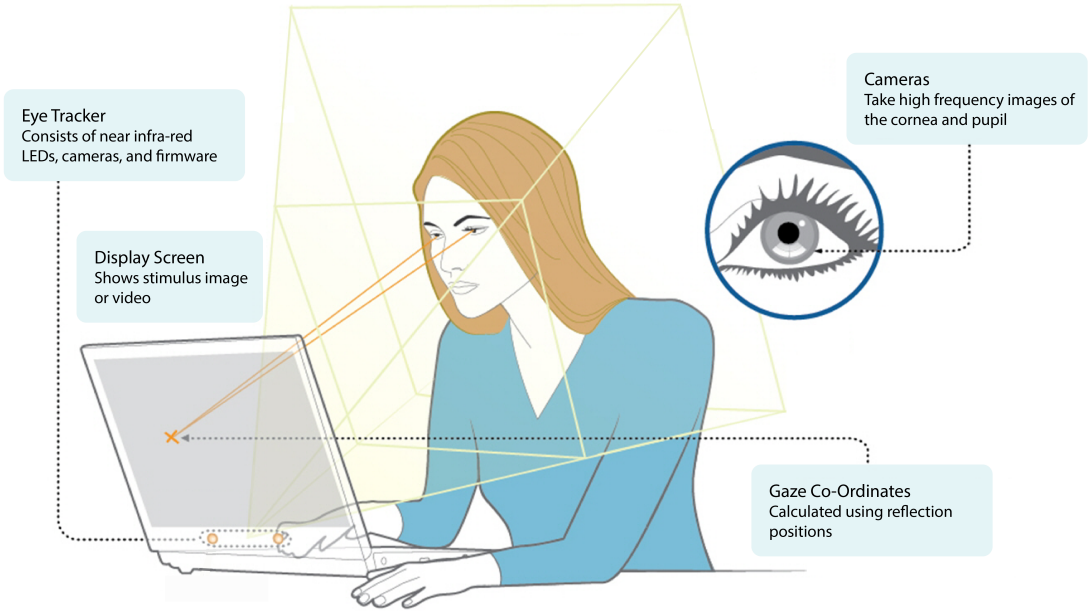


Figure 3.1: A schematic overview of the main components of a modern eye tracker, namely the illumination of the pupils and cornea with near infra-red light, the imaging of the eyes with infra-red cameras to identify corneal reflections, and a final step to calculate the direction of gaze and the gaze co-ordinates on a 2-D display screen. Figure adapted from Tobii[96].

.

the gaze co-ordinates on a 2-D display screen (Figure 3.1). Accuracies range from $0.4 - 1.5°$ visual angle (Figure 3.2), equivalent to $3.5 - 13.1mm$ when observers are positioned $0.5m$ from the eye tracker[95]. The experiments described in this thesis utilised an EyeTribe v1.0 (the Eye Tribe, Denmark) eye tracker, selected for its portability and open-source SDK.

Fixation Filtering

Raw gaze co-ordinates must be filtered into fixations, points on which the gaze lingers and which capture the observer's visual attention, and saccades, fast movements between fixations.

The selection of a filtering algorithm and associated thresholds is crucial. The widely used 'identification of velocity threshold' (I-VT) filter applies an angular velocity threshold[97], in degrees of visual angle per second, to raw eye movements (Figure 3.3). Points falling below the threshold are classified as fixations, and those

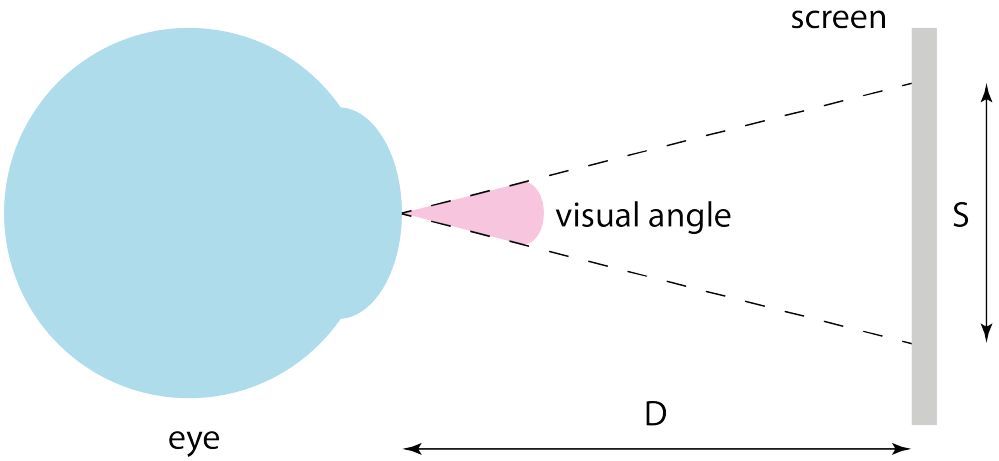


Figure 3.2: An illustration of the vertical visual angle, and the corresponding vertical portion of a 2-D screen subtended by an observer's gaze. Here S is the height of the area subtended by the visual angle, and D is the distance between the lens of the observer's eye and the screen surface

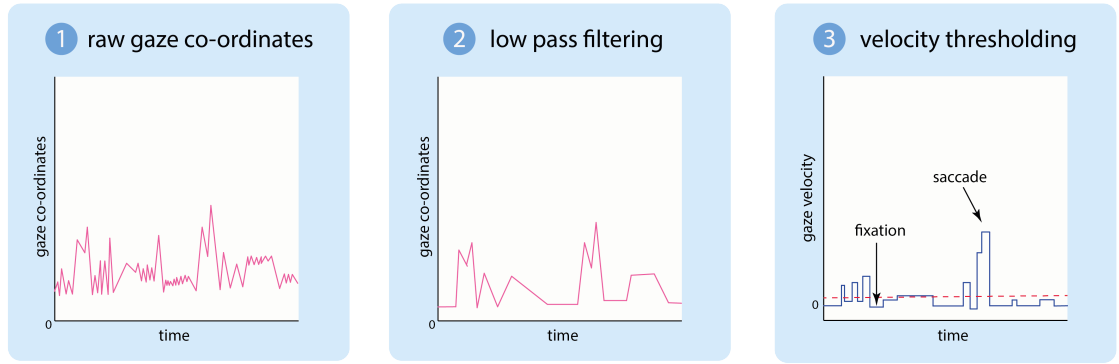
.

above as saccades. This method was used within Kundel and Nodine's work on the visual search strategies of mammographers and radiographers[98]. However, the robust calculation of angular gaze velocity relies on tracking or constraining the observer's distance from the display screen.

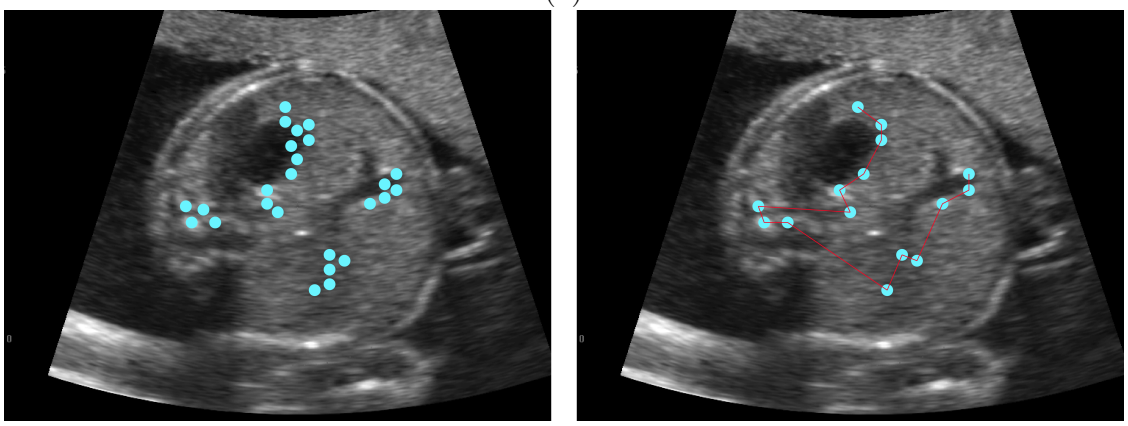
When this distance is not available the 'identification of dispersion threshold' (I-DT) filter is used. Here a duration threshold is applied to raw gaze co-ordinates, and the resulting points are clustered to form fixation points based on a dispersion threshold. In this thesis the I-VT filter is used, as observer-to-screen distance tracking is available via the chosen eye tracking hardware. This is discussed further in Chapter 5.

3.3 Modelling Visual Search

Observers searching for targets (such as tumours, fractures, or enlarged lymph nodes) in medical images have been found to exhibit varying visual search strategies, dependent on their level of expertise, the visual search task at hand, and the



(a)



(b)

Figure 3.3: An illustration of the I-VT fixation filter, with (a) A schematic overview of the steps involved in the I-VT filter, namely the low pass filtering of raw gaze co-ordinates to reduce noise, angular gaze velocity calculation, and velocity thresholding (b) An illustration of (left) Raw gaze co-ordinates (right) The same gaze data filtered into fixations (blue) and saccades (red).

stimulus type[99]. Various models of visual search have been proposed to characterise these different search strategies.

The importance of medical image perception, from a public health and expenditure perspective, is increasingly recognised[100]. Kundel and Nodine et al.[10,86,87,98,100,101] have focused on the visual search strategies of novice and experienced radiologists observing chest X-Rays. The authors[10,98,101] postulated the global-focal model of visual search for radiographers searching for abnormalities in medical images including chest radiographs and mammograms, consisting of three distinct stages of visual search:

1. The 'global impression' stage is an initial search using mainly peripheral vision, lasting for less than $200ms$.

2. The 'discovery search' phase, a detailed inspection of each target.

3. The 'reflective search' stage which involves the observer cross referencing against other targets and potential targets.

This model was further validated[86] in a study by Kundel et al., where ten radiologists were tasked with identifying abnormalities in 20 chest radiographs, with a viewing time of $200ms$ per image. The overall true positive rate of 70% supported the hypothesis that visual inspection of medical images begins with an initial 'global impression' phase whereby the observer establishes the overall content of the image and detects conspicuous abnormalities.

Kundel et al.[87] later built on the 'global-focal' model to propose the 'holistic' model of visual search, based on experiments involving nine mammographers searching for abnormalities in 40 mammograms. They found that more experienced mammographers undertook a more efficient 'global impression' or 'holistic' search phase, whereas less experienced observers relied more on the subsequent 'discovery' or 'search-to-find' phase to identify abnormalities.

In the first quantitative validation of Kundel and Nodine's[10,98,101] two-component search hypothesis, Leong et al.[102] tracked the eye movements of 25 consultant radiologists, consultant orthopaedic surgeons, orthopaedic registrars, and orthopaedic Senior House Officers searching for fractures in 33 hand radiographs. The authors employed a two-component Gaussian mixture model to model the distance between the fracture site and the each observer's gaze co-ordinates as a function of time. They found that expert observers employed more accurate and efficient search strategies than less experienced observers.

Mallett and Philips et. al.[103] quantified the differences in visual search between 27 expert and 38 novice radiologists tasked with identifying polyps in 23 3-D+t CT colonography videos. Experienced radiologists identified a higher proportion of polyps than novices, and the majority of false-negative interpretations of polyps

were attributable to decision and recognition errors, whereby a polyp was fixated but not reported by the observer.

Bertram et al.[104] investigated the effect of expertise on the visual search behaviour of seven radiologists, nine computed tomography (CT) radiographers and 22 psychology students when interpreting 9 3-D abdominal CT volumes. Radiologists required fewer fixations on salient regions in order to detect enlarged lymph nodes, performed shorter saccades, and detected enlarged lymph nodes with a greater accuracy than radiographers and students. Radiologists' fixations were of a longer duration than those of radiographers or psychology students, suggesting a more thorough search strategy.

Antonelli et al.[105] recorded the eye movements of five radiologists searching for abnormalities in 10 lung CT volumes, noting that many missed anomalies in medical images were subconsciously fixated upon by observers but not consciously registered. Local energy, entropy, and texture were computed around the radiologists' fixation points in real time. Salient regions in the images were identified, and used to provide live feedback to the radiologists as a means of guiding their search for abnormalities. There was no quantitative validation of whether this feedback method decreased the number of missed anomalies by visual inspection.

Despite the considerable work carried out on predicting visual saliency maps on natural scenes, and the visual search strategies of clinicians interpreting radiographs, prior to the work described in this thesis there has been no investigation into the generation of visual saliency maps for US images, and no eye tracking investigations using US images as stimuli.

3.3.1 Search Similarity

The use of metrics to analyse eye movements on an inter-observer basis is crucial for a standardised approach to image perception. Mathe and Sminchescu et al.[9] proposed the use of 'static consistency' and 'dynamic consistency' to quantify the similarity between the fixated locations and fixation sequences of a cohort of observers respectively.

Static Consistency

To compute the static consistency of a cohort of observers viewing one image, one observer's fixations are predicted using the summed fixations of all-but-one observers. The static consistency score is the accuracy, or area under the receiver-operating characteristic curve (AUC of ROC), of the summed fixation map in predicting an individual's fixations (Figure 3.4).

Dynamic Consistency

Dynamic consistency for a cohort of observers is computed by clustering fixations into discrete regions of interest (ROIs) and representing each observer's fixations as a string of sequential ROIs. Dynamic consistency is then the accuracy of the fixation sequences of all-but-one observers in predicting one observer's fixation sequence through training a Markov model as in Figure 3.5.

3.4 Predictive Visual Saliency Maps as Interest Point Operators

Determining which local and global image features attract the gaze, and producing predictive 'visual saliency' or attention maps for unseen images, is a crucial first step in harnessing eye tracking to build constraints and prior knowledge into image analysis frameworks. Predictive visual saliency maps have also been used as interest point operators on unseen images, as the first stage in a larger pipeline for natural image classification and action recognition[9].

Itti et al.[106] were amongst the first to produce a local feature driven model of visual saliency on natural images. Feature maps based on local image intensity, colour channels and oriented Gabor pyramids were computed at varying spatial scales. A weighted combination of these features provided the final saliency map, resulting in a fixation prediction accuracy of 75% across the MIT visual saliency benchmark dataset consisting of 300 natural images and ground truth fixations from 39 observers[107].

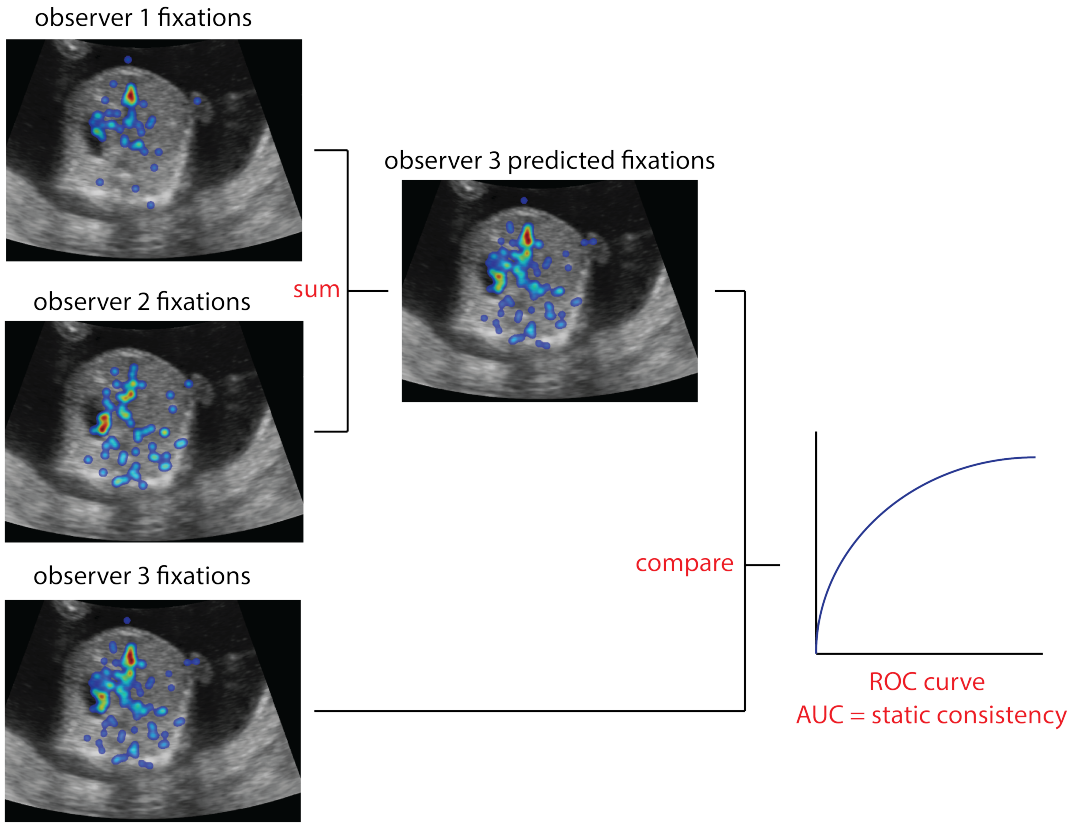


Figure 3.4: An illustration of static consistency calculation, showing the summing of the fixation maps of all-but-one observers to produce a predictive map of one observer's fixations on a single image.

.

Rajashekar et al.[108] investigated the effect of four local image descriptors on visual saliency namely luminance, contrast, and their bandpass filtered equivalents. The authors computed these feature vectors at varying spatial scales around fixation points and non-fixation points, and found that regions surrounding fixation points were characterised by higher bandpass contrast values than non-fixations.

Harel et al.[109] modelled local feature maps as fully connected graphs, with weighted edges representing the dissimilarity between neighbouring vertices in feature space. When modelled as Markov chains, the equilibrium distribution of these graphs provided a measure of saliency, resulting in a fixation prediction accuracy of 81% on the MIT dataset[107].

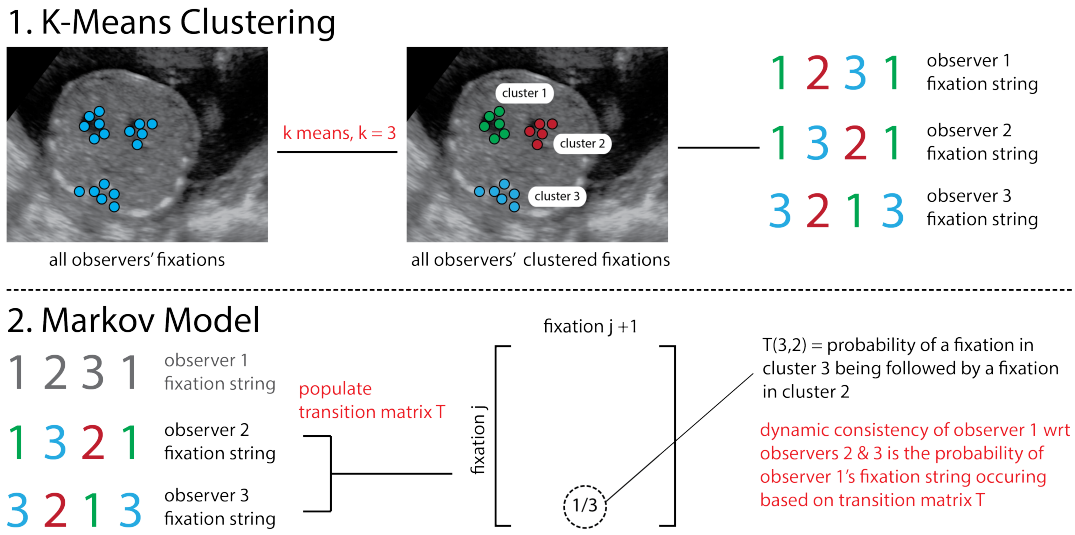


Figure 3.5: An illustration of dynamic consistency calculation, showing the clustering of the fixations of all observers into ROIs using k-means clustering, the generation of fixation strings, and the training of a Markov model on the fixation strings of all-but-one observers to predict one observer's fixation string.

.

Torralba et al.[82] derived a Bayesian model combining 'top-down' and 'bottom-up' scene information to compute visual saliency. The authors combined local and global image descriptors computed using steerable pyramids over small and large spatial scales respectively, to identify statistically improbable and therefore visually salient image regions. This model attained a fixation prediction accuracy of 68% on the MIT benchmark dataset[107].

Ehinger et al.[8] recorded the eye movements of 14 observers tasked with searching for human targets in 912 outdoor images. Using a weighted combination of low level features including gradient and intensity, mid level features including a Dalal-Triggs[73] person detector, and high level scene context features and constraints including a horizon detector, the authors trained a linear SVM to generate saliency maps, which predicted fixated image regions with a mean accuracy of 94%. However, the choice of local features was not optimised to determine which image features were the most accurate predictors of human fixations.

Mathe et al.[9] investigated visual saliency with videos of natural scenes. The authors recorded the eye movements of 16 observers viewing 497,000 video frames

and tasked with identifying human targets performing certain actions including running, hugging, and kicking a ball. Using local intensity, histograms of oriented gradients, and motion boundary histograms, the authors trained a linear SVM to predict fixations or saliency maps on video frames, effectively acting as a spatio-temporal interest point operator. When used in conjunction with a BoVW action classification framework, this method produced a mean accuracy of 91.5% and outperformed the Harris spatio-temporal interest point detector (86.6%) when classifying actions including walking, running, and kicking. The study was the first to show that visual saliency maps based on fixations can be used as interest point operators to boost action recognition.

Kienzle et al.[110] presented an alternative approach to predicting visual saliency maps on unseen images. They recorded the eye movements of 14 observers viewing 200 grayscale outdoor images. The authors computed local signal energy, the entropy of the local intensity histogram, the determinant and trace of the local autocorrelation matrix, and the determinant and trace of the local Hessian at varying scales surrounding fixation points. These features were used to train a linear SVM to classify fixation points on unseen images, again acting as a learned interest point operator.

Kienzle et al.[12] extended this approach to develop a spatio-temporal interest point operator. Here the authors collected the eye movements of ten observers viewing 22 videos featuring humans performing tasks such as playing tennis, and trained a feed-forward neural network to classify likely fixation points on new videos. The accuracy of this learned interest point operator (63.4%) was greater than the Harris (52.2%)[13] and Periodic (55.4%)[111] spatio-temporal interest point detectors in predicting human fixation points on video frames. When used within a BoVW action classification framework, the learned interest point operator yielded accuracies including 95%, 86% and 91% for the classification of running, boxing and clapping actions respectively.

3.5 Conclusions

Eye tracking has been used within the medical imaging fields of radiography and mammography to quantify differences between the visual search strategies of expert and novice observers performing image analysis tasks[86,87,102–105]. Prior to this thesis, these insights have not been harnessed to improve automated medical image analysis frameworks, nor has eye tracking research focused on improving US imaging.

Furthermore, it has been demonstrated that eye tracking inspired visual saliency models and spatio-temporal interest point operators for natural image processing applications can be a viable alternative to interest point operators based on hand-crafted features[9,12]. However, prior to this thesis, similar applications within medical image processing and, specifically, US image analysis remain unexplored.

This thesis presents eye tracking inspired methods for the automated analysis of US images, volumes and videos. Eye tracking experiments are conducted to determine which factors guide the visual attention of human observers interpreting fetal US stimuli. Machine learning frameworks are then implemented to mimic observers' visual search strategies for anatomical landmark localisation, standardised plane selection, and video classification.

Mr. Data, shut up.

— Captain Jean-Luc Picard, Star Trek: Nemesis (2002)

4 Datasets

Contents

4.1 Introduction

Three datasets were used for the eye tracking experiments and the training, validation and testing of the automated US analysis frameworks described within this thesis. They consisted of B-Mode US images, volumes and videos (Figure 4.1). All datasets showed healthy fetuses from $22 - 36$ weeks' GA, and were acquired by experienced sonographers using a Philips HD9 US machine with a V7-3 transducer as part of the global Intergrowth 21^{st} Project [112] with ethics approval in place for their acquisition

and use. The use of US datasets showing only healthy fetuses is a limitation of this thesis; the validation of the machine learning methods described in Chapters 5, 6 and 7 on growth restricted fetuses is therefore a potential area for future work.

All datasets were masked to show only the US scan window; this removed text showing study information and anonymised the data. The green RGB channel was eliminated from the datasets to remove annotations and callipers added by sonographers during acquisition. To account for differences in contrast, all datasets were converted to grayscale and normalised to zero mean and unit variance.

Bounding boxes were manually annotated by the author. High inter-observer agreements have been reported for biometric measurements obtained by manual calliper placement[113,114]; however relying solely on the author's annotations remains a source of variance, as these annotations did not necessarily reflect the true ground truth.

4.2 2-D Ultrasound Images

The first dataset is used within Chapter 5 of this thesis. It consists of 1500 clinical 2-D abdominal US images. All images showed the stomach bubble, umbilical vein, a circular abdominal wall and did not show the kidneys, scoring at least four points out of five according to the ISUOG criteria[1]. The images were upsampled from $745 \times 559px$ to $1439 \times 1080px$ through bicubic interpolation, preserving their aspect ratios. Each image was manually annotated by the author, with bounding boxes to give the ground truth positions of the abdominal wall, stomach bubble, umbilical vein, spine, and ribs where visible. The images were divided into independent datasets for use as stimuli in eye tracking experiments and for the training, validation and testing of the automated anatomical landmark localisation framework described in Chapter 5.

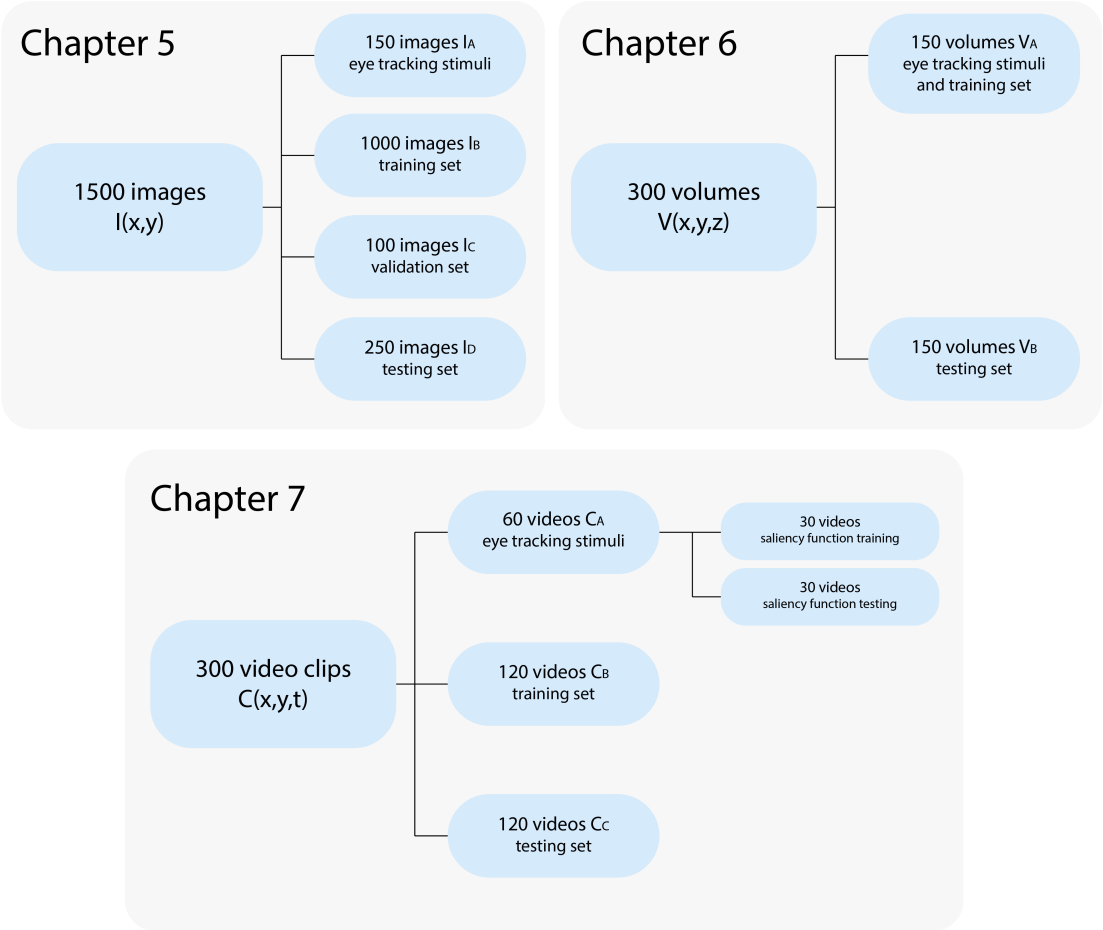


Figure 4.1: The three primary image, volume and video datasets used within this thesis for eye tracking experiments and training, testing and validating automated US analysis frameworks.

4.2.1 Eye Tracking

150 of these images were used as stimuli for eye tracking experiments, resulting in dataset $I_{A\,n}$ where $n = 1, ..., 150$, and alternate images in the sequence were reflected in the x and y axes respectively in order to randomise the positions of key anatomical landmarks (Figure 4.2).

The corresponding ground truth bounding boxes were $I^*_{A\,m,n}$ where $m = 1, ...5$ for the abdominal wall, stomach bubble, umbilical vein, spine and ribs respectively.

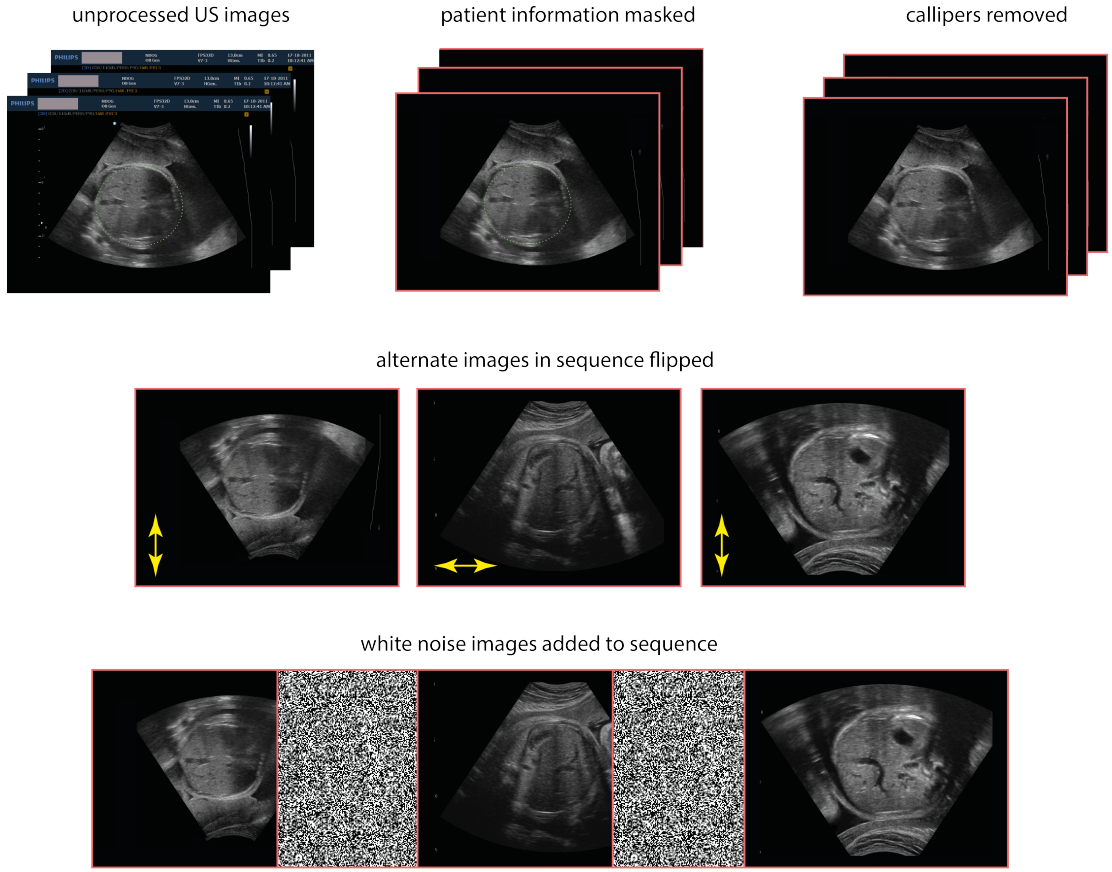


Figure 4.2: The main stages in preparing stimulus images for eye tracking: masking patient information, removing manually placed callipers, reflecting alternate images in the stimulus sequence in the x and y axes respectively, and displaying a Gaussian white noise image between each stimulus image

.

4.2.2 Training

1000 of these images were used as a training set, forming dataset $I_{B\,n}$ where $n = 1, ..., 1000$. The corresponding ground truth bounding boxes were $I^*_{B\,m,n}$ where $m = 1, ..., 4$ denoting bounding boxes for the abdominal wall, stomach bubble, umbilical vein and spine respectively.

4.2.3 Validation

100 images were used as a validation set, forming dataset $I_{C\,n}$ where $n = 1, ..., 100$. The corresponding ground truth bounding boxes were $I^*_{C\,m,n}$.

4.2.4 Testing

The remaining 250 images were used as a testing set, forming dataset $I_{D\,n}$ where $n = 1, ..., 250$. The corresponding ground truth bounding boxes were $I^*_{D\,m,n}$.

4.3 3-D Ultrasound Volumes

The second dataset is used within Chapter 6 of this thesis. It consists of 300 clinical 3-D abdominal US volumes. Each volume $V(x, y, z)$ had spatial dimensions x and y corresponding to a single transverse view of the fetal abdomen, and a third spatial dimension z corresponding to a progression along the longitudinal axis of the fetal abdomen (Figures 4.3 & 4.4) where pixels at a given z co-ordinate were defined as a single volume frame so that the f^{th} frame of the n^{th} volume was given by $V_{n,f}$. Volume frames were upsampled from $745 \times 559px$ to $1439 \times 1080px$ through bicubic interpolation, preserving their aspect ratios.

Each volume contained at least one frame which scored at least four out of five points according to the ISUOG criteria [16] for standardised abdominal plane selection, and each volume was manually labelled to give the ground truth z co-ordinates of these standardised abdominal planes. Each volume frame was also manually annotated by the author, with bounding boxes to give the ground truth positions of the abdominal wall, stomach bubble, umbilical vein, spine, and ribs where present. These 2-D bounding boxes representing the same anatomical structure on consecutive frames were grouped to form 3-D bounding 'envelopes'.

The volumes were divided into independent datasets for use as stimuli in eye tracking experiments and for the training and testing of the automated standardised plane selection framework described in Chapter 6.

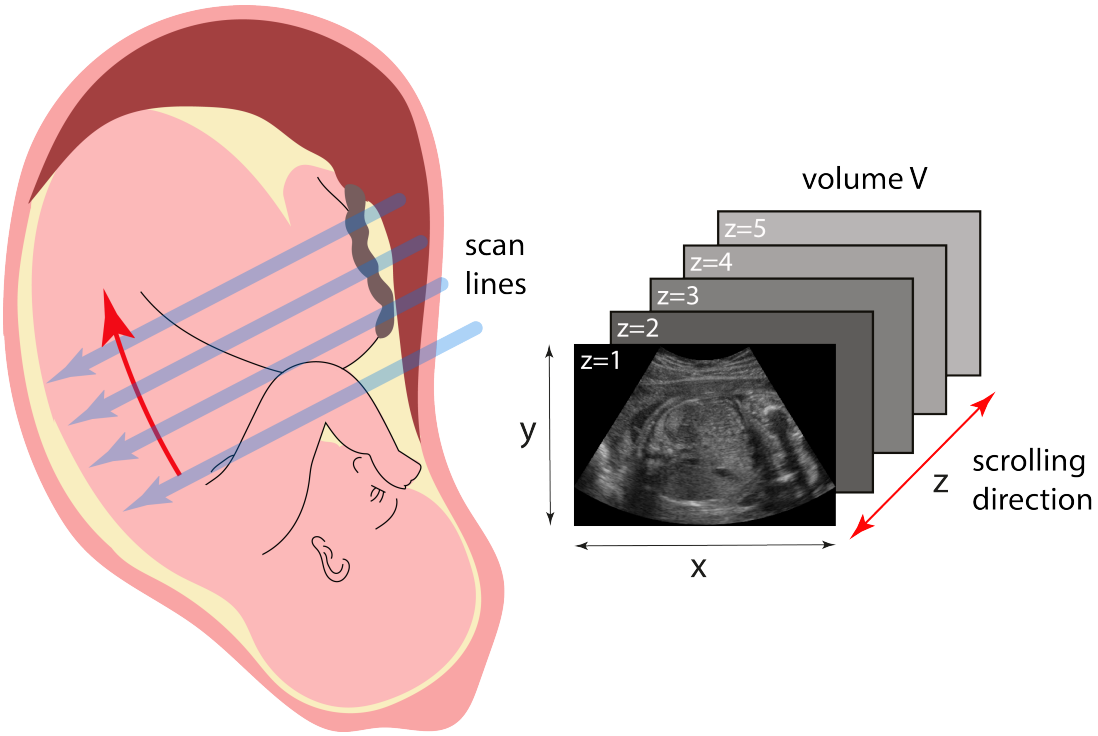


Figure 4.3: The acquisition of US volume stimuli, consisting of a sweep across the the fetal abdomen, resulting in a sequence of volume frames showing a cross-sectional progression through the longitudinal axis of the fetal abdomen. Transducer scan lines (blue) are parallel to the acquired image planes, and perpendicular to the direction of scrolling allowed through the volumes (red) by experimental observers in the eye tracking experiments.

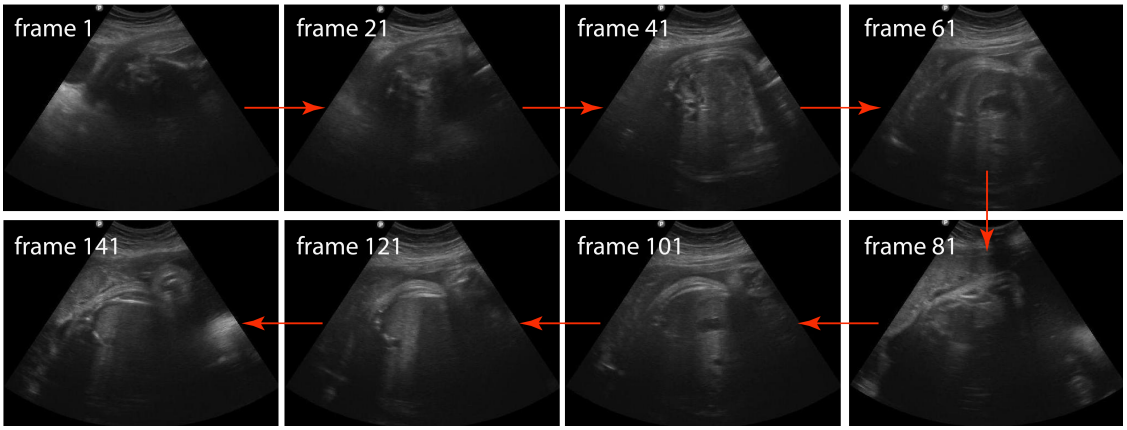


Figure 4.4: A sequence of frames taken from a 3-D fetal abdominal US volume used as an eye tracking stimulus. The frames are extracted from the longitudinal axis of the volume, and show a progression from the superior end of the fetal abdomen (frame 1) where the abdominal wall is not clearly visible, to the standardised abdominal plane (frame 61) to the inferior end of the fetal abdomen (frame 141) where the abdominal wall is once again not clearly visible.

4.3.1 Eye Tracking

150 of these volumes were used as stimuli for eye tracking experiments, and as a training set, resulting in dataset $V_{A\,n}$ where $n = 1, ..., 150$. The accompanying ground truth annotations for the f^{th} frame of the n^{th} volume were $V^*_{A\,m,n,f}$ where $m = 1, ..., 5$ for the abdominal wall, stomach bubble, umbilical vein, spine, and ribs respectively and $V^*_{A\,\text{plane}\,n,f} = 1$ for a ground truth standardised plane and 0 otherwise.

Bounding boxes representing the same anatomical structure on consecutive volume frames, with a Dice overlap coefficient ($Dice = \frac{2(GT \cap DT)}{GT+DT}$, where GT and DT are the sets of pixels defined by the ground truth and detected bounding boxes respectively) greater than 0.75, the threshold employed by Rahmatullah et al.[7], were grouped to form the 3-D ground truth bounding 'envelopes' $V^{*3D}_{A\,m,n}$.

Bounding boxes were manually annotated around artefacts and anatomical structures displaying similar appearances to the umbilical vein and spine, resulting in sets of bounding boxes $V^*_{A\text{uv}\,n,f}$ and $V^*_{A\text{sp}\,n,f}$ for umbilical vein-like and spine-like artefacts respectively. Bounding boxes representing the same artefact on consecutive volume frames, with a Dice overlap coefficient greater than 0.75, were grouped to form the 3-D ground truth bounding 'envelopes' where the p^{th} artefact envelopes in the n^{th} volume were represented by $V^{*3D}_{A\text{uv}\,n,p}$ and $V^{*3D}_{A\text{sp}\,n,p}$ for umbilical vein-like and spine-like artefacts respectively.

4.3.2 Testing

The remaining 150 volumes were used as a testing set, forming dataset $V_{B\,n}$ where $n = 1, ..., 150$. The corresponding ground truth annotations for the f^{th} frame of the n^{th} volume were $V^*_{B\,m,n,f}$, $V^{*3D}_{B\,m,n}$, and $V^*_{B\,\text{plane}\,n,f}$ where $m = 1, ...4$ for the abdominal wall, stomach bubble, umbilical vein and spine.

4.4 2-D+t Ultrasound Videos

The final dataset is used in Chapter 7 of this thesis, and consists of 100 clinical 2-D+t US videos acquired at $28fps$ via a sweep of the probe along the longitudinal axis of

the maternal abdomen (Figure 4.6). Video frames were upsampled from $745 \times 559px$ to $1439 \times 1080px$ through bicubic interpolation, preserving their aspect ratios. Each 'crown to rump' video was truncated into three or more shorter video clips showing:

1. A complete sweep across the fetal abdomen.
2. A complete sweep across the fetal head.
3. A sweep across another portion of the fetal anatomy, typically the legs or heart.

This resulted in a total of 300 video clips of which 100 showed the fetal head only, 100 showed the fetal abdomen only, and 100 showed other parts of the fetal anatomy only. For brevity these three classes of video clip will be referred to in this thesis as 'head', 'abdomen' and 'other' respectively.

As shown in Figure 4.5 each video clip consisted of a series of 2-D fetal US frames concatenated in the third, temporal direction. The truncated clips were of a mean length of $11s$ each, with frame sequences similar to those shown in Figure 4.7. All individual video frames were normalised to have zero mean and unit variance resulting in a dataset where each clip $C(x, y, t)$ had spatial dimensions x and y corresponding to a single video frame, and a temporal dimension t.

Each video was manually annotated by the author with the label 'head', 'abdomen', or 'other'. Each video frame was also manually annotated with bounding boxes to give the ground truth positions of anatomical regions of interest (ROIs), namely the head cavity in head videos and the abdominal cavity in abdominal videos. No equivalent ROI was annotated in video clips showing other parts of the fetal anatomy, as no single portion of the fetal anatomy was consistently visible across these video clips.

The video clips were divided into independent datasets for use as stimuli in eye tracking experiments and for the training and testing of the automated video classification framework described in Chapter 7.

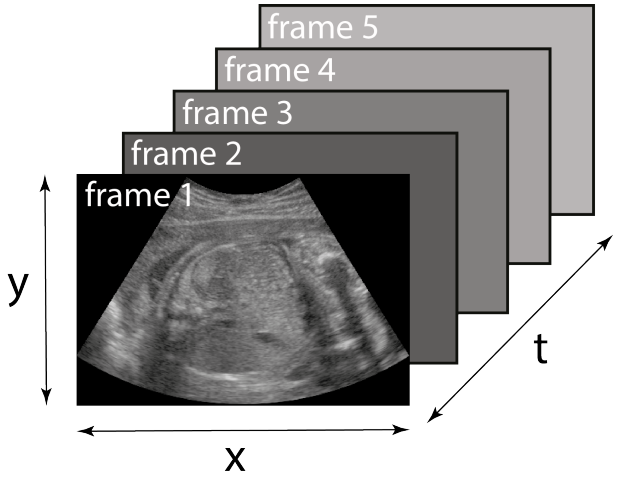


Figure 4.5: The dimensions of the truncated 2-D+t fetal US video clips used as experimental stimuli. Each video consisted of a series of 2-D fetal US frames, or images, concatenated in the third, temporal, direction.

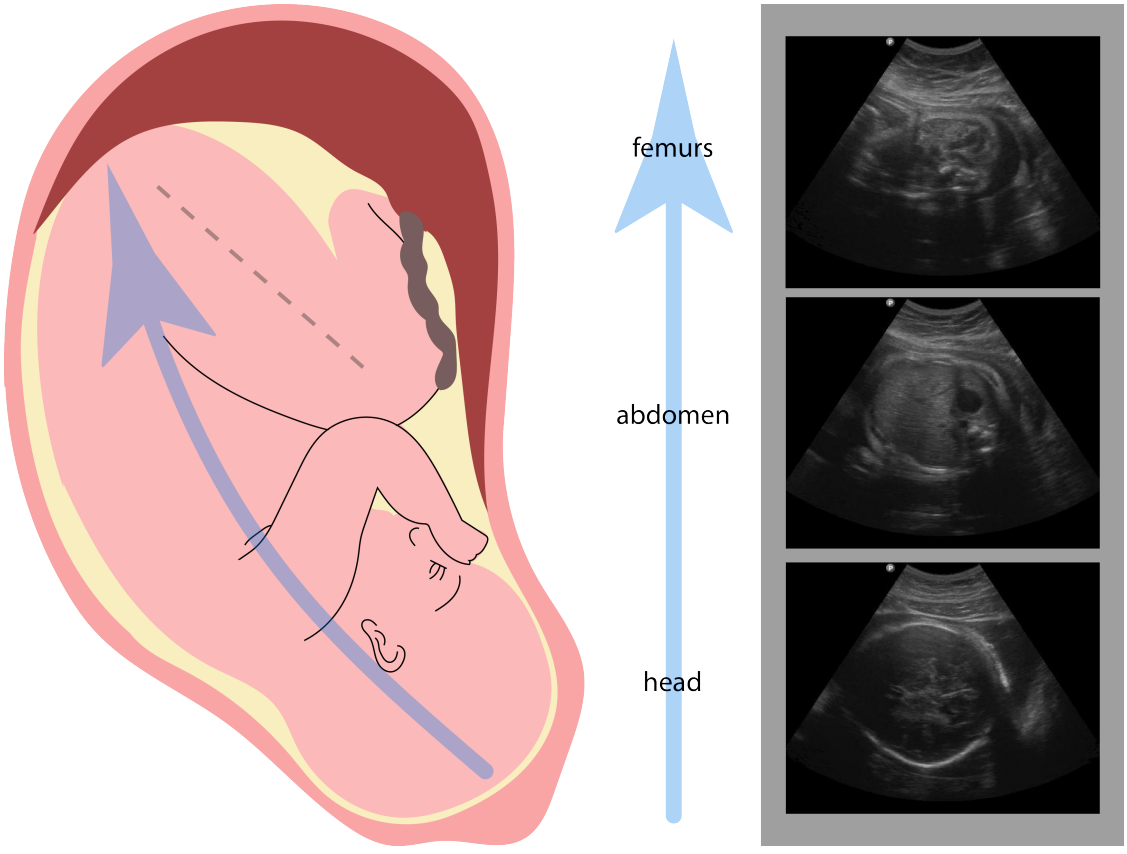


Figure 4.6: The acquisition of 2-D+t fetal US video clips. Each video clip consisted of a sweep of the US probe along the longitudinal axis of the maternal abdomen, for example resulting in a sequence of video frames showing a progression through the fetal head, abdomen, and femur along the vertical axis of the fetal body assuming the fetal position shown above.

4.4.1 Eye Tracking and Visual Saliency

60 truncated video clips were used as stimuli in eye tracking experiments, of which 20 each had the ground truth labels 'head', 'abdomen', and 'other'. They formed dataset $C_{A\,n}$ where $n = 1, ..., 60$. 50% of this dataset ($C_{A1\,n}$) was used for training a learned interest point operator, and the remaining 50% ($C_{A2\,n}$) was used for testing the operator and as a validation set. The corresponding ground truth class labels for each video were $C^*_{A\,n} \in 1, 2, 3$ denoting 'head', 'abdomen' and 'other' respectively and $C^*_{\mathrm{ROI}A\,m,n}$ with $m \in 1, 2$ corresponding to ground truth bounding boxes for the head cavity and abdominal cavity respectively.

4.4.2 Training

120 truncated video clips were used as a training set, of which 40 each had the ground truth labels 'head', 'abdomen', and 'other'. They formed dataset $C_{B\,n}$ where $n = 1, ..., 120$, and the corresponding ground truth class labels for each video were $C^*_{B\,n}$ and $C^*_{\mathrm{ROI}B\,m,n}$.

4.4.3 Testing

120 truncated video clips were used as a testing set, of which 40 each had the ground truth labels 'head', 'abdomen', and 'other'. They formed dataset $C_{C\,n}$ where $n = 1, ..., 120$, and the corresponding ground truth class labels for each video were $C^*_{C\,n}$ and $C^*_{\mathrm{ROI}C\,m,n}$.

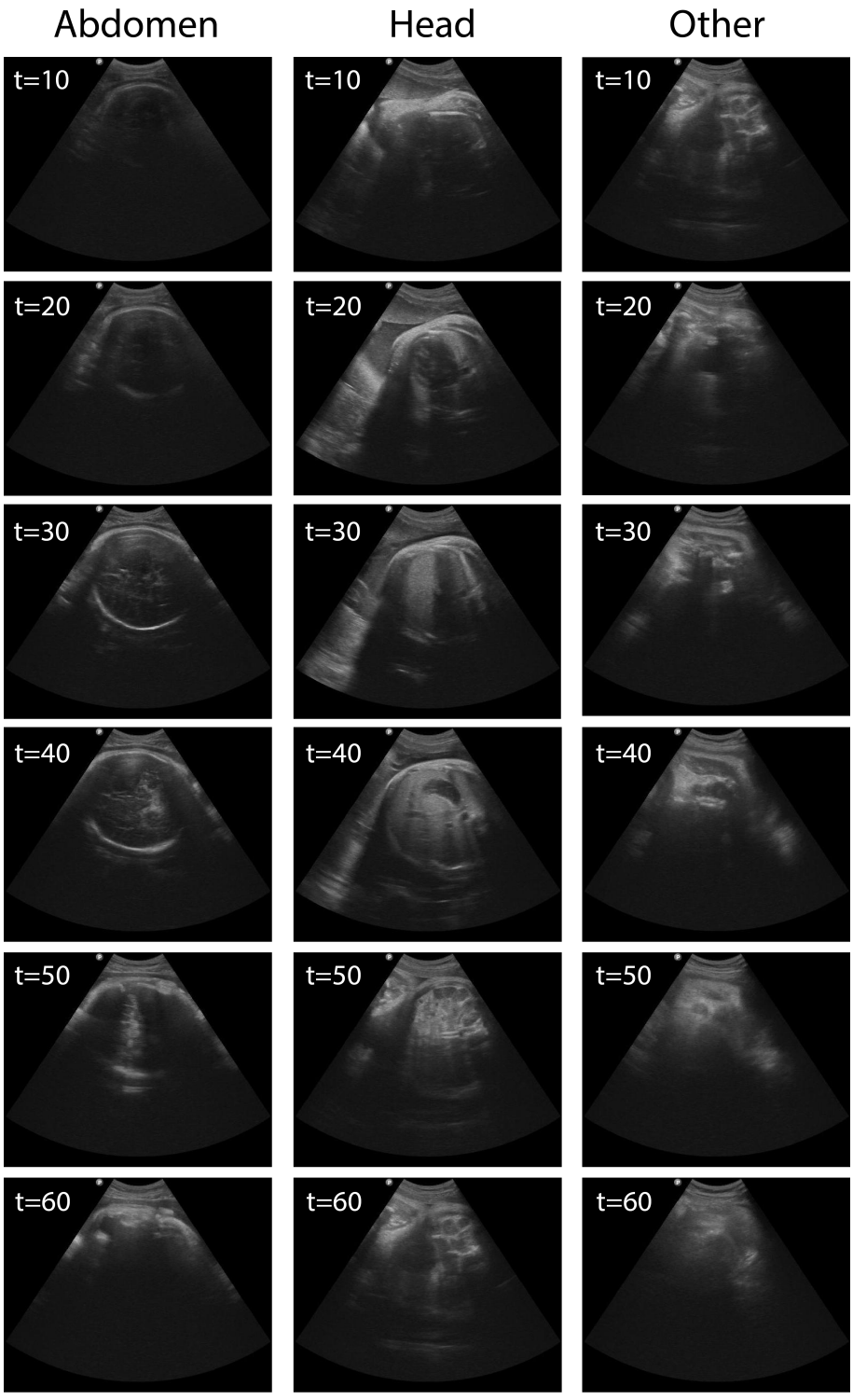


Figure 4.7: Examples of frame sequences within truncated fetal US video clips showing (left) the fetal head (centre) the fetal abdomen (right) the fetal femur.

The eyes have it.

— Philip K Dick, American Author, (1928-1982)

5

An Eye Tracking Inspired Method for Anatomical Landmark Localisation

Contents

5.1 Introduction

Localising the stomach bubble and umbilical vein in 2-D fetal abdominal US images is a first step in automating the process of standardised abdominal plane acquisition. As discussed in Chapter 2, existing approaches to this problem frequently misclassify the stomach bubble, umbilical vein, shadows and artefacts due to their similar appearance[24]. In this first contributions chapter, these limitations are addressed by introducing high level constraints used by the human visual system into a pictorial

structures model for stomach bubble and umbilical vein localisation.

This chapter focuses specifically on the abdominal plane due to the particular significance of the AC measurement for fetal growth monitoring[4], and the challenging nature of this plane compared to other standardised imaging planes. The lack of a well defined abdominal wall boundary, the similarities in appearance between the stomach bubble and umbilical vein, and challenges associated with hand crafted features leave significant potential for existing fetal abdominal US analysis methods to be improved through eye tracking. Additionally it is hypothesised that the presence of discrete anatomical landmarks within abdominal images confirms their suitability for eye tracking experiments, compared to other planes showing structures such as the fetal femur.

A series of eye tracking experiments were conducted whereby observers were asked to localise the stomach bubble and umbilical vein in 2-D abdominal US images. At the time of submission, this was the first eye tracking study conducted with US images as stimuli. In order to draw general conclusions on the visual search strategies of observers it was necessary to assess the degree of similarity between their eye movements. The spatial and temporal similarity between the eye movements of observers was computed found to be significantly higher than control values. An analysis of the gaze trajectories of observers was then carried out, demonstrating that observers exhibited a two-component visual search strategy and, despite not appearing in ISUOG[1,16] guidelines on standardised abdominal plane acquisition, the spine was fixated on by most observers and was utilised as a reference point.

These findings were used to derive a pictorial structures model to mimic the visual search strategies of observers, improving on the stomach bubble and umbilical vein localisation accuracies reported by Rahmatullah et al.[24]. Appearance priors for the stomach bubble, umbilical vein and spine were learned via a cascade of boosted decision stumps, and geometric priors were constructed using the relative positions of these structures across a set of training images.

5.2 Originality and Individual Role

Independently, I designed and wrote Python applications to interface with eye tracking hardware, acquired and post-processed eye movements, and analysed the resulting data. Using an open source Matlab library[115] I trained and tested sliding window detectors for the detection of candidate anatomical landmarks. Independently, I wrote applications to implement, train and test a 2-D pictorial structures model.

5.3 Eye Tracking

Eye tracking experiments were conducted to determine which anatomical features were fixated on by observers viewing 2-D abdominal US images, the degree of spatial and temporal similarity between their fixation points, and whether their visual search strategies adhered to a global-focal model of visual search.

5.3.1 Methods

Stimuli

The eye tracking stimuli consisted of dataset $I_{A\,n}$ as described in Chapter 4. Stimuli were presented to observers on a 25-inch $1920 \times 1080px$ LCD monitor, through a custom built user interface which received raw x and y gaze co-ordinates averaged across the left and right pupils, timestamps, and observer-to-screen distances at a sampling frequency of $30Hz$. The lack of stimulus images showing growth restricted or otherwise abnormal fetuses was a limitation of this experiment and may have led to observers learning likely orientations and appearances of anatomical landmarks rather than engaging in active visual search. The randomised orientations of stimulus images was, however, intended to mitigate this effect.

Hardware

Eye movements were recorded with an EyeTribe v1.0 (the Eye Tribe, Denmark) eye tracker.

Participants

Ten observers, with normal acuity, participated in this study. The observers were divided into two groups, experts and novices, based on their level of experience in US imaging and interpretation. The expert group consisted of one clinical fellow with three years' US experience and four second year PhD candidates in Biomedical Engineering. The novice group consisted of three first year PhD candidates in Biomedical Engineering, and two undergraduate clinical Medicine students. The distinction between expert and novice observers was defined by the author based on observers' familiarity with US image interpretation, and was not based on clinical guidelines or definitions. Ethics approval for this experiment was obtained via the University of Oxford Central University Research Ethics Committee (Reference: MS-IDREC-C1-2015-166).

A power analysis was conducted to determine the effect size that would be observable between the expert and novice groups assuming a power (or true positive probability) of 0.9 and a significance level (or false positive probability) of 0.05. As shown in Figure 5.1, the effect size given these parameters was 5.78% which is lower than the effect sizes reported between expert and novice observers in mammography eye tracking experiments conducted by Nodine et al.[116,117] (57.18% and 22.26%) and Bertram et al.[104] (32.43%). Ten observers was therefore deemed to be a sufficient sample size.

Experimental Procedure

Calibration was conducted for each observer prior to stimulus presentation. Observers were instructed to fixate on each of nine equally spaced circular targets; where the error between the observer's gaze position and a target centre was greater than 1.5° visual angle, the calibration process was repeated for that target until all errors fell below this threshold.

Each observer was presented with each stimulus image in a pre-determined sequence, with unlimited viewing time available for each image. Observers were

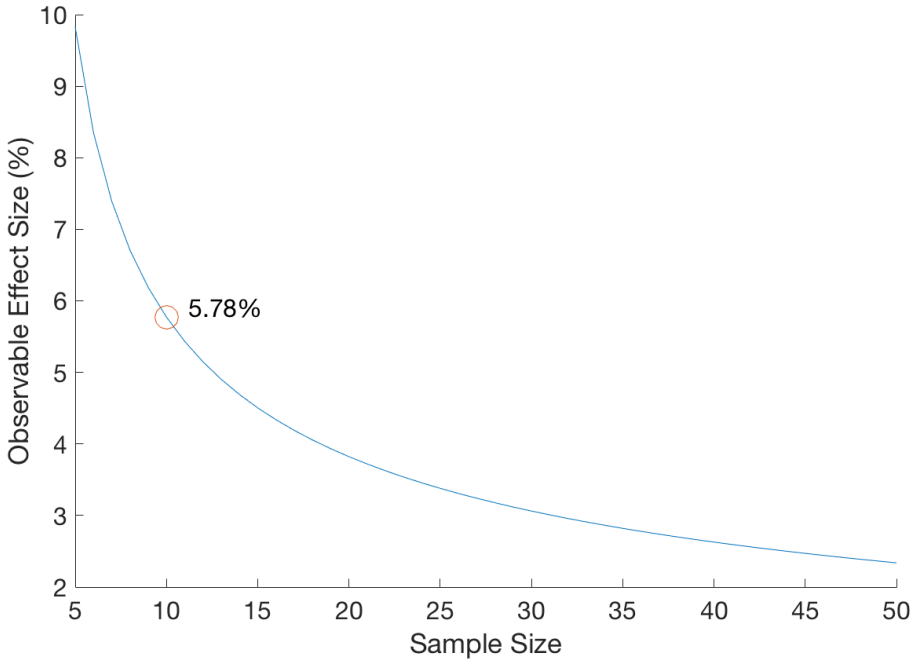


Figure 5.1: The results of a power analysis, showing the observable effect size between expert and novice observer groups against sample size. The effect size for a sample of ten observers is shown in red.

.

instructed to verbally report a score out of five for each image according to the ISUOG criteria[1] before proceeding to the next image using the mouse button. This experiment was designed to replicate the task, performed by sonographers, of applying Solomon's criteria to images acquired during an US scanning session to assess whether they qualify as standardised imaging planes. The verbal reporting of image scores was a deviation from clinical practise, and may have altered the visual search behaviours of observers. A $1920 \times 1080px$ white noise image was generated by sampling from a Gaussian distribution ($\mu = 0.5, \sigma = 0.2$) and displayed for $5s$ between each stimulus image to de-focus the observer's gaze. To avoid fatigue, observers were given a ten minute break between each group of 50 images.

This resulted in a set of raw gaze co-ordinates, where the j^{th} observer's gaze data on the n^{th} image was described by $R_{n,j}$ and each set of gaze co-ordinates $R(x, y, d_{\text{os}}, t)$ consisted of the mean x and y co-ordinates across the left and right pupils, the observer-to-screen distance d_{os} and a timestamp t.

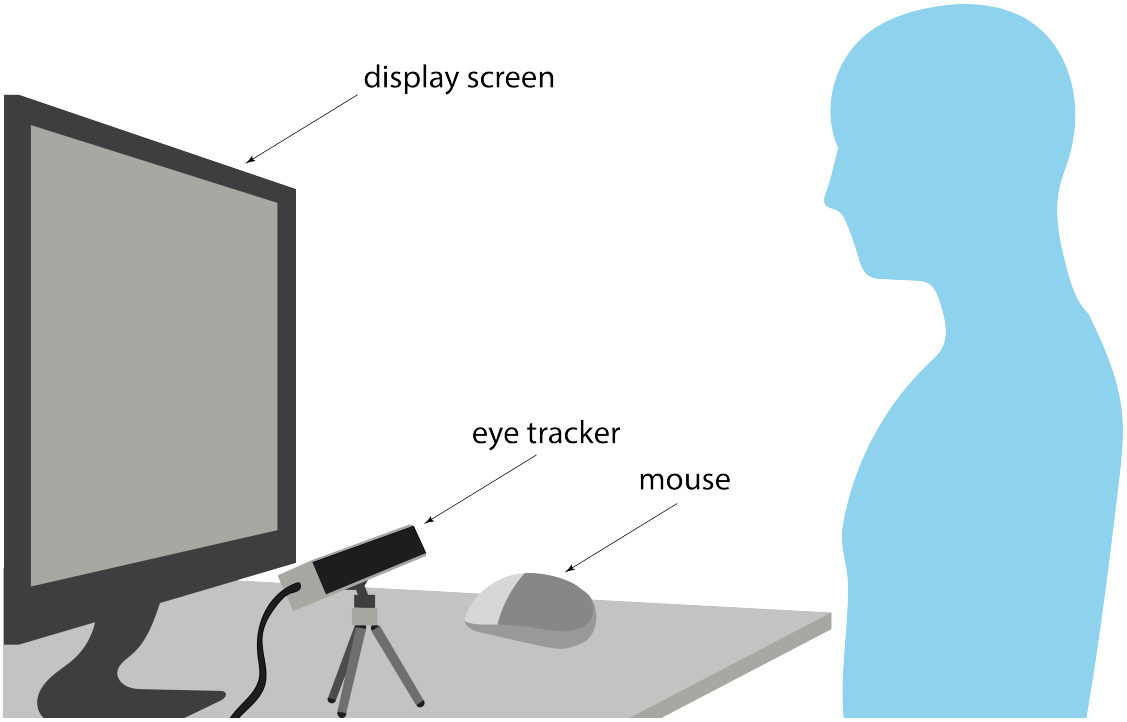


Figure 5.2: The experimental setup for tracking eye movements on 2-D abdominal US images, consisting of a display screen for stimulus presentation, a mouse for progressing through the sequence of stimulus images, and a USB connected EyeTribe eye tracking device.

.

Fixation Filtering

Raw gaze data was filtered into fixations and saccades using an I-VT algorithm[97]. Linear interpolation was used to rectify gaps in gaze data greater than $33ms$ (one sample) due to tracking errors or observer blinks. An averaging low pass filter was applied to the gaze data with a sliding window size of $99ms$ (three samples), to reduce high frequency noise caused by eye tremors. The angular velocity ($V = \frac{2\, arctan(\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}{)}(d_{\text{os}\,2} + d_{\text{os}\,1})(t_2 - t_1)$ where $x_i, y_i, d_{\text{os}\,i}$ and t_i are the mean x and y co-ordinates across the left and right pupils, observer-to-screen distance $d_{\text{os}\,i}$ and timestamps t_i respectively of two consecutive gaze points) of the gaze co-ordinates across $66ms$ (two samples) windows was calculated and a velocity threshold of $30°/s$ was applied in accordance with the fixation filter settings employed by Tobii[97]. Gaze points with angular velocities below

the threshold were classified as fixations, and those above the threshold classified as saccades. Fixations occurring less than 0.5° apart spatially were merged and classified as a single fixation, with a centroid calculated as the mean of the merged fixation points. Fixations shorter than $80ms$ in duration were discarded, again in accordance with the parameters employed by Tobii [97]. This resulted in a set of fixations, where the j^{th} observer's fixations on the n^{th} image were described by $F_{n,j}$ and each set of gaze co-ordinates $F(x, y, t)$ consisted of the fixation's x and y co-ordinates and a timestamp t.

Regions of Interest

The proportions of fixations falling inside the ground truth bounding boxes described in Chapter 4 ($I^*_{A\,m,n}$ where $m = 1, ...5$ for the abdominal wall, stomach bubble, umbilical vein, spine and ribs respectively) were computed for each stimulus image where present. These ROIs (Figure 5.3) were selected due to their consistent visibility in the majority of stimulus images.

Static Consistency

Static consistency was computed to assess the agreement between the fixated locations of observers. A set of binary fixation maps $B_{n,j} \in 0, 1$ (with images denoted by $n = 1, ..., 150$ and observers by $j = 1, ...10$) was computed for each observer and image such that pixel values corresponding to fixation points were incremented by 1, and all other pixels had value 0. As the human field of view typically extends 1.5° around a fixation point [9], the binary maps were convolved with a 2-D Gaussian kernel resulting in a set of attentional maps $A_{n,j} = B_{n,j} * G(\sigma)$ (where $\sigma = 15px$ corresponding to 1.5° visual angle with an observer-to-screen distance of $0.5m$) representing each observer's visual attention, or foveated area, on each stimulus image.

Static consistency across all observers on a particular image was then computed via a leave-one-out scheme. For each observer, the attentional maps of all other

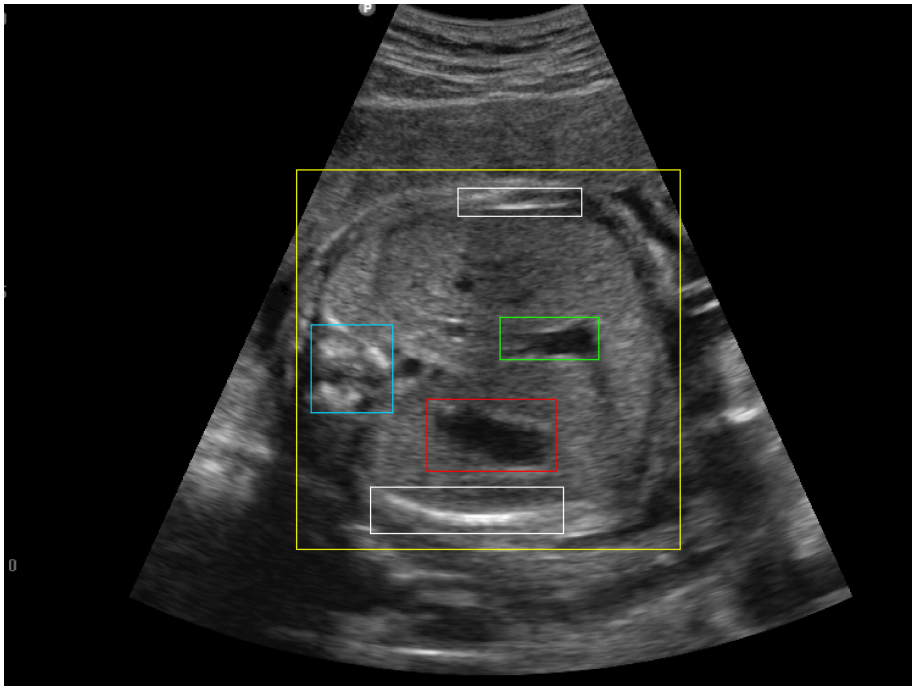


Figure 5.3: Manually annotated bounding boxes showing anatomical ROIs in stimulus images, namely the stomach bubble (red), umbilical vein (green), spine (blue), abdominal cavity (yellow), and ribs (white)

.

observers were summed so that $P_{n,J} = \Sigma_{1 \leq j \leq 10 j \neq J} \frac{A_{n,j}}{A_{max}}$ describes a predictive map for the J^{th} observer's fixations on the n^{th} image, normalised by the maximum value of the summed maps A_{max}. A varying threshold between 0 and 1 was applied to binarize the predictive map. At each threshold level, the sensitivity and specificity ($specificity = \frac{nTP}{nTP + nFN}$ and $sensitivity = \frac{nTN}{nTN + nFP}$, where nTP, nTN, nFP, nFN are true positive, true negative, false positive and false negative counts respectively) of the predictive map in predicting the current observer's attentional map were recorded. This resulted in a receiver-operator characteristic (ROC) curve with the area under curve (AUC) taken as the static consistency score. This was repeated across all observers and all images to compute a mean static consistency score.

Static consistency was computed for four sub-groups of observers: experts vs. novices (expert fixations predicted by novices), novices vs. experts (novice fixations predicted by experts), novices only, and experts only.

A cross-image control measure was calculated to assess how accurately the attentional map of one observer on one randomly selected image could be predicted by the summed attentional maps of all other observers on a different randomly selected image. Repeated for 150 randomly generated image and observer combinations, the mean cross-image static consistency acted as a baseline static consistency score. Due to biases in the stimuli, and oculomotor bias, for example the tendency to initially fixate on the centre of images[82], the control value of static consistency was expected to be greater than random chance (50%) but less than static consistency scores computed for many observers on a single stimulus image.

Dynamic Consistency

Dynamic consistency was computed to assess the agreement between the temporal sequences of the fixated locations of observers. A fixation sequence was generated for each observer on each image. This encoded the order in which observers fixated on the ROIs described in Section 5.3.1, namely the abdominal wall, stomach bubble, umbilical vein, spine, ribs, and background.

The fixation sequences were treated as Markov chains, where the probability of transitioning from a given state (or ROI) to the next was modelled as a Markov process. Dynamic consistency for a particular image across all observers was calculated on a leave-one-out basis. For each observer, the fixation sequences of all other observers were used to populate a transition matrix T, where the element $T(a,b)$ gives the probability of transitioning from state a to b ($T(a,b) = P(S_{t+1} = b|S_t = a) = \dfrac{1 + \Sigma_i\Sigma_j[f_i^j = a][f_i^{j+1} = b]}{C + \Sigma_i\Sigma_j[f_i^j = a]}$, where the j^{th} state of subject i is given by f_i^j and C is a Laplace smoothing term representing the total number of ROIs or states, accounting for sparsity in the training data used to populate the transition matrix).

The dynamic consistency score was computed as the posterior probability of the current observer's fixation string occurring ($P(g) = \Pi_j P(S_{t+1} = g^j|S_t = g^{j-1})$, where g is the current observer's fixation string), based on the Markov model trained on the fixation strings of all other observers.

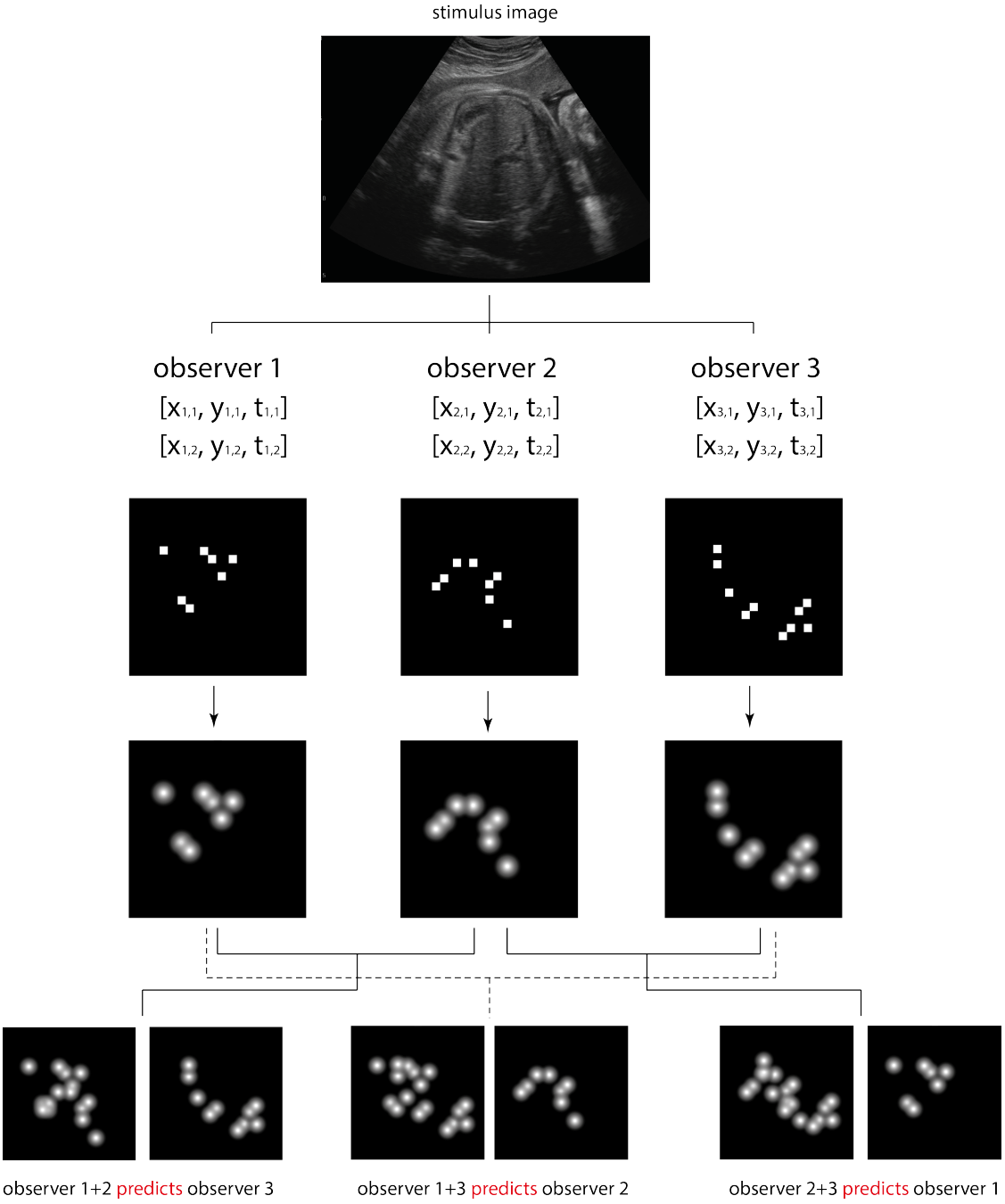


Figure 5.4: The main stages in computing static consistency: constructing individual fixation maps from a list of fixations on a given stimulus image, blurring to give attentional maps, and using a leave-one-out approach to predict each observer's attentional map using a sum of the attentional maps of all other observers, with the AUC of the resulting ROC curve taken as the static consistency.

.

Dynamic consistency was computed for four sub-groups of observers: experts vs. novices (expert fixation sequences predicted by novices), novices vs. experts (novice fixation sequences predicted by experts), novices only, and experts only.

As in Section 5.3.1, a cross-image control measure was calculated to assess how accurately the fixation sequence of one observer on one randomly selected image could be predicted by the fixation sequences of all other observers on a different randomly selected image. Repeated for 150 randomly generated image and observer combinations, the mean cross-image dynamic consistency acted as a baseline dynamic consistency score.

Global-Focal Search

Leong et al.'s[102] method for validating the discovery and reflective stages of visual search, as proposed by Kundel et al.[10], was applied to the collected gaze data.

For each image the Euclidean distance between each observer's gaze point and the centre of the stomach bubble bounding box (the most frequently fixated ROI), was computed as a function of time. Each observer's gaze trajectory with respect to the stomach bubble was modelled as a Gaussian mixture model (GMM), specifically a weighted sum of two Gaussian distributions ($m(x|c,\mu,\sigma) = \Sigma_{i=1}^{2} c_i g_i(x,\mu_i,\sigma_i)$ where $m(x|c,\mu,\sigma)$ is the sum of two Gaussian distributions with weights c_1, c_2, means μ_1, μ_2 and standard deviations σ_1, σ_2 and $g(x|\mu_i,\sigma_i)$ is the i^{th} Gaussian component of this mixture described by $g(x|\mu_i,\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp(-\frac{(x-\mu_i)^2}{2\sigma_i^2})$ with mean μ_i and standard deviation σ_i). The first Gaussian component represented the discovery phase of visual search where a detailed inspection of potential stomach bubble candidates was conducted, and the second component represented the reflective phase where candidate positions and appearances were cross-referenced and a final visual search decision was made.

The parameters for each distribution were found using the Expectation-Maximisation algorithm, with initial parameter values set to 0.5. The 'expectation' step ($P_i(x) = \frac{c_i^{old} g_i(x|\mu_i^{old},\sigma_i^{old})}{\Sigma_{k=1}^{2} c_k^{old} g_k(x|\mu_k^{old},\sigma_k^{old})}$, where $P_i(x)$ is the posterior probability that an observed data

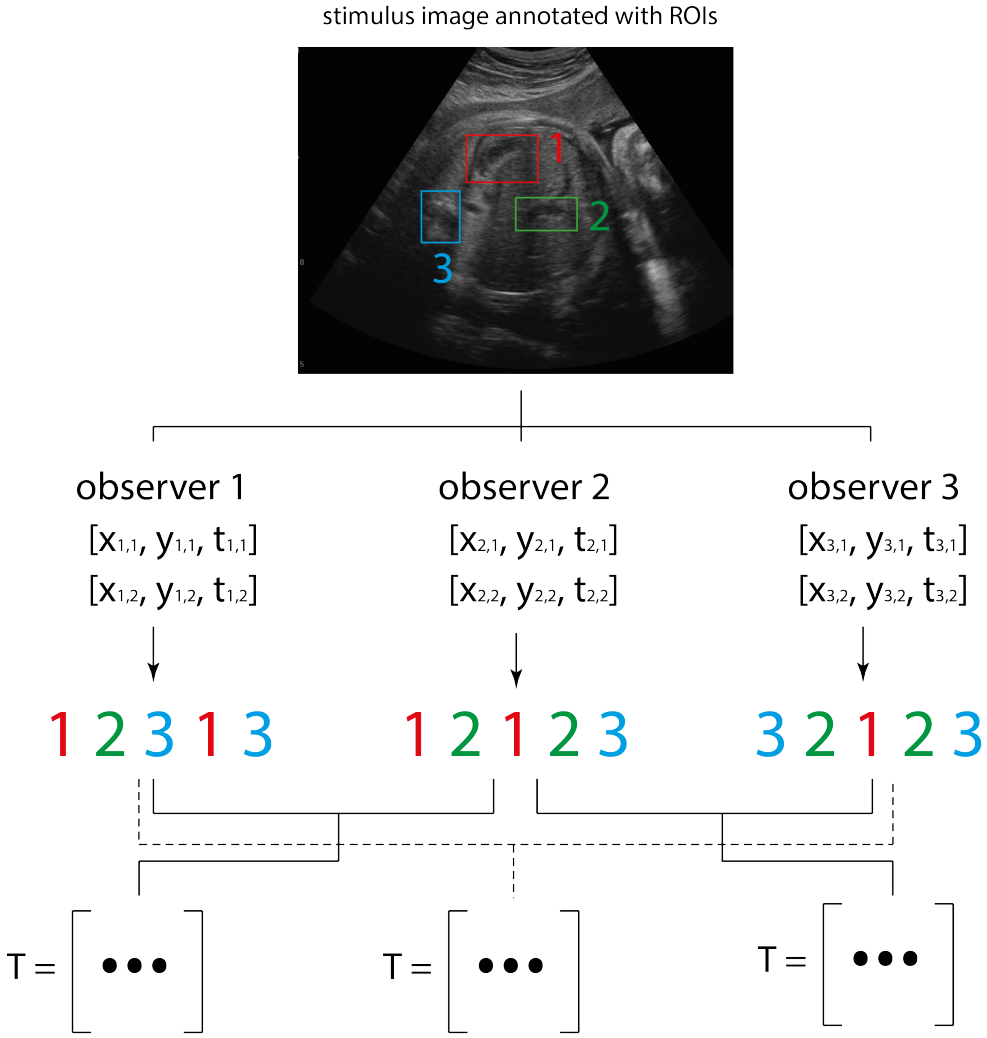


Figure 5.5: The main stages in computing dynamic consistency: constructing individual fixation sequences from a list of fixations on a given stimulus image, and using a leave-one-out approach to predict the posterior probability of each observer's fixation sequence. The posterior probability of the current observer's fixation sequence is taken as the dynamic consistency.

.

point x with value $y(x)$ belongs to component i of the model, and c_i^{old}, μ_i^{old}, and σ_i^{old} are the old estimates for the weights, means and standard deviations of each Gaussian component respectively) computed the posterior probability of the observed datapoints occurring given these initial parameter estimates.

The 'maximisation' step then computed updated parameter values to maximise the posterior distribution ($c_i^{new} = \frac{\Sigma_x y(x) P_i(x)}{\Sigma_x y(x)}$, $\mu_i^{new} = \frac{\Sigma_x y(x) P_i(x) x}{\Sigma_x y(x)}$, $\sigma_i^{new} = \frac{\Sigma_x y(x) P_i(x) [(x-\mu_i^{old})^T (x-\mu_i^{old})]}{\Sigma_x y(x) P_i(x)}$ where c_i^{new}, μ_i^{new}, and σ_i^{new} are the updated estimates for the weights, means and standard deviations of each Gaussian component respectively) and the process was iterated until the parameters converged with an error of less than 1%.

Relative entropy ($D(P||Q) = \Sigma_i P_i log(\frac{P_i}{Q_i})$, where P and Q are two distributions and a low relative entropy suggests a high level of similarity between distributions) was computed as a distance metric between GMMs for all pairings of observers, for each image. The mean value of relative entropy across all images was taken as a measure of similarity between GMMs, and described the extent to which observers exhibited similar two-component search strategies.

The reduced chi-squared statistic ($\chi^2_{red} = \frac{1}{(N-n)} \Sigma_{i=1}^{N} \frac{(O_i - E_i)^2}{\sigma^2}$, where O are observed gaze points, E are the points given by the GMM, σ^2 is the variance of the observed data, N is the number of observations and n is the number of model parameters) was computed for each two-component GMM fitted to each gaze trajectory dataset. This process was repeated for one, three, four and five-component GMMs in order to assess the validity of Kundel et al.'s[10] discovery-reflective search hypothesis and Leong et al.'s[102] subsequent use of a two-component GMM to model these phases of visual search, and to assess the goodness-of-fit of this two-component model compared to models containing greater or fewer Gaussian components.

As a control measure, relative entropies were computed between each possible pairing within 100 two-component GMMs with randomly generated weights, means and standard deviations. The mean relative entropy between these randomly generated distributions was taken as a baseline against which the relative entropy between the gaze trajectories of observers was assessed.

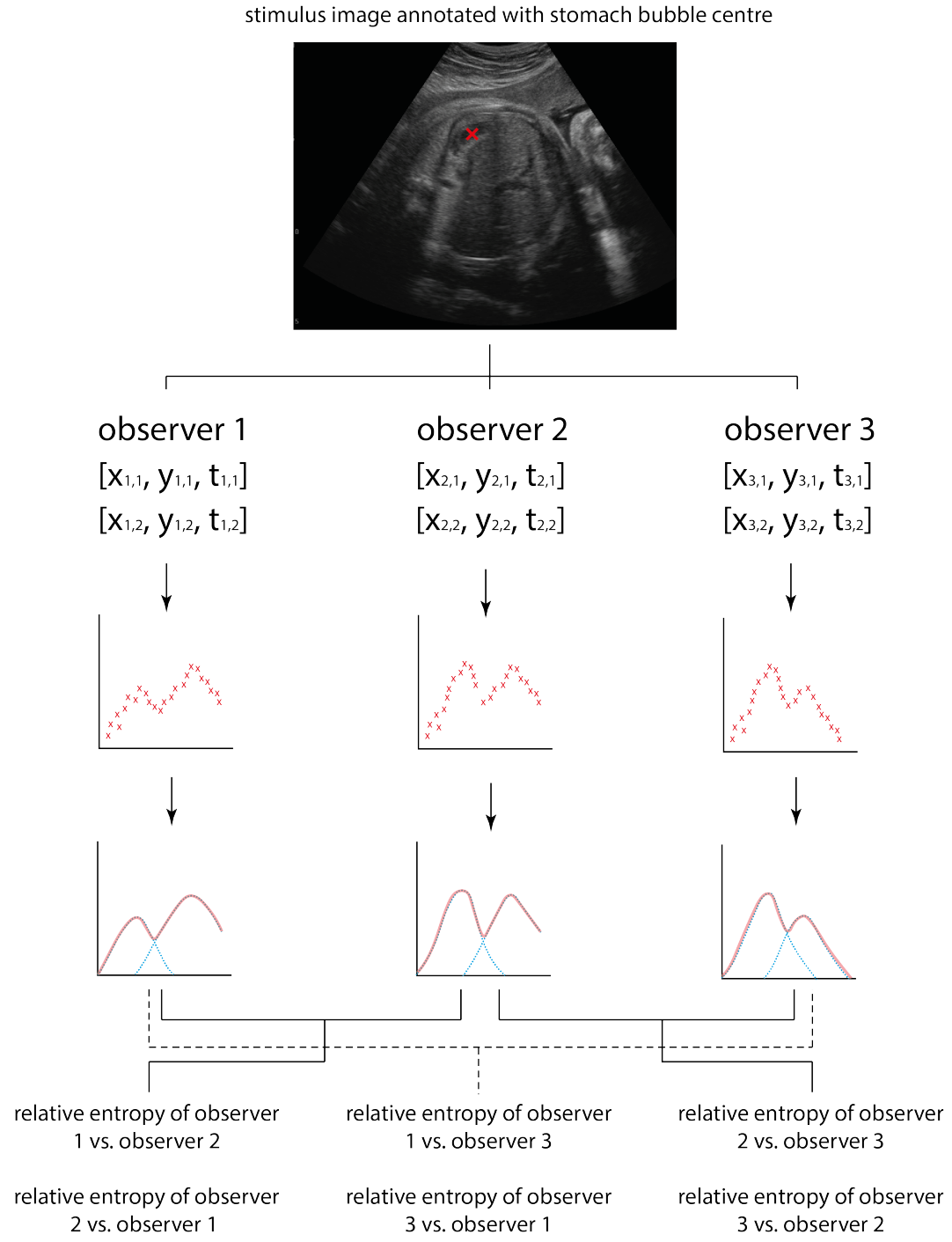


Figure 5.6: The main stages in validating the global-focal model of visual search: computing the Euclidean distance between the fixations of each observer and the stomach bubble as a function of time, fitting a two-component GMM to the data, and calculating the relative entropy between pairs of GMMs.

.

Group	Viewing Time (s)	Time to First ROI (s)	Fixation Length (s)	Fixations per Image
Experts	4.88 ± 1.70	0.92 ± 0.55	0.51 ± 0.23	6.90 ± 3.71
Novices	5.12 ± 1.84	1.01 ± 0.62	0.52 ± 0.25	7.10 ± 3.80

Table 5.1: Mean image viewing times, times to first ROI fixations, fixation lengths and fixations per image for expert and novice observers shown with standard deviations.

5.3.2 Results

Regions of Interest

Mean viewing time per image, mean time to the first fixation on an anatomical ROI (the stomach bubble, umbilical vein, spine or ribs), mean fixation length and the mean number of fixations per image are shown for experts and novices in Table 5.1. A two-sample t-test showed significant differences between expert and novice viewing times ($t(8) = 1.96, p = 0.05$ where the t value for 8 degrees of freedom was 1.96, with a p value of 0.05) and times to first ROI fixations ($t(8) = 2.07, p = 0.04$) but failed to show significant differences between expert and novice fixation lengths ($t(8) = 0.25, p = 0.80$) and fixations per image ($t(8) = 1.06, p = 0.29$). As shown in Table 5.2 and Figure 5.7, the anatomical landmark most frequently fixated on by expert and novice observers was the stomach bubble, followed by the umbilical vein and spine. The majority of fixations (99% for experts and 98.9% for novices) fell within the abdominal wall. Despite fixating on similar ROIs, novices fixated on the spine to a greater extent than experts. A two-sample t-test failed to show significant differences between expert and novice fixations on the stomach bubble ($t(8) = 0.82, p = 0.41$) or umbilical vein ($t(8) = 1.90, p = 0.06$) but demonstrated that novices fixated to a significantly greater extent on the spine than experts ($t(8) = 5.53, p < 0.01$).

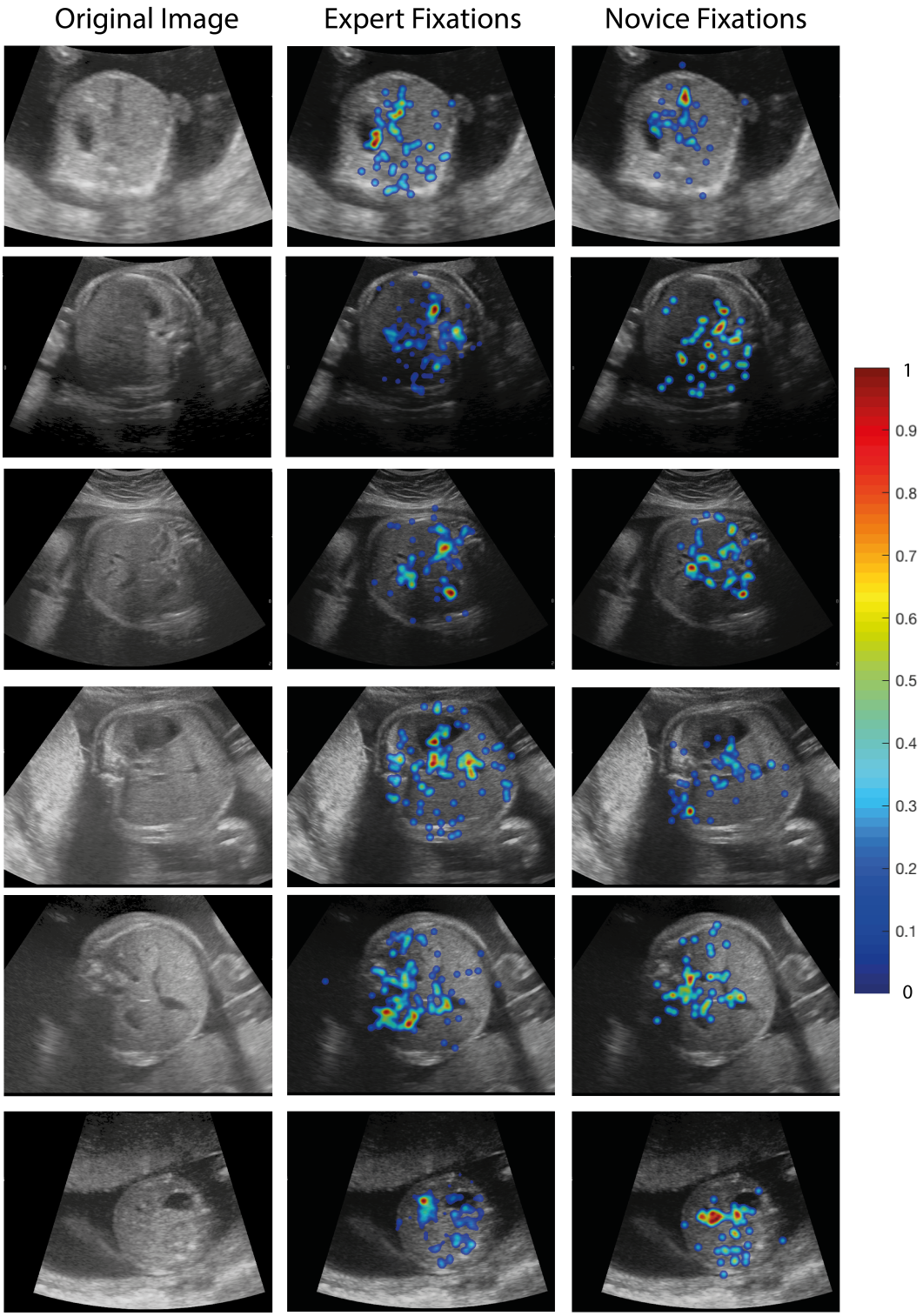


Figure 5.7: Examples of normalised fixation maps showing original stimulus images (left), all expert fixations (middle) and all novice fixations (right) where red signifies a higher number of fixations, and blue a lower number.

.

Group	Abdominal Cavity	Stomach Bubble	Umbilical Vein	Spine	Ribs	Background
Experts	59.08 ± 6.78	14.00 ± 8.94	15.60 ± 8.76	10.31 ± 2.11	0	1.01 ± 0.30
Novices	61.77 ± 8.01	12.10 ± 9.33	12.48 ± 9.41	11.42 ± 1.90	0	1.10 ± 0.44

Table 5.2: Mean percentages and standard deviations of fixation points falling within manually segmented bounding boxes around the abdominal cavity, stomach bubble, umbilical vein, spine, ribs, and background region for experts only, and novices only.

Predicted Group	Predicting Group	Static Consistency
Experts	Experts	0.76 ± 0.18
Experts	Novices	0.74 ± 0.21
Novices	Experts	0.73 ± 0.19
Novices	Novices	0.72 ± 0.23
Random	Random	0.63 ± 0.11

Table 5.3: Mean static consistency scores with standard deviations for sub-groups of observers: expert fixation maps predicted by experts only, expert fixation maps predicted by novices only, novice fixation maps predicted by experts only, and novice fixation maps predicted by novices only. The random cross-image baseline measure is also shown.

Static Consistency

Static consistency scores, as shown in Table 5.3 and Figure 5.8, were highest for expert fixation maps predicted by other experts only, suggesting that experts' fixated locations were more consistent than those of novices. Static consistency scores for all observer sub-groups were higher than the random baseline measure.

Dynamic Consistency

As shown in Table 5.4, the most common fixation sequences of both expert and novice observers involved cross-referencing between the stomach bubble and umbilical vein, with longer fixation sequences also involving cross-referencing with the spine. Dynamic consistency scores (Table 5.5) were highest for expert fixation sequences predicted by other experts only, suggesting that expert sequences were more consistent than those of novices. However, dynamic consistency scores for all observer sub-groups were higher than the random baseline measure.

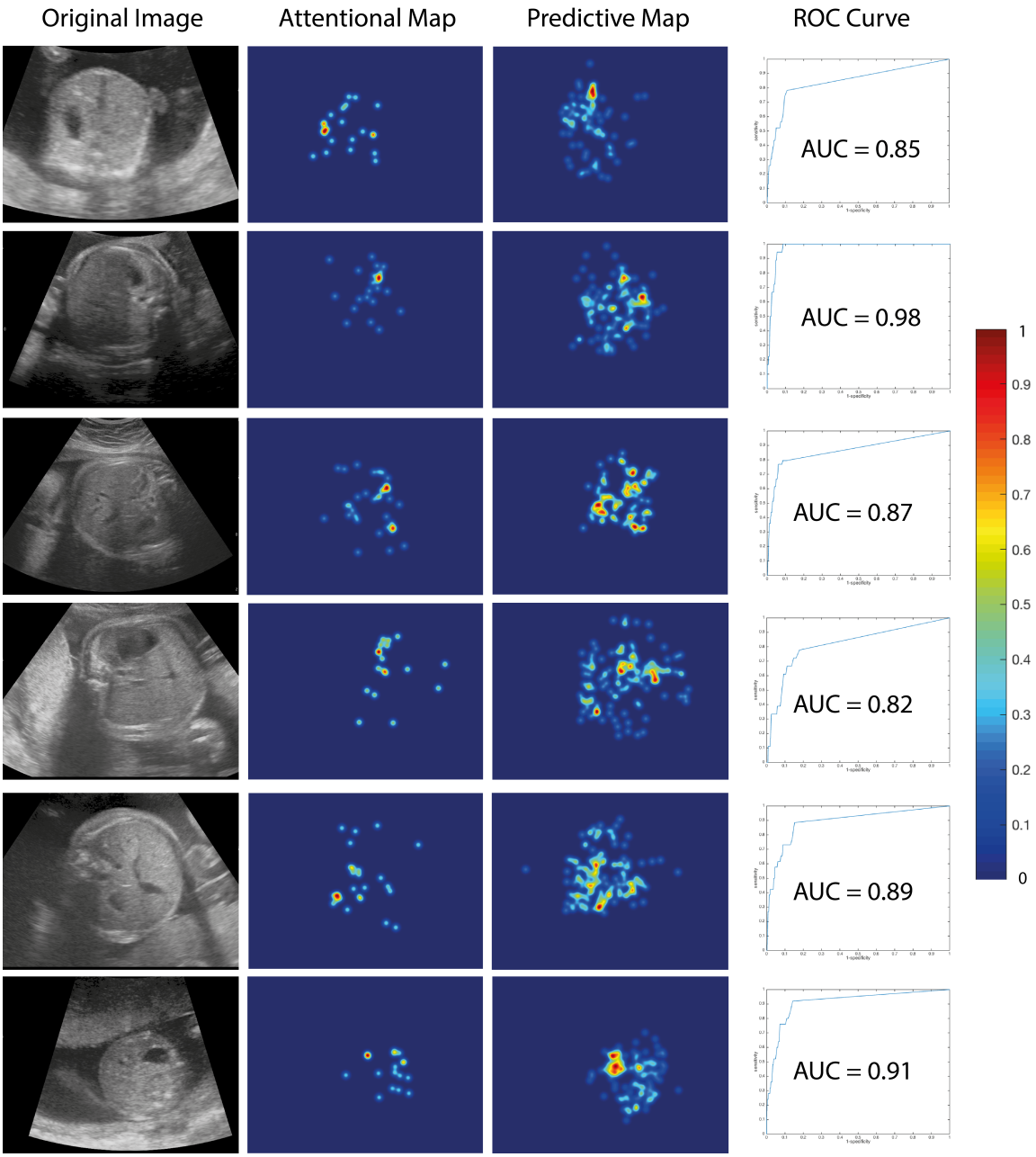


Figure 5.8: Stimulus images, individual attentional maps, leave-one-out summed predictive fixation maps and the resulting ROC curves for one expert observer's fixations predicted by all other observers. The static consistency score is given by the ROC curve AUC.

.

	Most Common Sequence of 3 Fixations		Most Common Sequence of 4 Fixations	
	1^{st}	2^{nd}	1^{st}	2^{nd}
Experts	$1-2-1$ (24%)	$2-1-2$ (19%)	$1-2-1-3$ (15%)	$2-1-2-3$ (14%)
Novices	$2-1-2$ (24%)	$1-2-1$ (22%)	$2-1-2-3$ (17%)	$1-2-1-3$ (13%)

Table 5.4: The 1^{st} and 2^{nd} most common sequences of 3 and 4 fixations for expert and novice observers, where ROI 1 denotes the stomach bubble, 2 denotes the umbilical vein and 3 the spine.

Predicted Group	**Predicting Group**	**Dynamic Consistency**
Experts	Experts	0.39 ± 0.11
Experts	Novices	0.38 ± 0.12
Novices	Experts	0.37 ± 0.10
Novices	Novices	0.37 ± 0.15
Random	Random	0.21 ± 0.18

Table 5.5: Mean dynamic consistency scores with standard deviations for sub-groups of observers: expert fixation sequences predicted by experts only, expert fixation sequences predicted by novices only, novice fixation sequences predicted by experts only, and novice fixation sequences predicted by novices only. The random cross-image baseline measure is also shown.

Global-Focal Search

Relative entropies between two-component GMMs fitted to expert gaze trajectories (Figure 5.9 and Table 5.6) were the lowest, showing a higher degree of similarity between expert gaze trajectories. Relative entropies between expert and novice groups were the highest, suggesting a low degree of similarity between expert and novice gaze trajectories. Relative entropies for all sub-groups were lower than the random baseline. The reduced chi-squared statistic (Table 5.7) for two-component GMMs exhibits $\chi^2_{red} \approx 1$ suggesting that this model best fitted the data. One and three-component GMMs displayed a poorer fit to the observed data, exhibiting $\chi^2_{red} > 1$. Four and five-component GMMs displayed $\chi^2_{red} < 1$ suggesting the model was over fitted to the data.

P Distribution	Q Distribution	Relative Entropy
Experts	Experts	35.4 ± 8.3
Experts	Novices	42.5 ± 6.5
Novices	Experts	39.3 ± 11.8
Novices	Novices	38.8 ± 10.5
Random	Random	200.6 ± 38.6

Table 5.6: Mean relative entropies with standard deviations of two-component GMMs representing the gaze trajectories of sub-groups of observers. Relative entropies are given for experts with respect to experts only, experts with respect to novices, novices with respect to experts, and novices with respect to novices only. A baseline measure of relative entropy calculated between randomly generated two-component GMMs is also shown. A lower entropy suggests a higher degree of similarity.

Gaussian Components	Reduced Chi-Squared Statistic
1	4.96
2	1.18
3	3.03
4	0.78
5	0.92

Table 5.7: Reduced chi-squared statistics for gaze trajectories represented by one, two, three, four and five-component GMMs, representing the goodness-of-fit of GMMs to observed data.

5.3.3 Discussion

As expected, ROI analysis showed that both novice and expert observers fixated on the stomach bubble and umbilical vein to a greater extent than any other anatomical landmark. This is perhaps not unexpected as they appear in ISUOG guidelines on standardised plane acquisition. However, the spine does not appear in ISUOG guidelines but was fixated on by expert observers to a greater extent than by novices, suggesting that the spine was used by more experienced observers as a reference point against which the positions of other anatomical landmarks were verified, given the consistent geometric relationship between the stomach bubble, umbilical vein and spine. In addition to making greater use of the spine as a reference point, experts exhibited more efficient search strategies than novices,

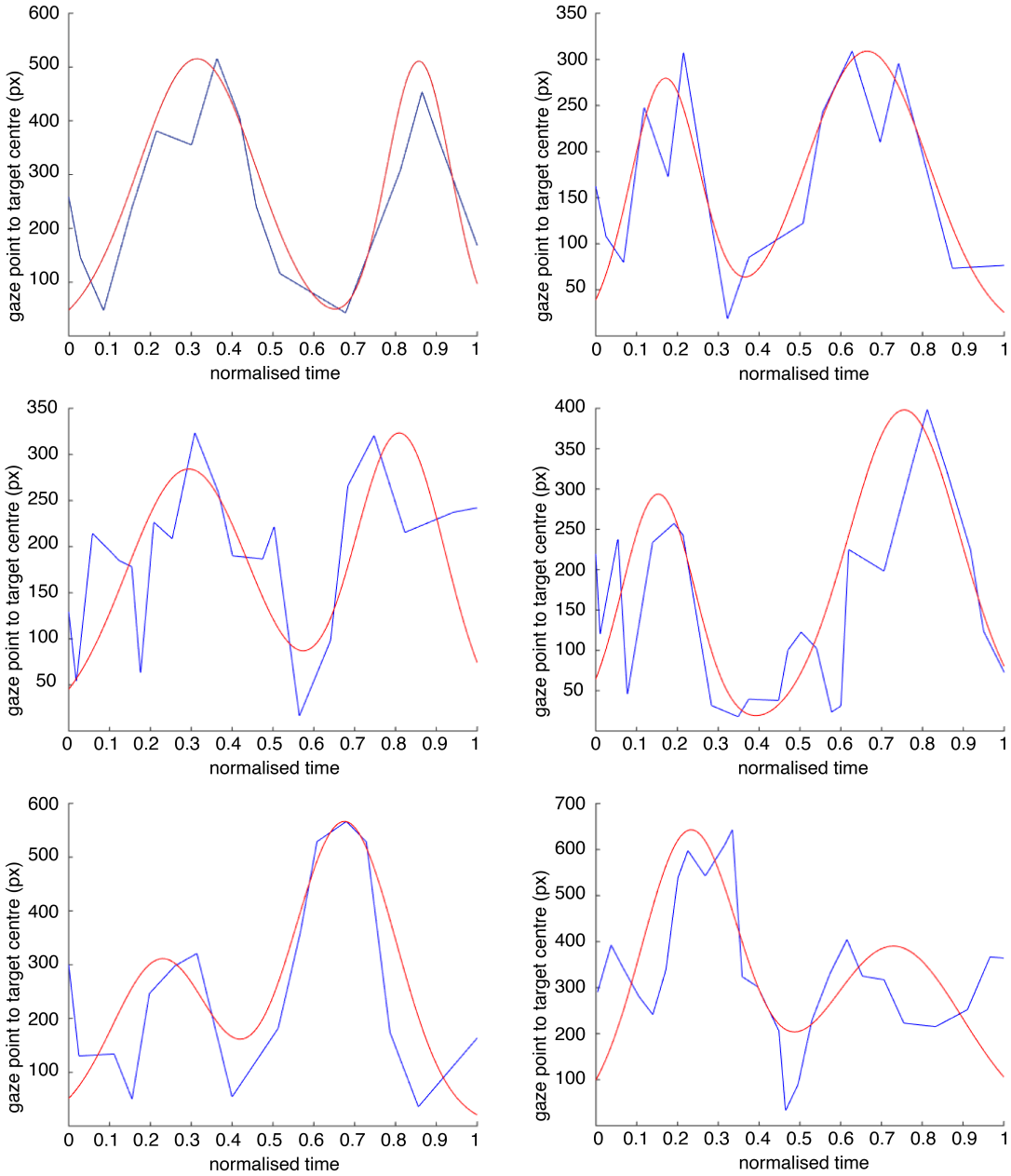


Figure 5.9: Two-component GMMs (red) fitted to the gaze trajectories (blue) of one expert observer viewing different stimulus images, where the gaze trajectory is plotted as the Euclidean distance between the gaze point and the centre of the stomach bubble bounding box against time

.

achieving a similar number of fixations per image than novices whilst spending on average less time viewing each image. Expert observers also appeared to be more skilled at identifying anatomically meaningful image regions than novices, displaying significantly lower times to first fixations on ROIs.

The similar values of static consistency for all observer sub-groups suggest that observers fixated largely on similar image regions, after accounting for oculomotor bias as demonstrated by the lower cross image baseline measure of static consistency. Fixated locations of experts were more consistent than those of novices, suggesting that the visual search strategy of the latter group was more disparate as novices were perhaps less familiar with the expected appearances and locations of target anatomical landmarks.

Both expert and novice observers displayed a high degree of cross-referencing between the stomach bubble, umbilical vein and spine with the most common fixation sequences characterised by repeated fixations on the stomach bubble and umbilical vein followed by a final fixation on the spine, reinforcing the role of the spine as a reference point and the importance of the geometric relationship between anatomical landmarks. The similar values of dynamic consistency for all observer sub-groups suggest that observers fixated on ROIs in similar sequences, with expert fixation sequences displaying a slightly higher degree of consistency than those of novices.

Goodness-of-fit analysis of varying component GMMs fitted to the gaze trajectories of observers demonstrated that two-component GMMs most accurately modelled the collected gaze data, validating Kundel et al.'s[10] discovery-reflective visual search hypothesis with respect to US images. Compared to the random baseline measure, the low relative entropies computed between two-component GMMs fitted to the gaze trajectories of observers demonstrated a high degree of consistency between the visual search strategies of observers. Relative entropies calculated between expert GMMs were lower than those calculated between novice GMMs, suggesting that the former group employed a more consistent two-component visual search strategy, in line with the findings of previous experiments on static and dynamic consistency.

Eye Tracking Result	Detector Design Decision
Two-phase search strategy displayed by observers	Two-phase detector consisting of sliding window detector followed by pictorial structures model
Observers fixated on the spine to almost the same degree as the stomach and umbilical vein, high degree of cross-referencing between these landmarks	Pictorial structures model was trained on the relative positions of the spine, stomach bubble and umbilical vein
Majority of fixations fell within the abdominal wall	Initial abdomen detection stage and subsequent masking of candidate bounding boxes falling outside this region

Table 5.8: Key findings of the eye tracking experiments outlined in Section 5.3, and corresponding decisions on the design of a detector for the automated localisation of the stomach bubble and umbilical vein.

5.4 Automated Landmark Localisation

The findings in Section 5.3 were harnessed to implement a framework for the automated detection of the stomach bubble and umbilical vein, incorporating the high-level constraints and visual search strategies employed by observers (Table 5.8). Specifically a two-stage detector was implemented, inspired by the two-component visual search strategy validated in Section 5.3.2. A sliding window detector consisting of a boosted set of decision stumps mimicked the 'discovery' phase of visual search by detecting many possible candidates for the abdomen, stomach bubble, umbilical vein and spine. A pictorial structures model mimicked the 'reflective' phase of visual search by finding the optimal configuration of candidate landmarks using prior knowledge of their geometry. The detector's use of the spine was inspired directly by the findings in Section 5.3.2 whereby the geometric relationship between the stomach bubble, umbilical vein and spine was harnessed by observers.

This framework was based on the work of Rahmatullah[7] with the addition of pyramid representations of image patches for efficient feature extraction, the arrangement of increasingly complex classifiers in a cascade for early rejection of negative image windows, and a pictorial structures model for constraining the relative positions of candidate bounding boxes.

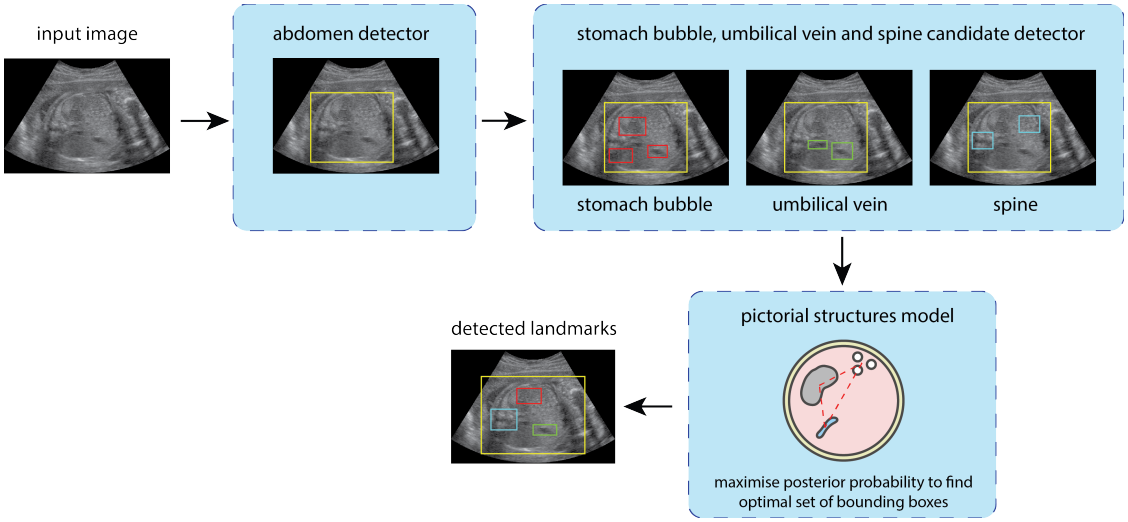


Figure 5.10: Schematic overview of a two-stage stomach bubble and umbilical vein detector, where the first stage identifies candidate bounding boxes for the stomach bubble and umbilical vein, and the second stage finds the optimal set of bounding boxes which maximise a posterior probability distribution based on candidate confidence scores and positions.

5.4.1 Methods

Training Data

The framework was trained using $I_{B\,n}$, the dataset described in Chapter 4 with corresponding ground truth bounding boxes $I^*_{B\,m,n}$ where $n = 1, ..., 1000$ and $m = 1, ..., 4$ for the abdomen, stomach bubble, umbilical vein and spine respectively.

Square image patches based on these bounding boxes were extracted with $10px$ padding in all directions, resulting in a set of positive training patches which were resized to $200 \times 200px$. 1000 square bounding boxes sampled randomly from image regions not containing positive training patches were also extracted with $10px$ padding, resulting in a set of negative training patches, also resized to $200 \times 200px$, with no overlap between positive and negative training patches. Rectangular ground truth bounding boxes necessarily contained portions of the image not showing structures of interest, therefore limiting the image regions eligible for selection as negative training patches. Annotation with elliptical boundaries would mitigate this; however a sufficient number of negative training patches were extracted from

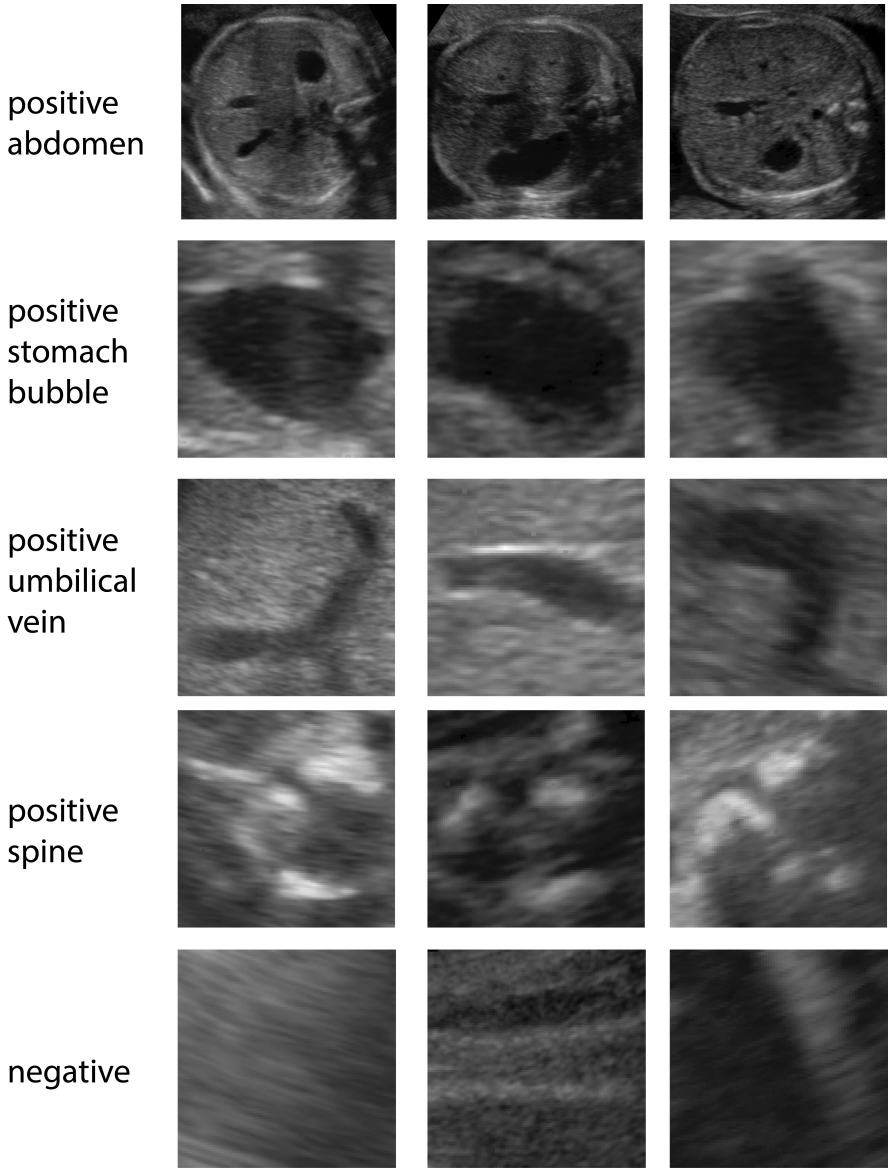


Figure 5.11: Examples of positive training patches showing the abdomen, stomach bubble, umbilical vein and spine, and negative training patches showing no anatomical landmarks.

.

the set of training images. Additionally, the work of Rahmatullah et al.[24] on which this framework is based and which is used as a benchmark, employs rectangular ground truth bounding boxes and square training patches.

Feature Extraction

The local descriptors used to train the automated framework were based on those employed by Rahmatullah et al.[24], with the addition of pyramid representation

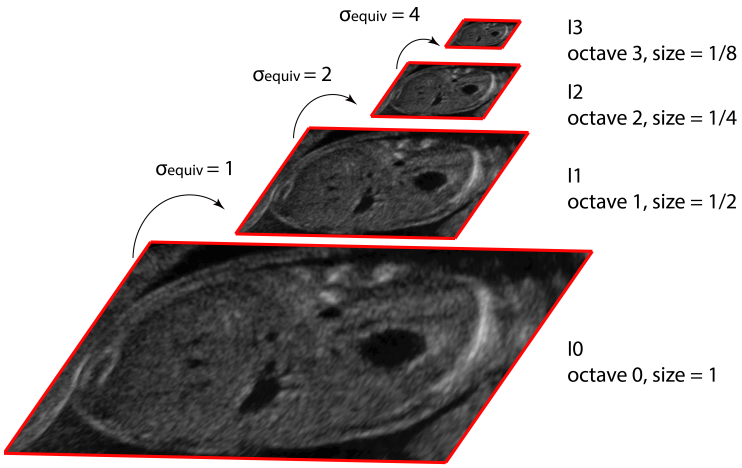


Figure 5.12: Examples of a three-octave Gaussian image pyramid constructed from a positive training patch showing the abdomen. Each octave reduces the image size by a factor of 2, and the effective Gaussian smoothing kernels at each octave, σ_{eff}, are $1, 2$ and 4 respectively. This results in the images $I0$, $I1$, $I2$ and $I3$.

to allow the efficient computation of features at mutiple scales as demonstrated by Dollar et al.[118].

Pyramid Representation Features were extracted from the positive and negative training patches in a Gaussian pyramid (Figure 5.12). An image pyramid was constructed for each training patch by convolution with a Gaussian kernel of $\sigma = 1px$ and subsequent downsampling by a factor of two. This was repeated to a minimum size of $16 \times 16px$, with eight scales per octave corresponding to a downsampling factor of $2^{-\frac{1}{8}} \approx 0.917$ for each octave or level of the image pyramid.

Gradient Magnitude and Gradient Orientation Gradient magnitudes ($G = \sqrt{I_x^2 + I_y^2}$, where I_x and I_y are the horizontal and vertical components of the image gradient at a particular pixel) for the training patches were computed from the image pyramids. Histograms of oriented gradients were also computed from the image pyramids; gradient orientations ($\theta = tan^{-1}\frac{I_x}{I_y}$) within $4 \times 4px$ blocks were quantised into 8 bins, giving the distributions of gradients at different orientations around each pixel within the training patches. Each $4 \times 4px$ block was therefore encoded by a final HoG descriptor consisting of a 1×8 feature vector giving the relative weights of gradients at each of the eight quantised orientations.

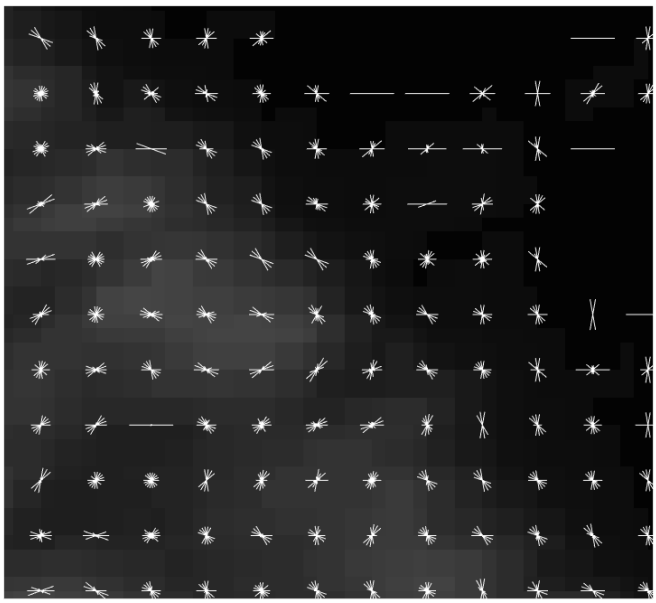


Figure 5.13: HoG descriptors extracted from $4 \times 4px$ cells in a positive abdomen training patch, where gradients are quantised into 8 bins.

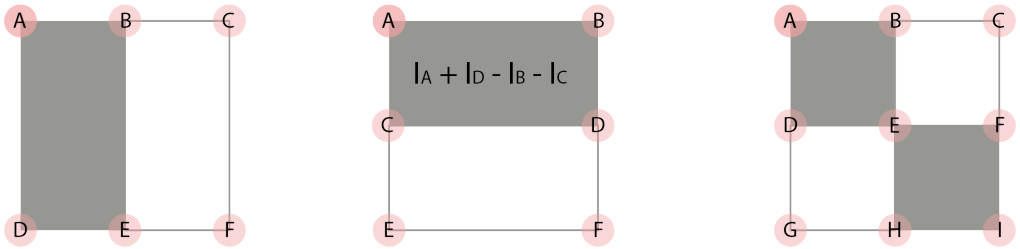


Figure 5.14: Haar-like features extracted from integral representations of image pyramids, constructed from positive and negative training patches.

Haar-Like Features Haar-like features were computed from integral image representations ($I_\Sigma(x, y) = \Sigma_{x' \leq x\, y' \leq y} I(x', y')$) of the image pyramids, where the final descriptor value for each of the $4 \times 4px$ Haar-like features in Figure 5.14 was the difference between summed pixel values underlying the dark and light windows. Integral image representation reduced the computation of summed pixel values within a rectangle to four operations as shown in Figure 5.14 where I_A, I_B, I_C, I_D are integral image values at rectangle vertices.

Boosted Decision Stumps

An AdaBoost framework as described by Viola et al.[80] was used to independently train four classifiers, each consisting of a set of boosted decision stumps, to localise candidate bounding boxes for the abdomen, stomach bubble, umbilical vein and spine. Each weak classifier in an ensemble was a single decision stump thresholding on a single feature to perform the binary classification task of separating positive and negative training samples for the anatomical landmark (the abdomen, stomach bubble, umbilical vein or spine) being detected as in $h_j(x) = \{^{1\ if\ f_j(x)>\theta_j}_{0\ otherwise}$, where $f_j(x)$ is the j^{th} feature extracted from some training patch x, and θ_j is the associated threshold, resulting in a weak classifier output $h_j(x)$.

The weak classifiers were trained in sequence, with each being trained on a weighted form of the training data, with training datapoints that were misclassified by previous weak classifiers assigned a higher weight when training subsequent weak learners. Given i training points, of which l are positive and m are negative, consisting of image patches x and binary labels y, points are assigned equal starting weights of $w_i = \frac{1}{2m}$ for negative training patches and $w_i = \frac{1}{2l}$ for positive patches.

Each weak classifier t in a set of T was trained sequentially, finding a threshold value of a single feature $h_j(x_i)$ which minimised the error function $\varepsilon_j = \Sigma_i w_i |h_j(x_i) - y_i|$.

Weights were then updated for each data point to give $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$, where $e_i = 1$ for points that were correctly classified and $e_i = 0$ for those incorrectly classified by the previous weak learner, and $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$.

The final strong classifier, after all T weak learners had been trained, was a weighted combination of decision stumps as given by $h(x) = \{^{1\ \Sigma_{t=1}^{T}\alpha_t h_t(x) \geq \frac{1}{2}\Sigma_{t=1}^{T}\alpha_t}_{0\ otherwise}$, where $\alpha_t = log\frac{1}{\beta_t}$.

Each of the final trained detectors operated on a sliding window basis with a spatial stride of $4px$ between windows, and window dimensions $360 \times 360px$, $100 \times 100px$, $60 \times 60px$, and $50 \times 50px$ corresponding to the mean dimensions

Stages	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Accuracy	0.75	0.77	0.83	0.92	0.93	0.94	0.94

Table 5.9: The variation of abdominal wall detection accuracy with the number of cascade stages across a validation set, and defining a correct detection as one where the maximal confidence score bounding box has a Dice overlap coefficient greater than 0.75 with the manually labelled ground truth bounding box. The number of boosted decision stumps increased by a factor of four with each additional cascade stage, with each stage in the seven stage detector consisting of $32, 128, 512, 2048, 8192, 32768$ and 131072 boosted decision stumps respectively.

of the ground truth bounding boxes $I^*_{B\,1,n}$, $I^*_{B\,2,n}$, $I^*_{B\,3,n}$, $I^*_{B\,4,n}$ for the abdomen, stomach bubble, umbilical vein and spine respectively. Positive window instances were returned with a bounding box confidence score, computed as the difference between the weighted sums of positive and negative weak classifier votes, or margin, as a proportion of the total number of classifiers. This resulted in a set of detected candidate bounding boxes where the r^{th} ranking bounding boxes by confidence score on the n^{th} testing image were described by $DT_{\mathrm{ab}\,r,n}$ $DT_{\mathrm{st}\,r,n}$ $DT_{\mathrm{uv}\,r,n}$ and $DT_{\mathrm{sp}\,r,n}$ for the abdomen, stomach bubble, umbilical vein and spine respectively.

A series of ensemble classifiers, increasing in complexity, were arranged in a cascade to rapidly focus computational attention on image windows most likely to contain anatomical landmarks, and to discard negative sub-windows at an early stage.

Validation over the dataset $I_{C\,n}$ described in Chapter 4 was used to determine the optimal number of cascade stages. The number of stages was selected based on abdominal wall detection accuracy, where a correct detection was one in which the Dice overlap coefficient ($Dice = \frac{2(GT \cap DT)}{GT+DT}$, where GT and DT are the sets of pixels defined by the ground truth and detected bounding boxes respectively) between the maximal confidence score bounding box ($DT_{\mathrm{ab}\,1,n}$) and the ground truth bounding box ($I^*_{C\,1,n}$) was greater than 0.75, the same threshold employed by Rahmatullah et al.[24]. As shown in Table 5.9, significant improvements in accuracy were not achieved beyond four cascade stages. Therefore the final configuration of the cascade consisted of four stages comprising $32, 128, 512$ and 2048 boosted decision stumps.

A target false negative rate was set for each stage of the cascade, with lower target false negative rates for earlier stages to ensure the majority of negative

windows were correctly rejected. Threshold confidence scores of $0.07, 0.11, 0.13$ and 0.21 were set for each stage of the cascade through validation over $I_{C\,n}$, to achieve false negative rates of $4\%, 6\%, 8\%$ and 10% respectively for abdominal wall detection.

Testing Data

The detectors were tested over $I_{D\,n}$, the dataset described in Chapter 4. For each testing image in $I_{D\,n}$ non-maximal suppression was used to select $DT_{\mathrm{ab}\,1,n}$, the detected abdominal bounding box with the highest confidence score. To implement abdominal masking the stomach bubble, umbilical vein and spine bounding boxes falling within the maximal confidence abdominal bounding box (for which $DT_{\mathrm{st}\,r,n} \in DT_{\mathrm{ab}\,1,n}$, $DT_{\mathrm{uv}\,r,n} \in DT_{\mathrm{ab}\,1,n}$, $DT_{\mathrm{sp}\,r,n} \in DT_{\mathrm{ab}\,1,n}$) were retained and all others discarded.

A correct anatomical landmark detection was defined as a detected bounding box with a Dice overlap coefficient greater than 0.75 (the threshold employed by Rahmatullah et al.[6]) with the ground truth bounding box. Sensitivity and specificity were calculated at varying threshold levels for bounding box confidence scores, and a ROC curve generated for each detector across all $I_{D\,n}$. Standalone accuracies for the abdominal wall, stomach bubble, umbilical vein and spine detectors were defined as the proportion of testing images where the maximal confidence score bounding boxes $DT_{\mathrm{ab}\,1,n}, DT_{\mathrm{st}\,1,n}, DT_{\mathrm{uv}\,1,n}, DT_{\mathrm{sp}\,1,n}$ respectively displayed a Dice coefficient greater than 0.75 with the ground truth bounding boxes $I^{*}_{D\,m,n}$.

Pictorial Structures Model

The final stage of the detector computed the optimal configuration of candidate bounding boxes for the stomach bubble, umbilical vein and spine by modelling these landmarks as rigid parts connected by deformable springs as described by Fischler et al.[11]. The statistical framework described by Felzenszwalb et al.[119] was used to determine the optimal configuration of parts by maximising the posterior probability based on individual part appearances (represented by detected bounding

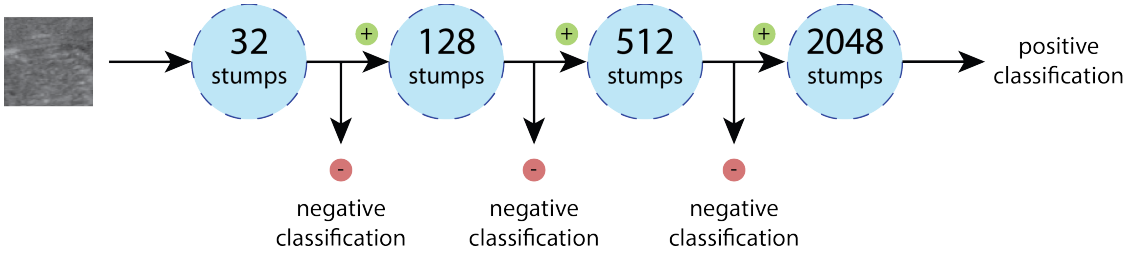
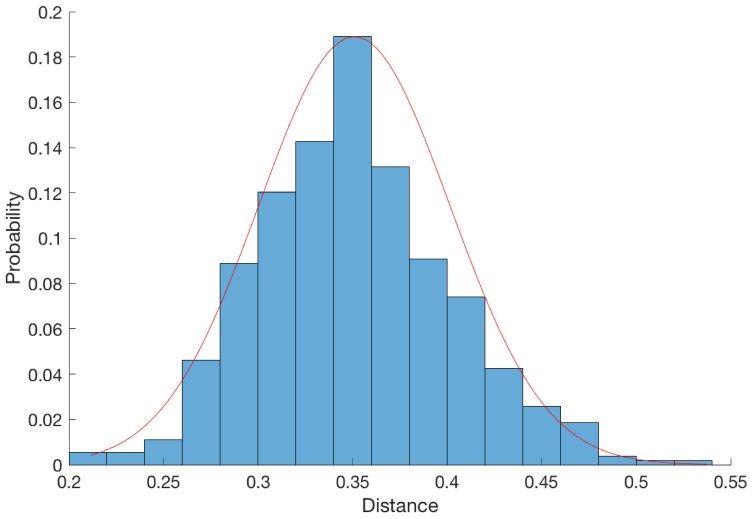


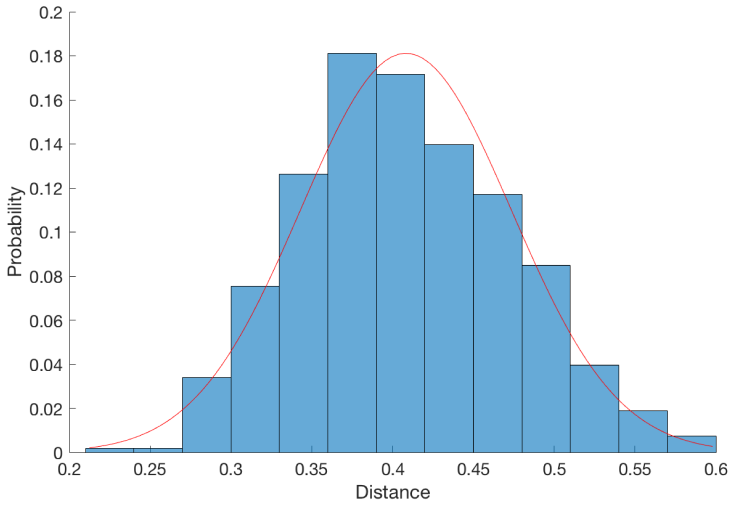
Figure 5.15: A cascade of four sets boosted decision stumps, where each stage of the cascade is a ensemble classifier trained using the AdaBoost algorithm. Any sub-window classified as negative by any stage of the cascade is immediately classified as negative, whereas a sub-window must pass through every stage of the cascade in order to be classified as positive; this process quickly focuses computational attention on image sub-regions that are likely to contain anatomical landmarks.

box confidence scores) and positions (represented by the Euclidean distances between pairs of parts as a proportion of abdominal width). Specifically, $p(L|I,\theta) \propto \Pi_{i=1}^{n} p(I|l_i,u_i)\Pi_{v_i,v_j} p(l_i,l_j|c_{ij})$ gives the posterior probability $p(L|I,\theta)$ of a particular configuration of bounding boxes matching the ground truth, as a function of confidence scores and how closely the relative locations of the parts adhered to the geometric model. Here, L is a particular configuration of parts, I is an image, θ are model parameters, l_i is a part at position i with a corresponding match score u_i, and c_{ij} is a cost function based on the distance between two parts at positions i and j.

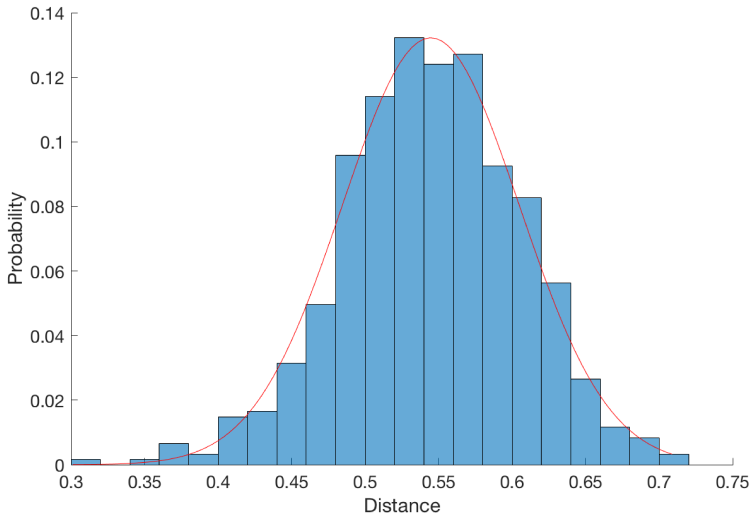
These probability distributions were constructed using dataset $I_{C\,n}$ as described in Chapter 4. $p(I|l_i,u_i)$ was treated as a Gaussian distribution modelling the probability of correctly detected bounding boxes for the stomach bubble, umbilical vein and spine across varying confidence scores (Figure 5.16). Here, the confidence scores of correctly detected bounding boxes were obtained for each of these landmarks using the detectors described in Section 5.4.1. $p(l_i,l_j|c_{ij})$ was treated as a Gaussian distribution modelling the joint probabilities of the Euclidean distances between ground truth bounding box centres for the stomach and umbilical vein, stomach and spine, and spine and umbilical vein as proportions of ground truth abdominal bounding box width (Figure 5.16). The model was tested over $I_{D\,n}$, the dataset described in Chapter 4.



(a)



(b)



(c)

Figure 5.16: Gaussian distributions modelling the joint probabilities of Euclidean distances as a proportion of abdominal width of the (a) Stomach bubble and umbilical vein, (b) Stomach bubble and spine, (c) Spine and umbilical vein.

.

5.4.2 Results

Standalone detection accuracies with no geometric priors for the abdomen, stomach bubble, umbilical vein and spine using non-maximal suppression of bounding box confidence scores are shown in Table 5.10 (where a Dice coefficient greater than 0.75 between the maximal confidence score bounding box and the ground truth bounding box constituted a correct detection). The associated ROC curves are shown in Figure 5.23 and individual bounding box Dice coefficients and confidence scores are shown in Figure 5.24. Examples of correct and incorrect detections are shown in Figures 5.17, 5.18, 5.19, and 5.20. Detection accuracies following abdominal masking and after the application of geometric constraints via the pictorial structures model are also shown in Table 5.10, alongside benchmark accuracies for stomach bubble and umbilical vein detection as reported by Rahmatullah et al.[6] (where a Dice coefficient greater than 0.75 between the geometrically optimal bounding box and the ground truth bounding box constituted a correct detection). The confusion matrix in Figure 5.25 demonstrates that of the predicted stomach bubble bounding boxes, all incorrect localisations were ground truth umbilical veins. However of all predicted umbilical vein bounding boxes, incorrect localisations consisted of both ground truth stomach bubbles and other errors including localising only a portion of the ground truth umbilical vein and misclassifying artefacts or blob-like anatomical structures as the umbilical vein. All incorrect predicted spine bounding boxes were attributable to other errors including the misclassification of bright structures including the ribs as the spine. The pictorial structures model displayed a mean run-time of $2.02s$ per image on a MacBook Pro $2.8GHz$ Intel Core i7 processor, with $16GB$ $1600MHz$ DDR3 RAM.

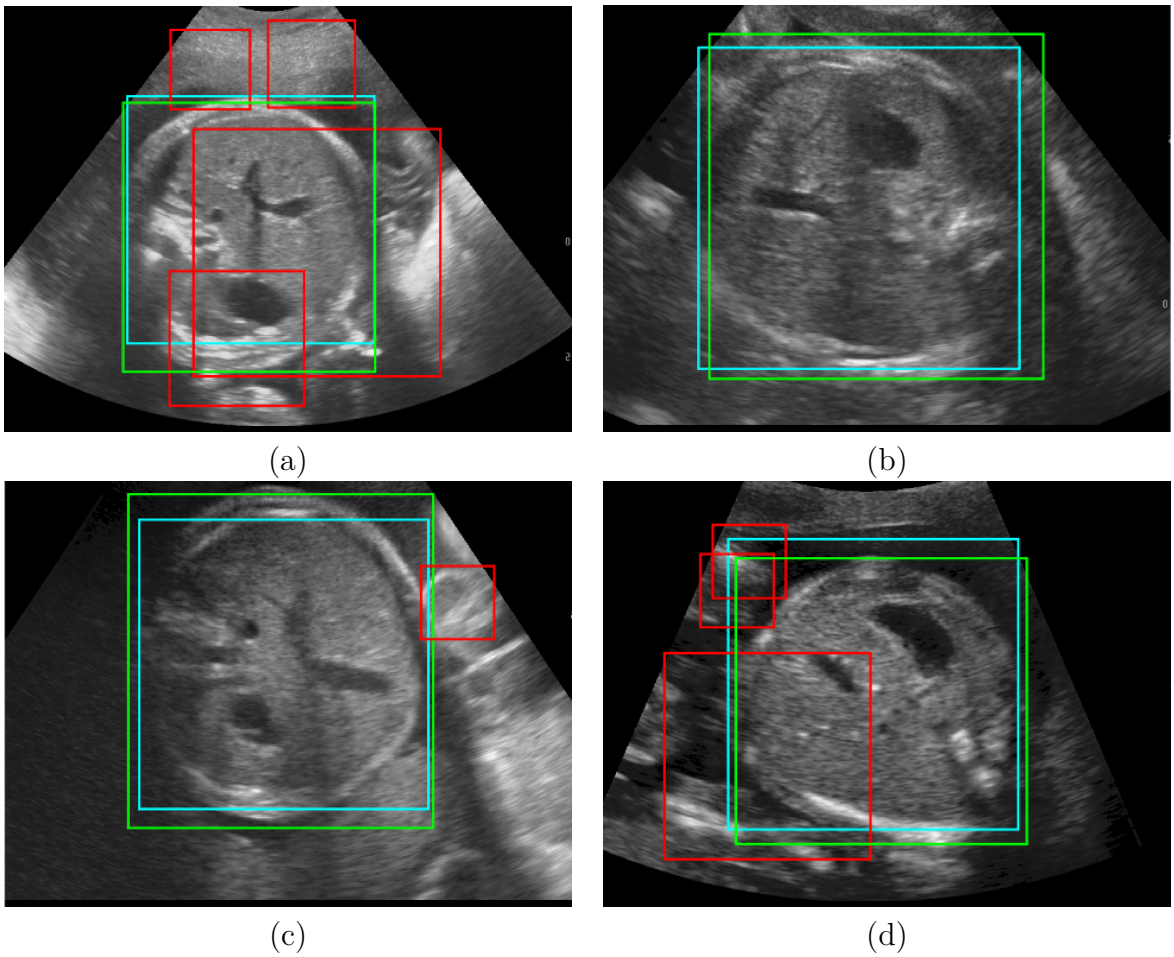


Figure 5.17: Correctly detected abdominal walls using non-maximal suppression of bounding box confidence scores. Ground truth bounding boxes (green) have a Dice overlap coefficient greater than 0.75 with the highest confidence ranked detected bounding boxes (blue), with lower ranked false positive bounding boxes (red) disregarded

.

Landmark	**Benchmark** (%)	**Standalone** (%)	**Abdomen Masking** (%)	**Pictorial Structures Model** (%)
Abdomen	-	92.40	-	-
Stomach bubble	78.94	75.60	83.60	87.20
Umbilical vein	62.80	63.60	73.20	83.20
Spine	-	60.80	70.40	71.60

Table 5.10: The accuracy in stomach bubble, umbilical vein and spine detection at each stage in the development of the anatomical landmark detector described above. Standalone detector accuracies with no geometric priors are given alongside standalone detector accuracies after abdominal masking, and after the application of geometric constraints through the pictorial structures model. Benchmark stomach bubble and umbilical vein accuracies as reported by Rahmatullah et al.[6] are also shown for reference.

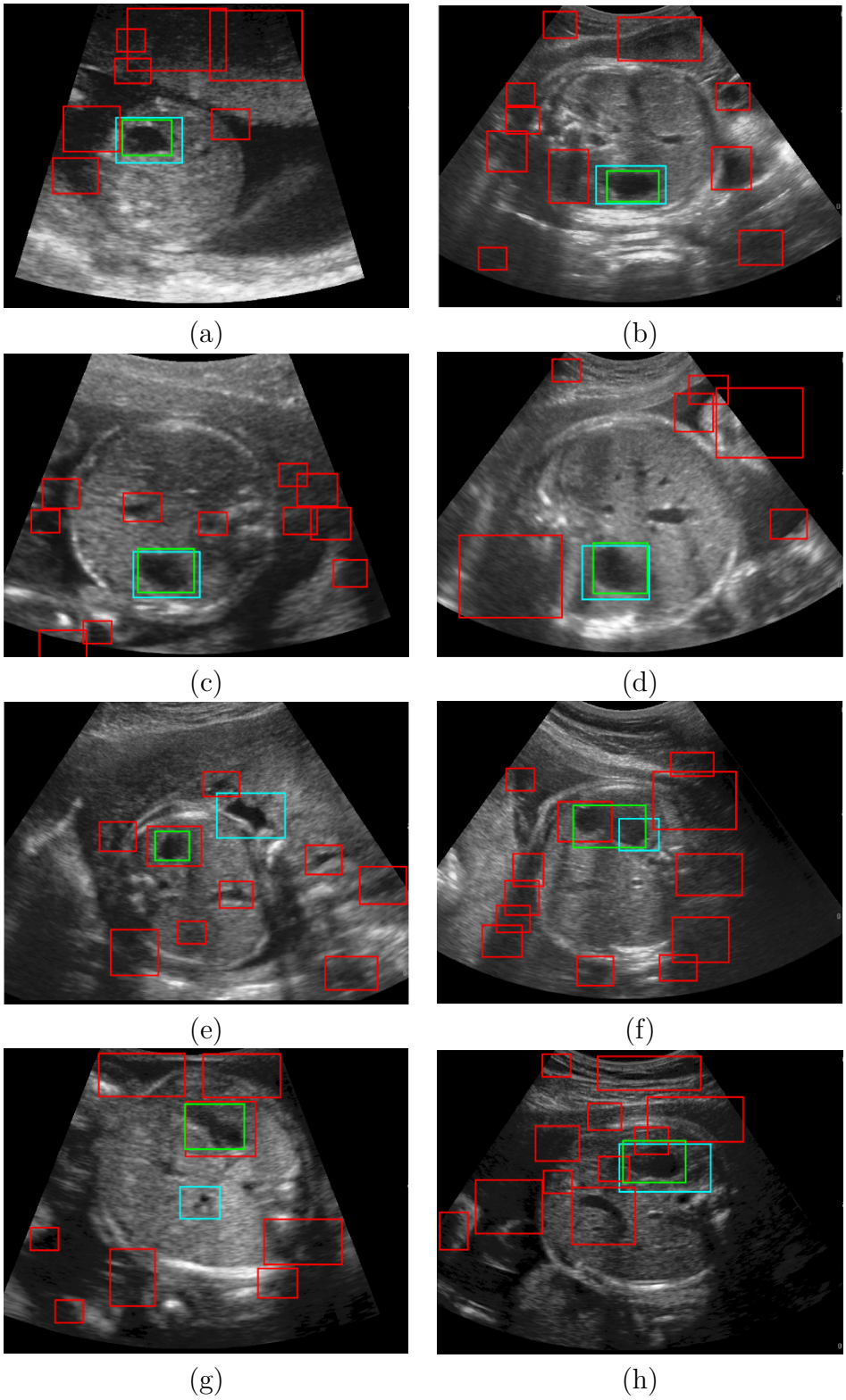


Figure 5.18: (a)-(d) Correctly and (e)-(h) Incorrectly detected stomach bubbles using non-maximal suppression of bounding box confidence scores and no geometric constraints. For correct detections, ground truth bounding boxes (green) have a Dice overlap coefficient greater than 0.75 with the highest confidence ranked detected bounding boxes (blue), with lower ranked false positive bounding boxes in red. For incorrect detections, the Dice coefficient between ground truth and highest ranked bounding boxes is less than 0.75.

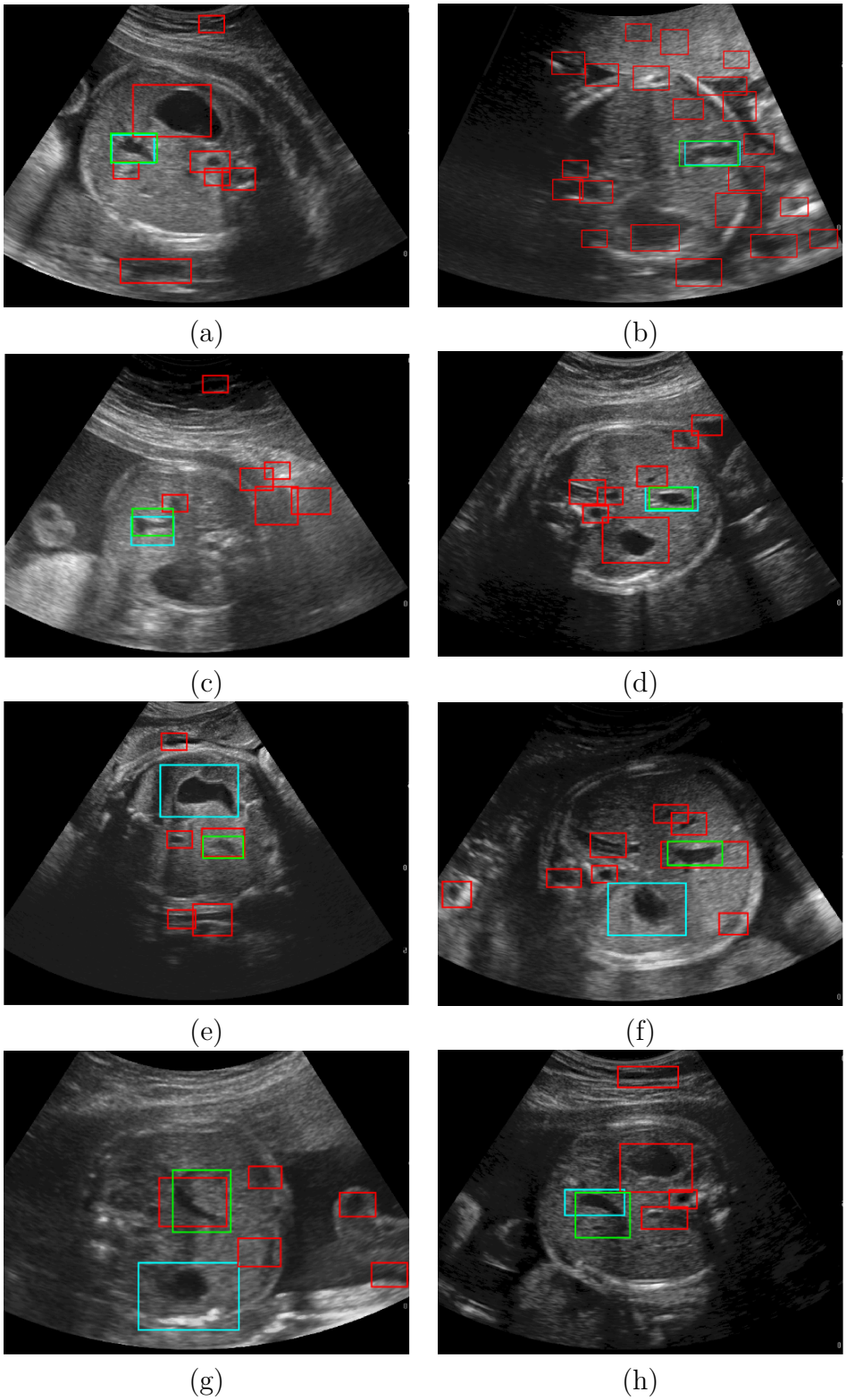


Figure 5.19: (a)-(d) Correctly and (e)-(h) Incorrectly detected umbilical veins using non-maximal suppression of bounding box confidence scores and no geometric constraints. For correct detections, ground truth bounding boxes (green) have a Dice overlap coefficient greater than 0.75 with the highest confidence ranked detected bounding boxes (blue), with lower ranked false positive bounding boxes in red. For incorrect detections, the Dice coefficient between ground truth and highest ranked bounding boxes is less than 0.75.

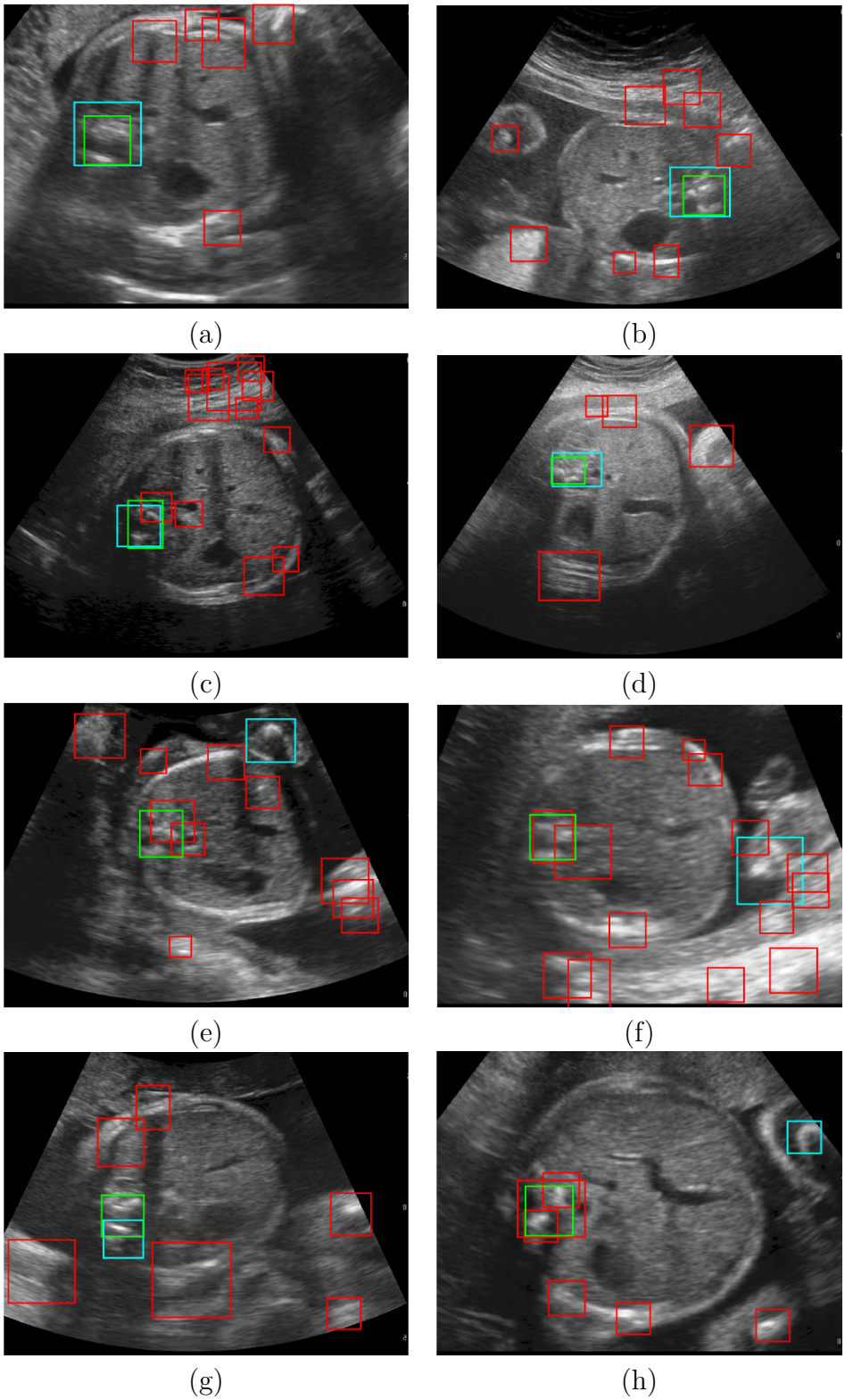


Figure 5.20: (a)-(d) Correctly and (e)-(h) Incorrectly detected spines using non-maximal suppression of bounding box confidence scores and no geometric constraints. For correct detections, ground truth bounding boxes (green) have a Dice overlap coefficient greater than 0.75 with the highest confidence ranked detected bounding boxes (blue), with lower ranked false positive bounding boxes in red. For incorrect detections, the Dice coefficient between ground truth and highest ranked bounding boxes is less than 0.75.

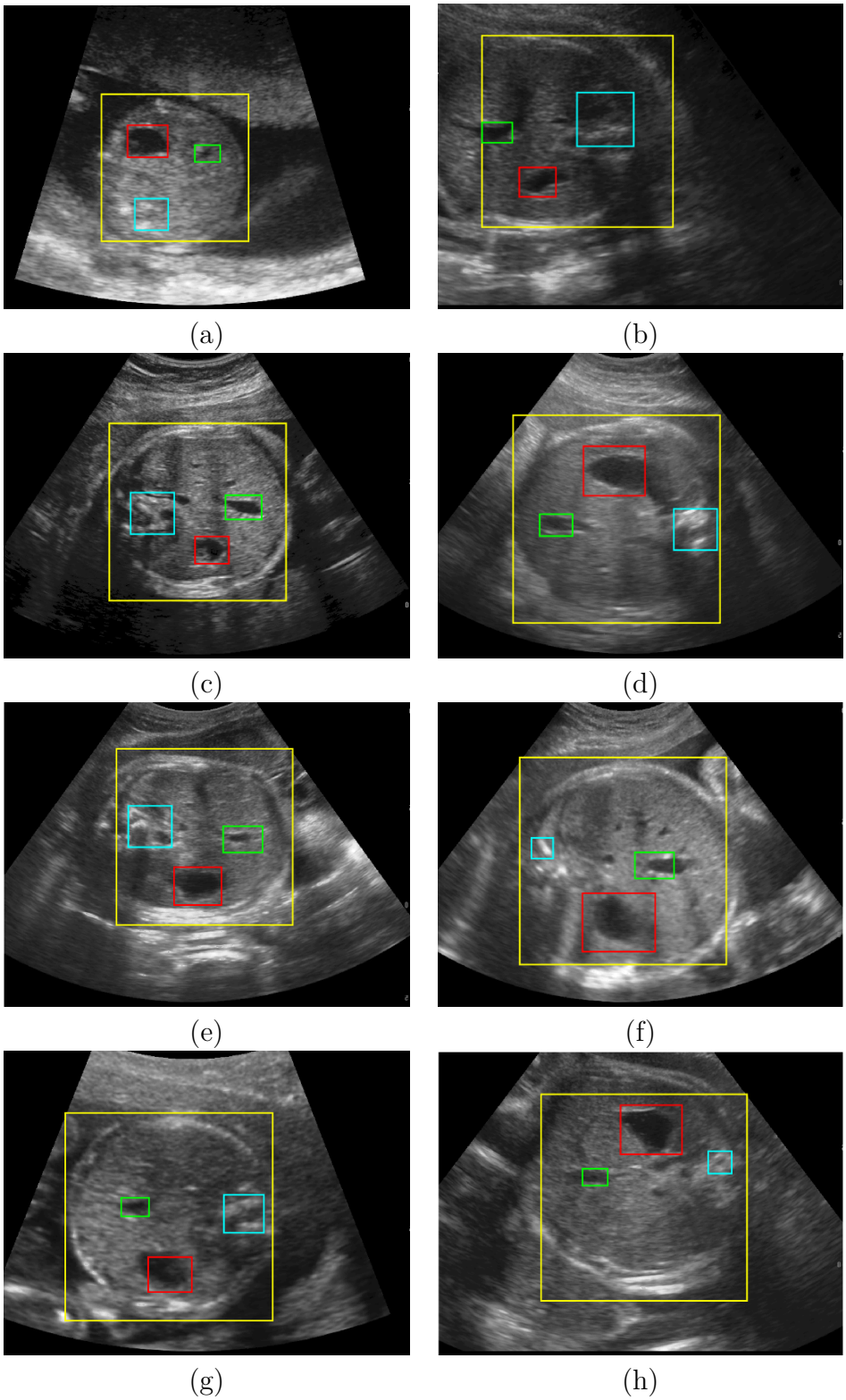


Figure 5.21: Correctly detected abdominal wall (yellow) stomach bubble (red), umbilical vein (green) and spine (blue) bounding boxes using the pictorial structures model, where the bounding boxes shown have a Dice overlap coefficient greater than 0.75 with the ground truth bounding boxes.

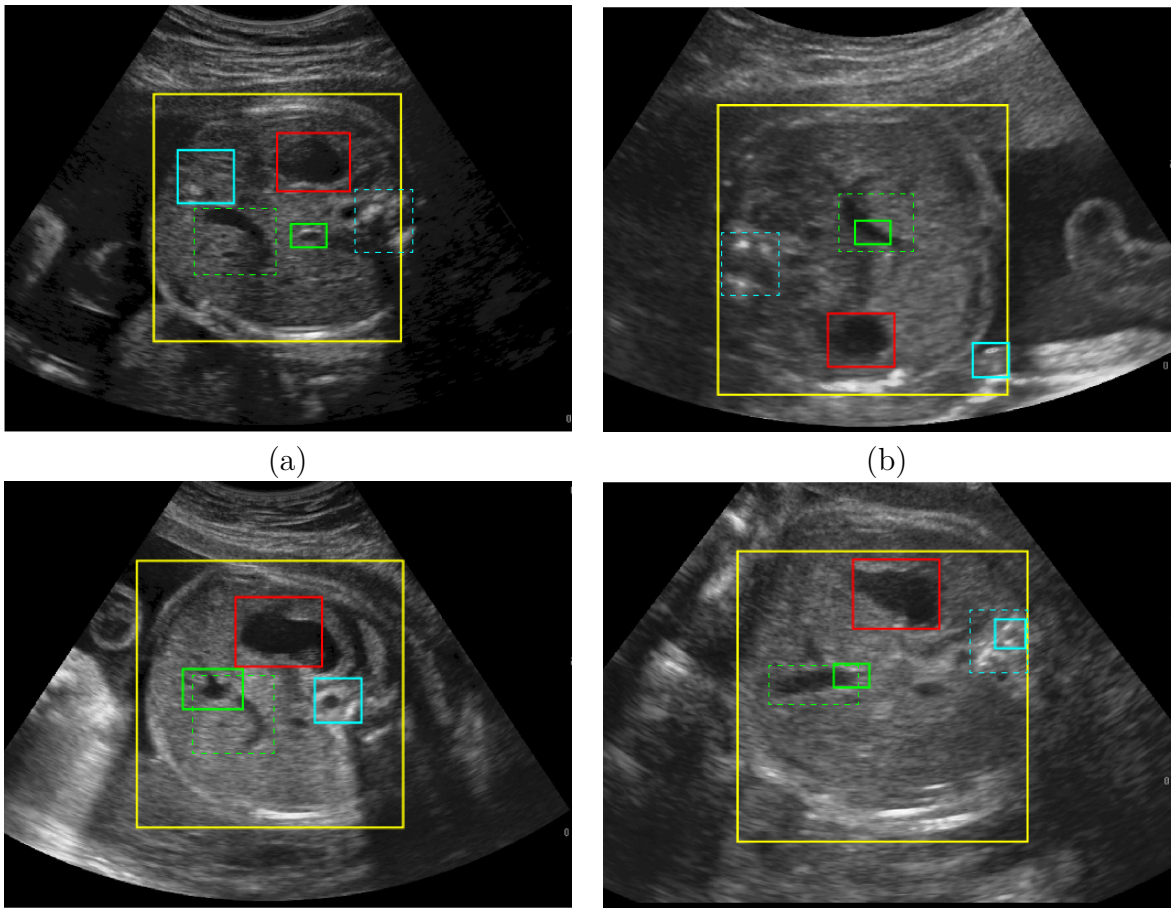


Figure 5.22: Incorrectly detected stomach bubbles (solid red), umbilical veins (solid green) and spines (solid blue) using the pictorial structures model described above, in particular (a) Incorrectly localised spine and umbilical vein (b) Incorrectly detected spine and partially detected umbilical vein (c) Partially detected umbilical vein (d) Partially detected umbilical vein and spine, where the incorrectly detected bounding boxes have a Dice overlap coefficient less than 0.75 with the ground truth bounding boxes (dashed).

.

5.4.3 Discussion

It has been demonstrated that the use of geometric priors obtained via eye tracking experiments, specifically abdominal masking and a pictorial structures model, lead to improved anatomical landmark detection accuracies in 2-D fetal abdominal US images compared to the benchmark method of Rahmatullah et al.[7], which does not harness geometric constraints.

For the stomach bubble, umbilical vein and spine, the greatest increase in

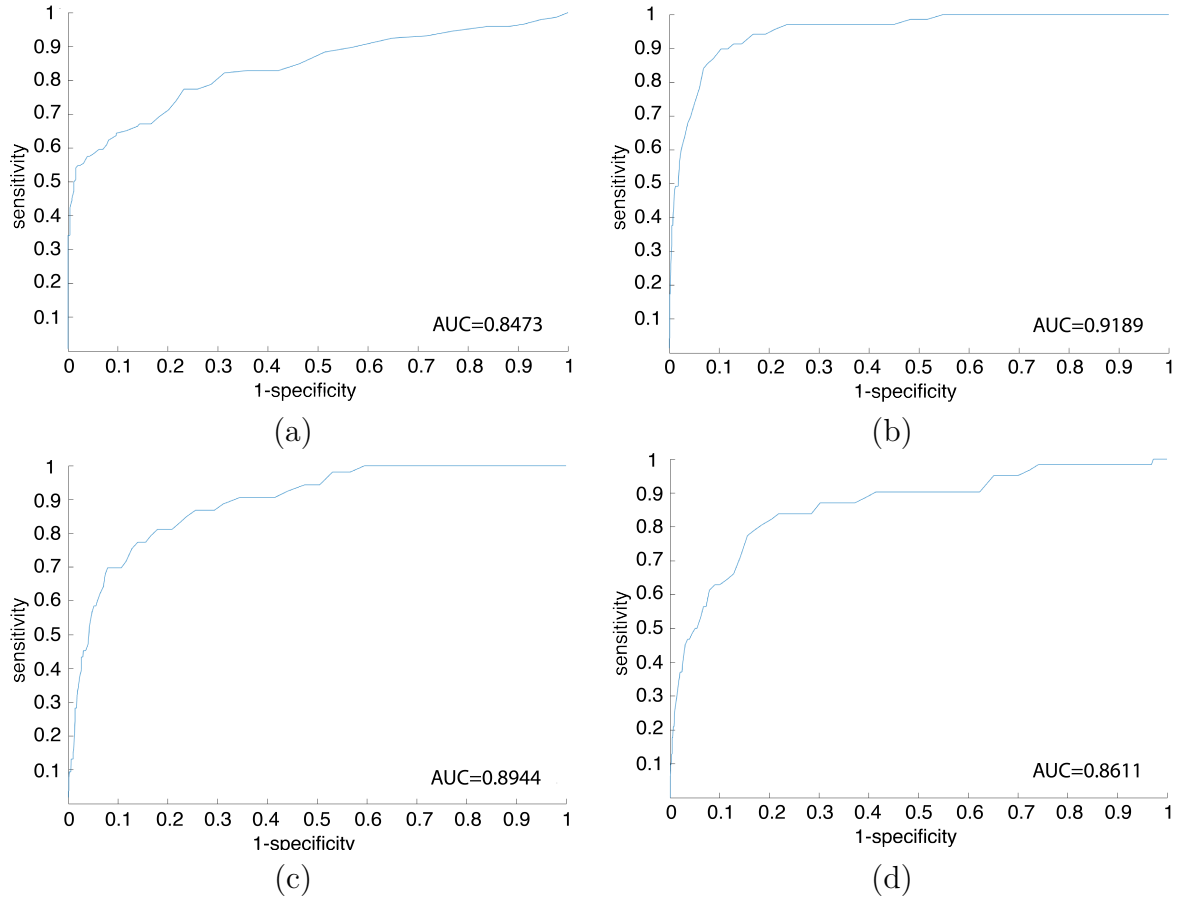
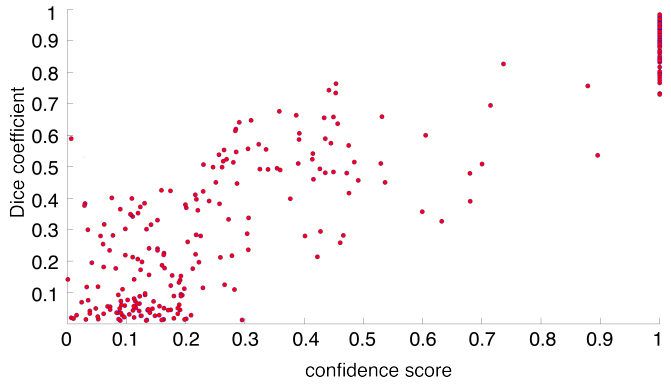


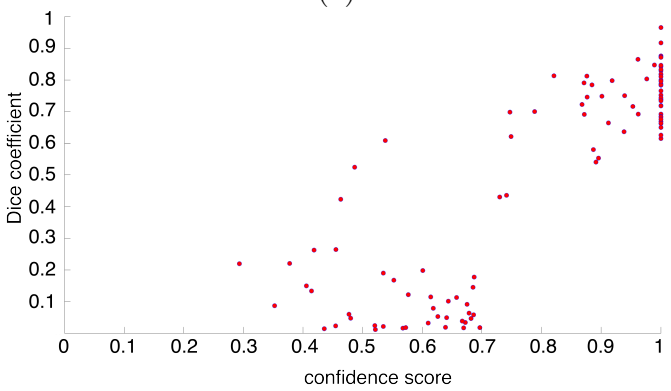
Figure 5.23: ROC curves showing the sensitivity, specificity and accuracy (AUC) of standalone detectors to localise the (a) Abdomen ($AUC = 0.8473$) (b) Stomach bubble ($AUC = 0.9189$) (c) Umbilical vein ($AUC = 0.8944$) (d) Spine ($AUC = 0.8611$) in the absence of any geometric priors.

detection accuracy was achieved via abdominal masking (Table 5.10), which reduced false positives caused by artefacts and anatomical structures falling outside the abdominal wall, and caused the greatest increase in spine detection accuracy.

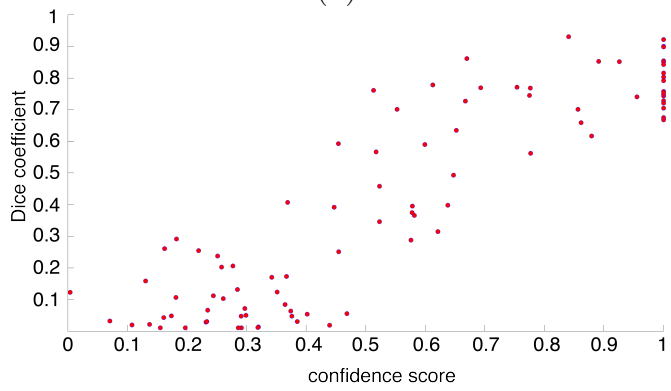
The inclusion of geometric constraints through the pictorial structures model caused a significant increase in umbilical vein detection accuracy, reducing false positives caused by artefacts and abdominal structures with a similar shape and appearance to the umbilical vein but in anatomically less probable positions relative to the stomach bubble and umbilical vein. A common failure mode of the pictorial structures model was false positive umbilical vein detection, whereby small artefacts or anatomical structures were incorrectly detected as the umbilical vein due to their



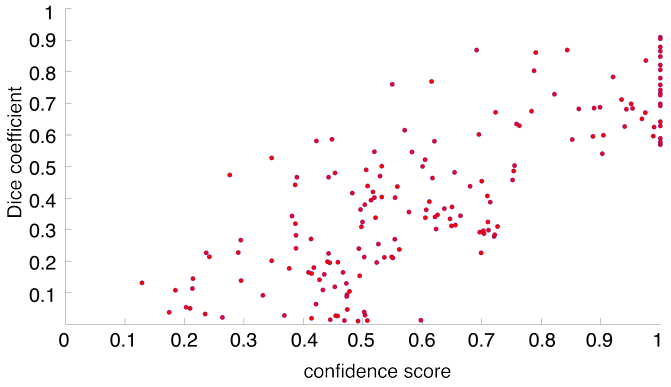
(a)



(b)



(c)



(d)

Figure 5.24: Bounding box Dice coefficients shown against confidence scores, for standalone detectors with no geometric constraints for the (a) Abdomen (b) Stomach bubble (c) Umbilical vein and (d) Spine.

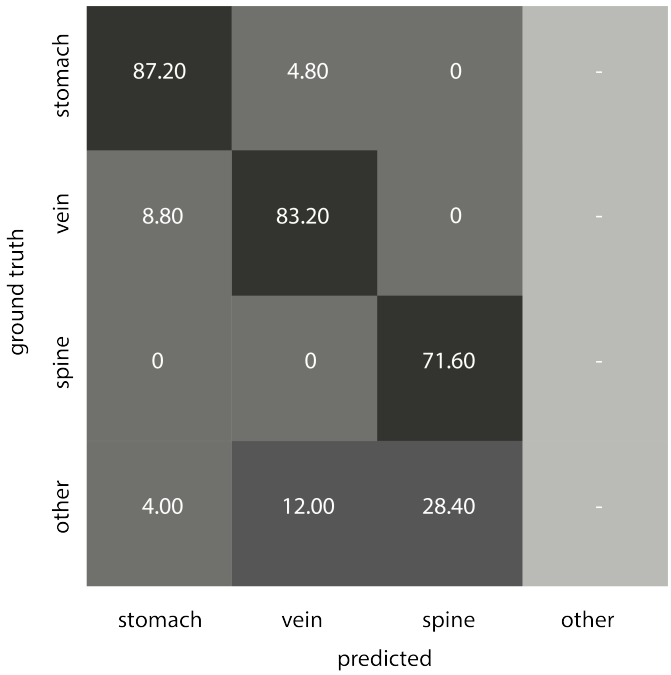


Figure 5.25: A confusion matrix showing percentage misclassifications of the pictorial structures model between the stomach bubble, umbilical vein, spine and other structures including artefacts or the fetal ribs.

positions relative to the stomach bubble and spine. In other cases, only a portion of the umbilical vein was detected due to changes in umbilical vein contrast or poorly defined umbilical vein boundaries. The addition of size constraints to the pictorial structures model to reduce instances of this failure mode was considered, but would be ineffective due to the large variation in umbilical vein size in 2-D US images.

Of the three anatomical landmarks used in the pictorial structures model, the spine displayed the lowest detection accuracy and the lowest standalone AUC due to the appearance of bright artefacts within the abdominal wall which were detected as false positives.

It's the eye of the tiger, it's the thrill of the fight.

— Survivor, American rock band, (1982)

6

An Eye Tracking Inspired Method for Standardised Abdominal Plane Selection

Contents

6.1 Introduction

The performance of the automated stomach bubble and umbilical vein localisation method implemented by Rahmatullah et al.[24] was assessed by computing its standardised plane selection accuracy across a set of 3-D fetal abdominal US

volumes. This chapter presents an extension of the 2-D pictorial structures model derived in Chapter 5 for the selection of standardised abdominal planes from 3-D fetal abdominal US volumes through the introduction of 3-D constraints.

A series of eye tracking experiments were conducted whereby observers were asked to localise the standardised abdominal plane whilst scrolling through 3-D abdominal US volumes. This was, at the time of submission, the first eye tracking study conducted with US volumes as stimuli. The spatial and temporal similarity of fixations and scrolling velocities were computed. It was demonstrated that observers exhibited a two-component visual search strategy, and the spine and umbilical vein were the most frequently fixated anatomical landmarks.

Informed by these findings, 3-D constraints were incorporated into the 2-D pictorial structures model derived in Chapter 5. The position and length of the fetal spine was constrained through the application of optical flow and length priors, and the length of the umbilical vein was similarly constrained. A dynamic programming algorithm was implemented to increase standardised abdominal plane selection speeds, with the final framework improving on the benchmark standardised abdominal plane selection accuracies reported by Rahmatullah et al.[66].

6.2 Originality and Individual Role

Independently, I designed and wrote Python applications to interface with eye tracking hardware, acquired and post-processed eye movements, and analysed the resulting data. Independently, I designed and wrote applications to implement, train and test a 3-D pictorial structures model, a dynamic programming algorithm, and derived probabilistic models based on optical flow and length to constrain the fetal spine and umbilical vein.

6.3 Eye Tracking

Eye tracking experiments were conducted to determine which anatomical features were fixated on by observers scrolling through 3-D abdominal US volumes, the

degree of similarity between their scrolling strategies, and whether their visual search strategies adhered to a two-component model of visual search.

6.3.1 Methods

Stimuli

The eye tracking stimuli consisted of dataset $V_{A\,n}$ as described in Chapter 4. Stimuli were presented to observers as described in Chapter 5. The consistent position of standardised abdominal planes within the central portion of each stimulus volume was a limitation of this experiment, and may have led to observers learning the likely positions of standardised planes and navigating to these positions as a default behaviour, rather than engaging in active scrolling and visual search behaviours.

Hardware

Eye movements were recorded with an EyeTribe v1.0 (the Eye Tribe, Denmark) eye tracker as described in Chapter 5.

Participants

Ten observers, with normal acuity, participated in this study. The observers were divided into expert and novice groups based on their level of experience in US imaging and interpretation. The expert group consisted of three clinical fellows with three years' US experience, and two PhD candidates in Biomedical Engineering. The novice group consisted of five clinical medical students. As in Chapter 5, a power analysis demonstrated that an effect size of 5.78% would be observable between the expert and novice groups, assuming a power (or true positive probability) of 0.9 and a significance level (or false positive probability) of 0.05.

Experimental Procedure

Calibration was conducted as described in Chapter 5. Each observer was presented with each stimulus volume in a pre-determined sequence, with unlimited viewing time available for each volume. Observers were instructed to scroll in the z-direction through each volume, along the longitudinal axis of the fetal abdomen at

a maximum speed of 14 frames per second (fps) using the keyboard. This maximum scrolling speed was chosen to ensure that fixations would still be detectable with an eye tracker sampling rate of $30Hz$. Observers were instructed to report, via the keyboard, when they had identified the volume frame which most closely adhered to the ISUOG criteria for standardised abdominal plane acquisition [1], before automatically proceeding to the next volume. A Gaussian white noise image was displayed for $5s$ between each volume to de-focus the observer's gaze as described in Chapter 5. To avoid fatigue, observers were given a ten minute break between each group of ten volumes. Ethics approval for this procedure was obtained via the University of Oxford Central University Research Ethics Committee (Reference: MS-IDREC-C1-2015-166).

This resulted in a set of raw gaze co-ordinates, where the j^{th} observer's gaze data on the n^{th} volume was described by $R_{n,j}$ and each set of gaze co-ordinates $R(x, y, z, d_{\text{os}}, t)$ consisted of the mean x and y co-ordinates across the left and right pupils, the z co-ordinate of the volume frame displayed on the screen, the observer-to-screen distance d_{os} and a timestamp t.

Fixation Filtering

Raw gaze co-ordinates $R_{n,j}$ were filtered into fixations and saccades using the I-VT [97] algorithm described in Chapter 5. Fixations occurring less than 0.5° apart in the x and y spatial dimensions were merged and classified as a single fixation, with a centroid calculated as the mean of the merged fixation points. Similarly fixations occurring on consecutive frames and less than 0.5° apart in the x and y spatial dimensions were indexed as members of the same fixation 'group'. This resulted in a set of fixations, where the j^{th} observer's fixations on the n^{th} volume were described by $F_{n,j}$ and each set of gaze co-ordinates $F(x, y, z, t, i)$ consisted of the fixation's x, y and z co-ordinates, timestamp t and fixation group index i.

Regions of Interest

The proportions of fixations falling within the 3-D ground truth bounding envelopes described in Chapter 4 ($V^{*3D}_{A\,m,n}$ where $m = 1, ..., 5$ for the abdominal wall, stomach bubble, umbilical vein, spine and ribs respectively) were computed for each stimulus volume where present. These ROIs were selected due to their consistent visibility in the majority of stimulus volume frames.

Static Consistency

Static consistency was computed to assess the agreement between the fixated locations of observers.

A set of 3-D binary fixation maps $B_{n,j} \in 0, 1$ (with volumes denoted by $n = 1, ..., 150$ and observers by $j = 1, ...10$) was computed for each observer and volume such that voxel values corresponding to fixation points were incremented by 1, and all other voxels had value 0. Each map $B(x, y, z)$ had equal spatial dimensions to the corresponding volume $V(x, y, z)$. As the human field of view typically extends 1.5° around a fixation point [9] in the x and y directions and each fixation 'group' as defined in Section 6.3.1 spanned a mean of 3.2 volume frames in the z-direction, the binary maps were convolved with a 3-D Gaussian kernel resulting in a set of attentional maps $A_{n,j} = B_{n,j} * G(\sigma_{x,y}, \sigma_z)$ (where $\sigma_{x,y} = 15$ voxels corresponding to 1.5° visual angle with an observer-to-screen distance of $0.5m$ and $\sigma_z = 3$ voxels corresponding to sustained visual attention across multiple frames in the z-direction) representing each observer's visual attention on each stimulus volume (Figure 6.1).

Static consistency across all observers on a particular volume was then computed via a leave-one-out scheme. For each observer, attentional maps for all other observers were summed so that $P_{n,J} = \Sigma_{1 \leq j \leq 10 j \neq J} \frac{A_{n,j}}{A_{\max}}$ describes a predictive map for the J^{th} observer's fixations on the n^{th} volume, normalised by the maximum value of the summed maps $A_{\max}$. A varying threshold between 0 and 1 was applied to binarize the predictive maps. At each threshold level, the sensitivity and specificity of the predictive map in predicting the current observer's attentional map were recorded. This resulted in a ROC curve with the AUC taken as the

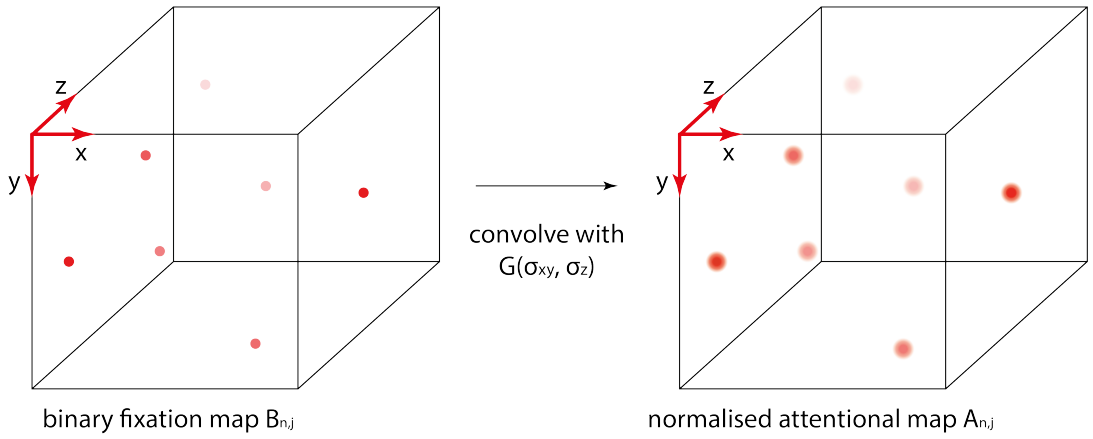


Figure 6.1: The convolution of a 3-D binary fixation maps with a 3-D Gaussian kernel to produce a 3-D normalised attentional map, representing a given observer's visual attention on a single stimulus US volume.

static consistency score. This was repeated across all observers and all volumes to compute a mean static consistency score.

Static consistency was computed for four sub-groups of observers: experts vs. novices (expert fixations predicted by novices), novices vs. experts (novice fixations predicted by experts), novices only, and experts only.

A cross-volume control measure was calculated to assess how accurately the attentional map of one observer on one randomly selected volume could be predicted by the summed attentional maps of all other observers on a different randomly selected volume. Repeated for 100 randomly generated volume and observer combinations, the mean cross-volume static consistency acted as a baseline static consistency score. As in Chapter 5, the control measure of static consistency was expected to be greater than random chance (50%) but less than static consistency scores computed for many observers on a single stimulus volume.

Dynamic Consistency

Dynamic consistency was computed to assess the agreement between the temporal sequences of the fixated locations of observers.

A fixation sequence was generated for each observer on each volume. This encoded the order in which observers fixated on the 3-D bounding envelopes described in Section 6.3.1 namely the abdominal cavity, stomach bubble, umbilical vein, spine, ribs, and background.

As in Chapter 5, the fixation sequences were treated as Markov chains, and dynamic consistency score was computed as the posterior probability of the current observer's fixation string occurring, based on a Markov model trained on the fixation strings of all other observers.

Dynamic consistency was computed for four sub-groups of observers: experts vs. novices (expert fixation sequences predicted by novices), novices vs. experts (novice fixation sequences predicted by experts), novices only, and experts only.

A cross-volume control was calculated to assess how accurately the fixation sequence of one observer on one randomly selected volume could be predicted by the fixation sequences of all other observers on a different randomly selected volume. Repeated for 100 randomly generated volume and observer combinations, the mean cross-volume dynamic consistency acted as a baseline dynamic consistency score.

Two Component Search

A two-component GMM was fitted to the 3-D US gaze data, in order to establish whether observers viewing US volumes displayed an equivalent discovery-reflective search strategy to that employed by observers viewing 2-D US images (Chapter 5). As Kundel et al.'s[10] discovery-reflective model[10] describes gaze trajectories with respect to targets in 2-D images, it may not be directly applicable to 3-D volumes where anatomical structures are evolving in shape and intensity as observers scroll through the stimuli. Developing a model of visual search on dynamic stimuli such as volumes and videos is therefore a potential avenue for further research; measuring observers' gaze trajectories relative to the 3-D spine bounding envelope constitutes an initial attempt to model search strategies in 3-D US volume stimuli.

For each volume, the Euclidean distance between the gaze trajectory of each observer and the centroid of the 3-D spine bounding envelope (the most frequently

fixated ROI), was computed as a function of time. As in Chapter 5, each observer's gaze trajectory was modelled with respect to the spine as a weighted sum of two Gaussian distributions representing the discovery and reflective phases of visual search respectively. The parameters for each distribution were found by iteratively solving an Expectation-Maximisation algorithm as described in Chapter 5, with initial parameter values set to 0.5. Relative entropy was computed as in Chapter 5 as a distance metric between GMMs for all pairings of observers, for each volume. The mean relative entropy across all volumes was taken as a measure of similarity between GMMs for pairs of observers, and described the extent to which observers exhibited similar two-component search strategies.

The reduced chi-squared statistic was computed as in Chapter 5 for each two-component GMM fitted to each gaze trajectory dataset. χ^2 was computed for one, three, four and five-component GMMs, to assess the goodness-of-fit of this two-component model compared to models containing greater or fewer Gaussian components.

As a control measure, relative entropies were computed between each possible pairing within 100 two-component GMMs with randomly generated weights, means and standard deviations. The mean relative entropy was taken as a baseline against which the relative entropy between the gaze trajectories of observers was assessed.

Scrolling Strategies

The velocity (frames/sec) at which observers navigated through each volume in the z-direction, along the longitudinal axis of the fetal abdomen, was computed where the j^{th} observer's scrolling velocities on the n^{th} volume were given by $S_{n,j}$.

It was hypothesised that expert observers would decelerate faster than novices as the first standardised abdominal plane in each volume became visible. It was also hypothesised that expert observers would perform fewer scrolling actions (where a single action is characterised by a period of scrolling through the volume at the constant maximum scrolling velocity of $14fps$) than novice observers prior to deciding on the final standardised abdominal plane, as a result of novices more

frequently under or over-shooting the set of standardised abdominal planes (for which $V^*_{A\,\text{plane}\,n,f} = 1$).

To quantify differences between expert and novice scrolling strategies, the ideal observer scrolling strategy was modelled as a Heaviside step function $H[t] = \{^{v\ n<t_s}_{0\ n\geq t_s}$. Here, the step function $H[t]$ is equal to the maximum scrolling velocity v of $14fps$ prior to t_s, the time at which the first plane where $V^*_{A\,\text{plane}\,n,f} = 1$ becomes visible, and equal to 0 thereafter. This characterised a single scrolling action at the maximum allowable scrolling velocity of $14fps$ followed by an immediate decrease in velocity to $0fps$ when the first standardised abdominal plane was visible.

Relative entropy was computed as a distance metric between each observer's scrolling velocities on each volume $S_{n,j}$, and the Heaviside function $H[t]$. The mean values of relative entropy between expert observers and the Heaviside function, and novice observers and the Heaviside function quantified the extent to which the scrolling strategies of expert and novice observers resembled the ideal Heaviside scrolling strategy.

6.3.2 Results

Regions of Interest

Mean viewing time per volume, mean time to the first fixation on an anatomical ROI (the stomach bubble, umbilical vein, spine or ribs), mean fixation length and the mean number of fixations per volume are shown for experts and novices in Table 6.1. A two-sample t-test showed significant differences between expert and novice viewing times ($t(8) = 21.75, p < 0.001$ where the t value for 8 degrees of freedom was 21.75, with a p value less than 0.001) and times to first ROI fixations ($t(8) = 11.59, p < 0.001$) but failed to show significant differences between expert and novice fixation lengths ($t(8) = 1.61, p = 0.11$) and fixations per volume ($t(8) = 1.59, p = 0.11$). As shown in Table 6.2 and Figure 6.2, the anatomical ROI most frequently fixated on by expert and novice observers was the spine. The majority of fixations (98.17% for experts and 92.27% for novices) fell within the abdominal wall, and the spine was the most commonly fixated ROI. A two-sample

Group	Viewing Time (s)	Time to First ROI (s)	Fixation Length (s)	Fixations per Volume
Experts	42.18 ± 11.35	10.68 ± 3.55	0.65 ± 0.34	28.30 ± 10.66
Novices	54.52 ± 12.48	14.77 ± 5.98	0.64 ± 0.30	25.93 ± 15.82

Table 6.1: Mean image viewing times, times to first ROI fixations, fixation lengths and fixations per volume for expert and novice observers shown with standard deviations.

Group	Abdominal Cavity	Stomach Bubble	Umbilical Vein	Spine	Ribs	Background
Experts	29.33 ± 8.08	13.81 ± 4.29	21.50 ± 6.23	33.06 ± 9.83	1.10 ± 0.75	1.83 ± 0.68
Novices	27.40 ± 9.21	11.76 ± 6.96	20.49 ± 5.88	30.18 ± 10.03	2.44 ± 0.74	7.73 ± 2.05

Table 6.2: The percentages of fixation points falling within manually segmented bounding boxes around the abdominal cavity, stomach bubble, umbilical vein, spine, ribs, and background region.

t-test showed significant differences between expert and novice fixations on the stomach bubble ($t(8) = 15.96, p < 0.01$) umbilical vein ($t(8) = 3.45, p < 0.01$) and spine ($t(8) = 6.23, p < 0.01$).

Static Consistency

Static consistency scores, as shown in Table 6.3 and Figure 6.2, were highest for expert fixation maps predicted by other experts only, suggesting that experts' fixated locations were significantly more consistent than those of novices ($t(8) = 7.89, p < 0.01$). Static consistency scores for all observer sub-groups were significantly higher than the random baseline measure according to a two-sample t-test with $t(8) = 15.47, p < 0.01$, $t(8) = 7.47, p < 0.01$, $t(8) = 15.77, p < 0.01$, $t(8) = 11.96, p < 0.01$ respectively for the experts vs. experts, experts vs. novices, novices vs. experts, and novices vs. novices static consistency scores against the random baseline.

Predicted Group	Predicting Group	Static Consistency
Experts	Experts	0.76 ± 0.13
Experts	Novices	0.73 ± 0.18
Novices	Experts	0.74 ± 0.12
Novices	Novices	0.71 ± 0.11
Random	Random	0.66 ± 0.26

Table 6.3: Static consistency scores for sub-groups of observers, namely expert fixation maps predicted by experts only, expert fixation maps predicted by novices only, novice fixation maps predicted by experts only, and novice fixation maps predicted by novices only. The random cross-image baseline measure is also shown.

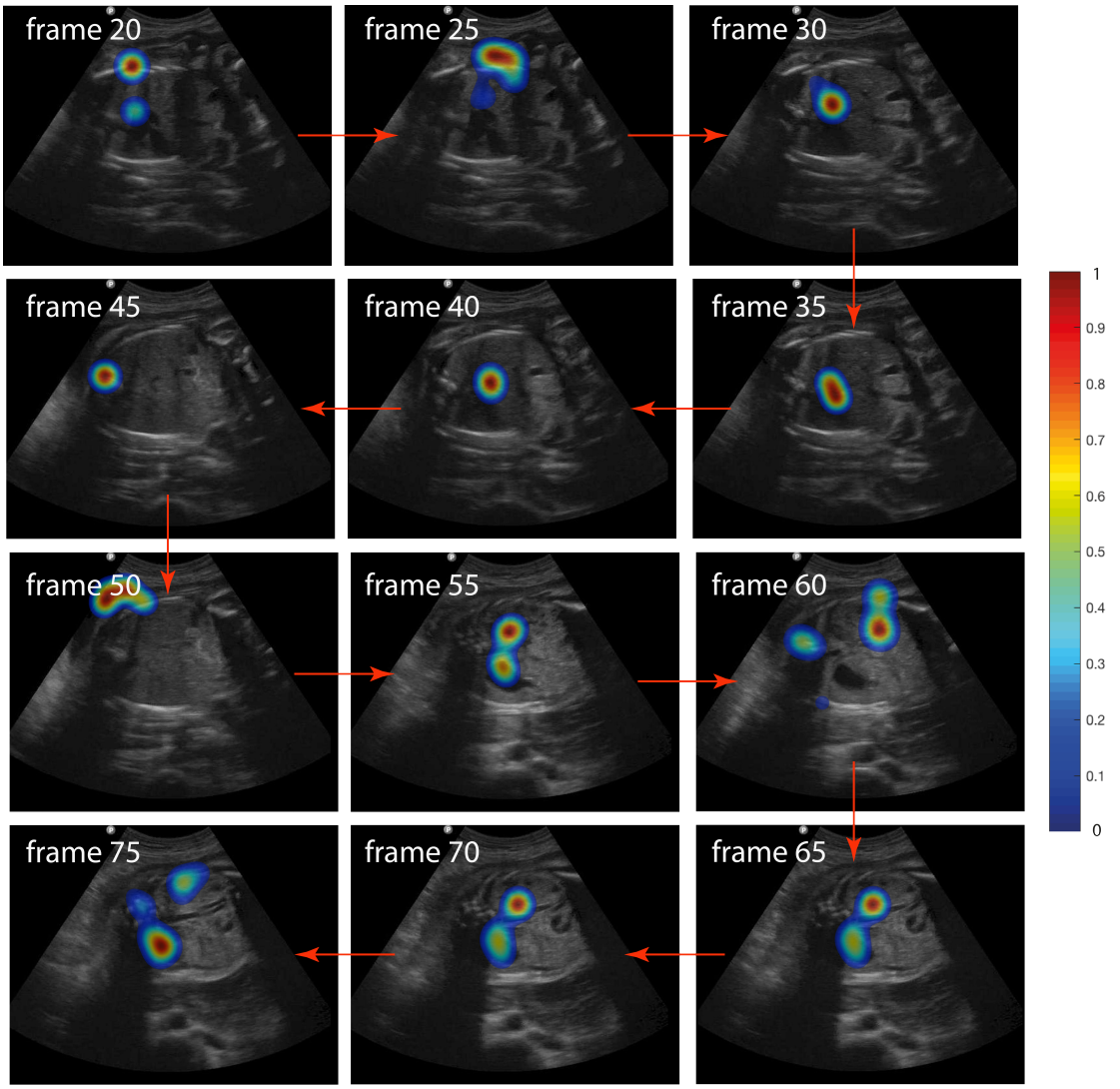


Figure 6.2: Heat maps showing the fixations of all observers on a sequence of frames in a single 3-D abdominal US volume. In particular, when key anatomical landmarks were not visible (frames 20, 25, 50) fixations fell primarily on the abdominal wall and ribs. When the stomach bubble and umbilical vein were visible (frames 35, 40, 60, 65, 70, 75) fixations fell on these landmarks. In some instances (frames 30, 45, 50) fixations fell solely on the fetal spine.

	Most Common Sequence of 3 Fixations		Most Common Sequence of 4 Fixations	
	1^{st}	2^{nd}	1^{st}	2^{nd}
Experts	$3-2-3$ (22%)	$2-3-2$ (21%)	$3-2-3-1$ (16%)	$2-3-2-1$ (14%)
Novices	$3-2-3$ (26%)	$2-3-2$ (23%)	$3-2-3-1$ (14%)	$2-3-2-1$ (13%)

Table 6.4: The 1^{st} and 2^{nd} most common sequences of 3 and 4 fixations for expert and novice observers, where ROI 1 denotes the stomach bubble, 2 denotes the umbilical vein and 3 the spine.

Dynamic Consistency

As shown in Table 6.4, the most common fixation sequences employed by both the expert and novice observer groups involved cross-referencing between the umbilical vein and spine, with longer fixation sequences also involving cross-referencing with the stomach.

Dynamic consistency scores (Table 6.5) were highest for expert fixation sequences predicted by other experts only- suggesting that expert sequences were significantly more consistent than those of novices ($t(8) = 9.68, p < 0.01$). However, dynamic consistency scores for all observer sub-groups were higher than the random baseline measure according to a two-sample t-test with $t(8) = 24.49, p < 0.01$, $t(8) = 23.57, p < 0.01$, $t(1498) = 19.01, p < 0.01$, $t(8) = 12.62, p < 0.01$ respectively for the experts vs. experts, experts vs. novices, novices vs. experts, and novices vs. novices static consistency scores against the random baseline.

Two Component Search

Relative entropies between two-component GMMs fitted to the gaze trajectories of observers are given in Table 6.6. Mean relative entropy for the experts vs. experts group was the lowest, suggesting a high degree of visual search similarity within this observer group (Figure 6.3). Mean relative entropy for the novices vs. novices group was the highest, suggesting a lower degree of visual search similarity within this group (Figure 6.4) and demonstrating that the former group employed significantly more consistent search strategies than the latter group ($t(8) = 12.48, p < 0.01$).

Predicted Group	Predicting Group	Dynamic Consistency
Experts	Experts	0.42 ± 0.14
Experts	Novices	0.39 ± 0.11
Novices	Experts	0.38 ± 0.14
Novices	Novices	0.33 ± 0.18
Random	Random	0.20 ± 0.17

Table 6.5: Dynamic consistency scores for sub-groups of observers, namely expert fixation sequences predicted by experts only, expert fixation sequences predicted by novices only, novice fixation sequences predicted by experts only, and novice fixation sequences predicted by novices only. The random cross-image baseline measure is also shown.

P Distribution	Q Distribution	Relative Entropy
Experts	Experts	30.30 ± 8.73
Experts	Novices	33.65 ± 6.54
Novices	Experts	37.11 ± 7.72
Novices	Novices	39.89 ± 12.46
Random	Random	154.33 ± 39.87

Table 6.6: Relative entropies of two-component GMMs representing the gaze trajectories of sub-groups of observers. Relative entropies are given for experts with respect to experts only, experts with respect to novices, novices with respect to experts, and novices with respect to novices only. A baseline measure of relative entropy calculated between randomly generated two-component GMMs is also shown. A lower entropy suggests a higher degree of similarity.

However, relative entropies for all sub-groups were significantly lower than the random baseline according to a two-sample t-test with $t(8) = 125.03, p < 0.01$, $t(8) = 118.61, p < 0.01$, $t(1498) = 112.07, p < 0.01$, $t(8) = 106.67, p < 0.01$ respectively for the experts vs. experts, experts vs. novices, novices vs. experts, and novices vs. novices static consistency scores against the random baseline. The reduced chi-squared statistic (Table 6.7) for two-component GMMs $x^2_{\text{red}} \approx 1$ suggesting that this model best fits the data. One and three-component GMMs displayed a poorer fit to the observed data with $x^2_{\text{red}} > 1$ and the reduced chi-squared statistic $x^2_{\text{red}} < 1$ for four and five-component GMMs suggested that in these cases, the model was over-fitted to the data.

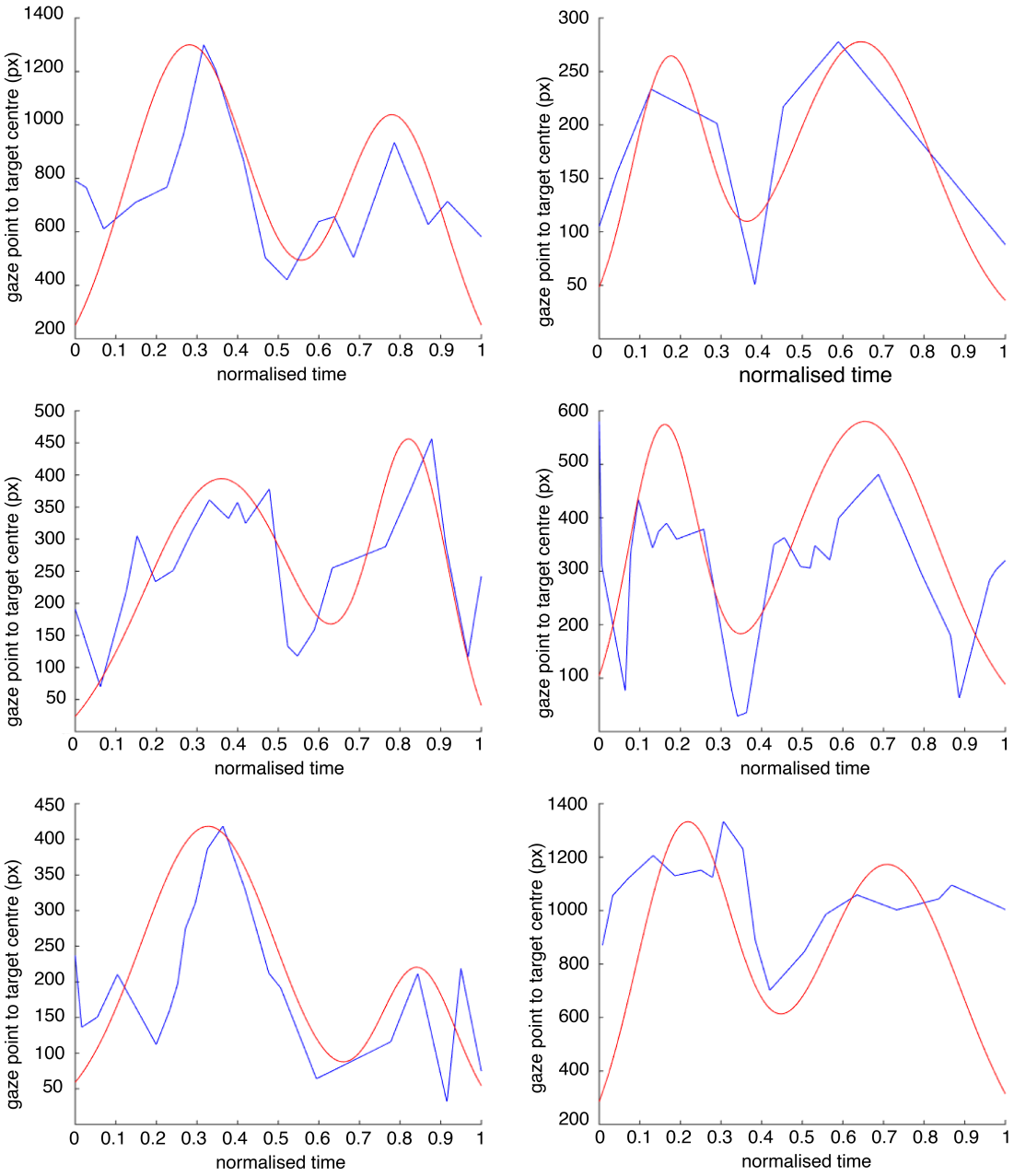


Figure 6.3: Two-component GMMs (red) fitted to the gaze trajectories (blue) of one expert observer viewing six examples of different stimulus volumes, where the gaze trajectory is plotted as the Euclidean distance between the gaze point and the centre of the spine bounding box against time, across the whole volume.

.

Gaussian Components	Reduced Chi-Squared Statistic
1	4.22
2	1.25
3	4.02
4	0.89
5	0.86

Table 6.7: Reduced chi-squared statistics for gaze trajectories represented by one, three, four and five-component GMMs, representing the goodness-of-fit of GMMs to observed data.

Scrolling Strategies

Relative entropies between the scrolling velocity profiles of observers ($S_{n,j}$) and the ideal Heaviside scrolling profile ($H[t]$) are given in Table 6.8. Mean relative entropy for the experts vs. experts group was the lowest, suggesting a high degree of scrolling behaviour similarity within this observer group (Figure 6.5). Mean relative entropy for the novices vs. novices group was the highest, suggesting a lower degree of scrolling behaviour similarity within this group (Figure 6.6) and demonstrating that the former group employed significantly more consistent scrolling strategies than the latter group ($t(1498) = 12.48, p < 0.01$). Additionally mean relative entropy for the experts vs. Heaviside group was significantly lower than for the novices vs. Heaviside group ($t(1498) = 4.11, p < 0.01$), suggesting that expert scrolling strategies were more similar to the ideal scrolling pattern defined by the Heaviside function.

Expert observers performed a mean of 8.9 ± 3.3 scrolling actions prior to reporting on the presence of the standardised abdominal plane, whereas novice observers performed a mean of 10.3 ± 5.4 scrolling actions further demonstrating that expert scrolling strategies were significantly ($t(1498) = 6.20, p < 0.01$) more efficient than those of novices.

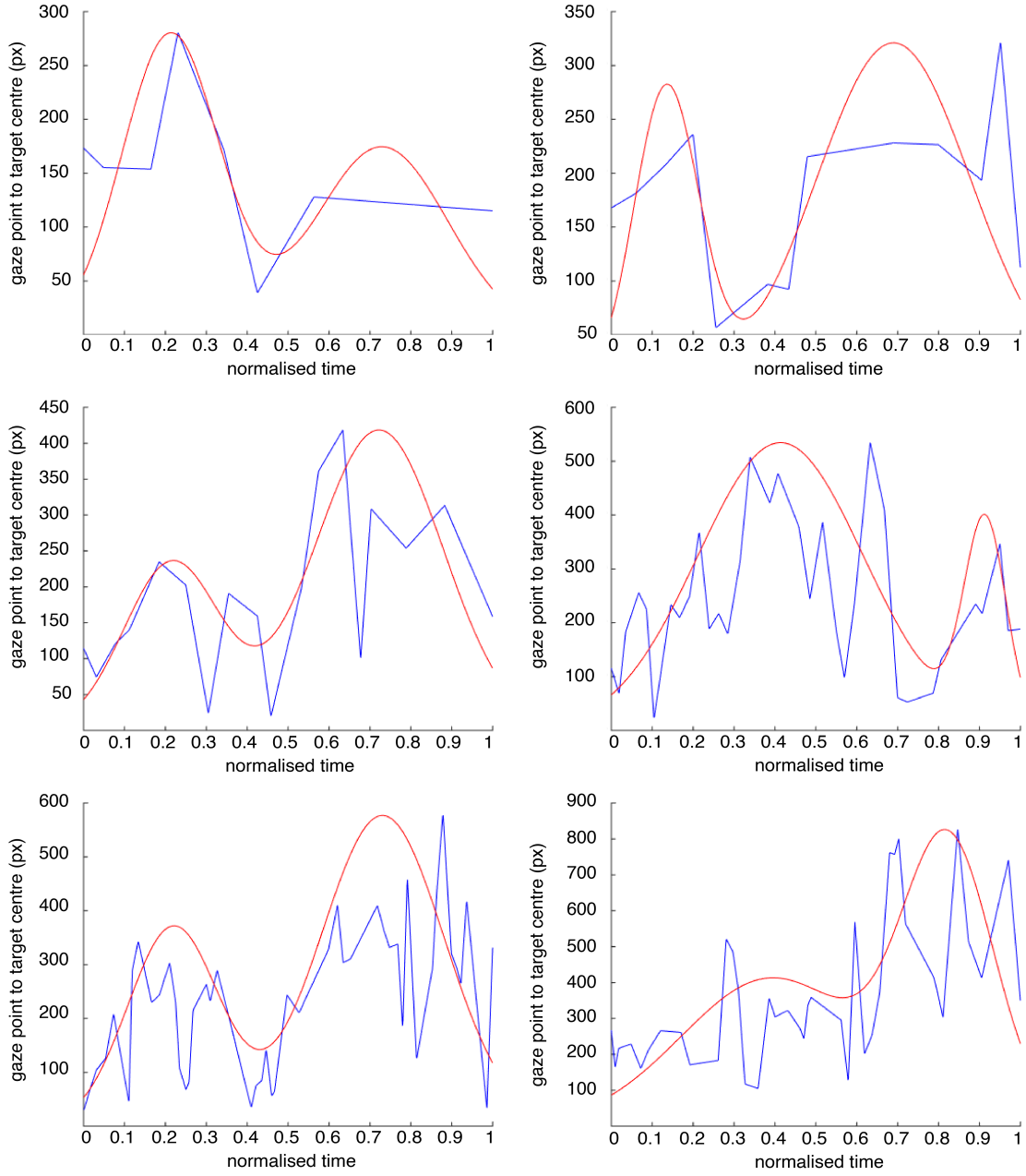


Figure 6.4: Two-component GMMs (red) fitted to the gaze trajectories (blue) of one novice observer viewing the same stimulus volumes shown in Figure 6.3, where the gaze trajectory is plotted as the Euclidean distance between the gaze point and the centre of the spine bounding box against time, across the whole volume.

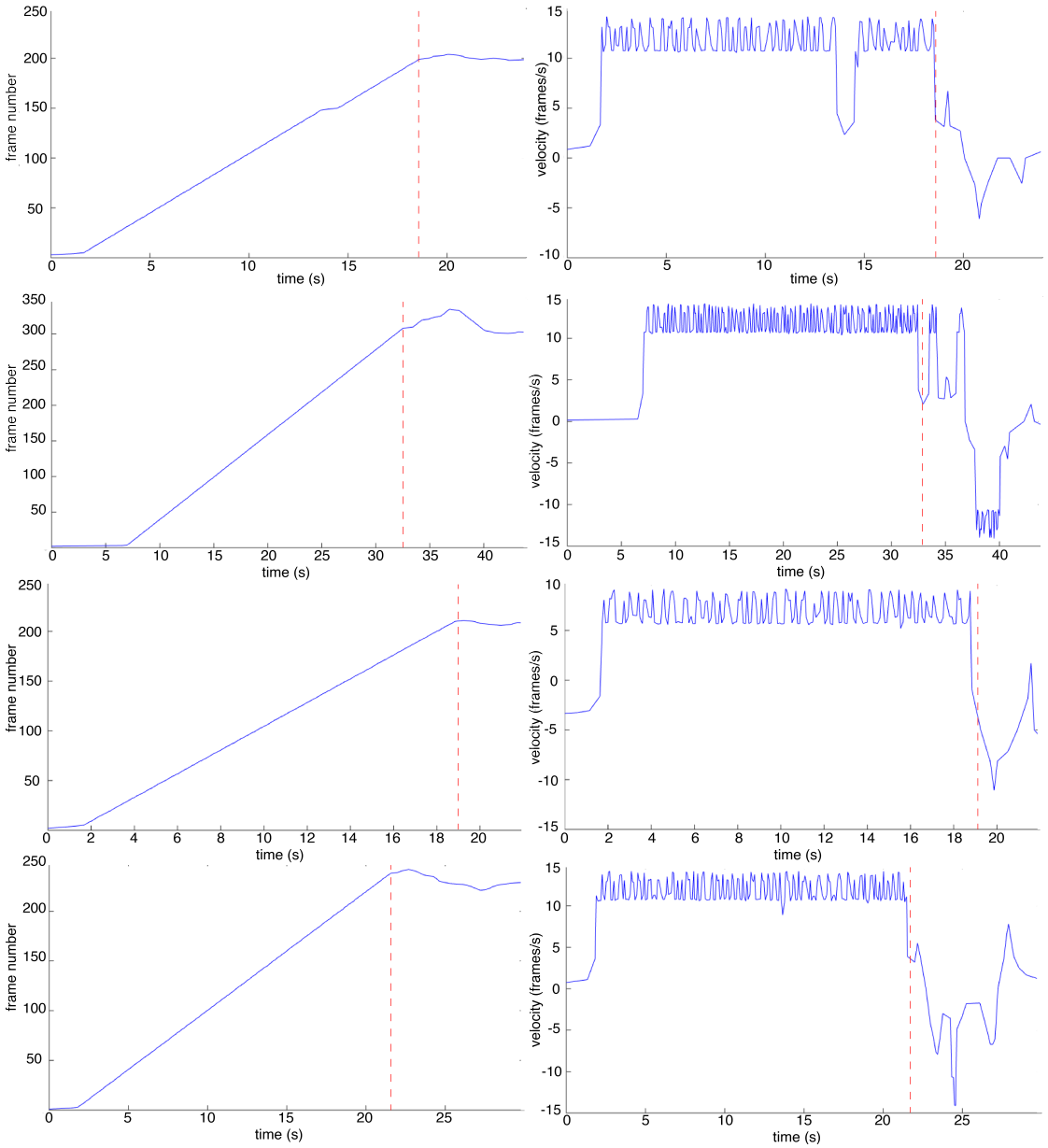


Figure 6.5: The scrolling velocities of an expert observer searching for the standardised abdominal plane in four examples of 3-D fetal abdominal US volumes, with (left) Frame number as a function of time (b) Scrolling velocity as a function of time. The time at which the standardised abdominal plane, as per the ISUOG criteria [16], becomes visible is shown in red.

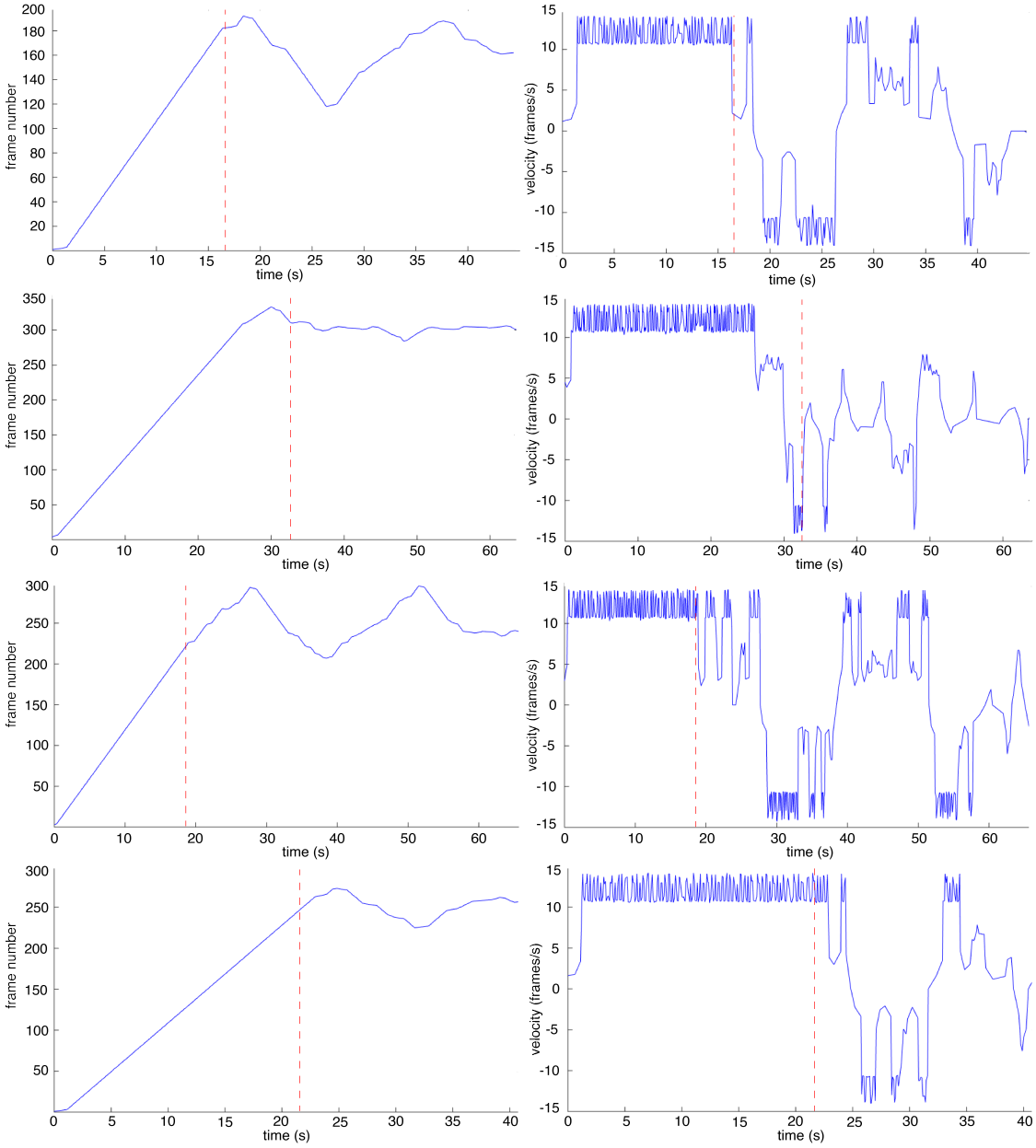


Figure 6.6: The scrolling velocities of a novice observer searching for the standardised abdominal plane in the same 3-D fetal abdominal US volumes shown in Figure 6.5, with (left) Frame number as a function of time (b) Scrolling velocity as a function of time. The time at which the standardised abdominal plane, as per the ISUOG criteria [16], becomes visible is shown in red.

P Distribution	Q Distribution	Relative Entropy
Experts	Experts	30.68 ± 11.37
Experts	Heaviside	40.53 ± 14.88
Experts	Novices	32.44 ± 13.94
Novices	Experts	34.39 ± 18.63
Novices	Novices	36.20 ± 17.66
Novices	Heaviside	44.35 ± 13.56

Table 6.8: Relative entropies of the scrolling strategies of sub-groups of observers. Relative entropies are given for experts with respect to experts only, experts with respect to the ideal Heaviside model, experts with respect to novices, novices with respect to experts, novices with respect to novices only, and novices with respect to the ideal Heaviside model. A lower entropy suggests a higher degree of similarity.

6.3.3 Discussion

ROI analysis demonstrated that observers fixated on the spine to a greater extent than the stomach bubble and umbilical vein. This is in contrast to fixations on 2-D stimuli as described in Chapter 5, where the spine was the least frequently fixated anatomical landmark. This suggests that when viewing 3-D US stimuli, observers employed the spine as a reference point against which to cross-reference the positions of other anatomical landmarks to a greater extent than when viewing 2-D US stimuli. It was hypothesised that the relatively constant appearance, intensity and position of the spine between consecutive volume frames explains its use as a 3-D constraint to confirm the positions of the stomach bubble and umbilical vein, both of which can have varying positions and appearances between consecutive volume frames in the z-direction.

The umbilical vein was the second most frequently fixated anatomical landmark, with both expert and novice observers displaying a high degree of cross referencing between the umbilical vein and spine. The most common fixation sequences were characterised by repeated fixations on the umbilical vein and spine followed by a final fixation on the stomach bubble, reinforcing the role of the spine and umbilical vein as 3-D reference points between consecutive volume frames.

The role of the umbilical vein as a 3-D constraint for observers is perhaps unexpected due to its varying appearance and position between frames. It was hypothesised that the mean length of the umbilical vein along the longitudinal axis of the fetal abdomen is significantly greater than the mean length of artefacts between consecutive volume frames. Therefore fixating on candidate umbilical veins across consecutive volume frames may have allowed observers to distinguish the umbilical vein from artefacts which may have resembled the umbilical vein in individual volume frames, but did not span a sufficient number of volume frames in the z-direction to be identified as viable umbilical vein candidates.

Expert observers exhibited significantly lower viewing times per volume than novice observers whilst performing a similar number of fixations per volume suggesting, as expected, that expert observers employed more efficient visual search strategies in order to locate the standardised abdominal plane in 3-D volumes. Static consistency scores suggested that the fixated volume regions of experts were significantly more consistent than the fixated regions of novices, suggesting that the visual search strategies of the latter group were more disparate as novices may not have been familiar with the expected appearances and locations of target anatomical landmarks.

Goodness-of-fit analysis of varying component GMMs fitted to gaze trajectories demonstrated that two-component GMMs most accurately modelled the collected gaze data, validating Kundel et al.'s[10] discovery-reflective visual search hypothesis with respect to 3-D US volumes. Relative entropies computed between the GMMs of expert observers were significantly lower than those calculated between novice GMMs, suggesting that the former group employed more consistent two-phase visual search strategies.

The scrolling strategies of expert observers adhered more closely to the ideal scrolling strategy, as modelled by the Heaviside function, than the scrolling strategies of novices. Additionally, scrolling strategies within the expert observers group were more consistent than those within the novice group, as demonstrated by the lower mean relative entropy for the former group. The significantly lower number of

scrolling actions per volume performed by experts compared to novices provides further evidence that expert observers employed more efficient scrolling strategies when navigating through the stimulus volumes.

It has been demonstrated that the fetal spine and umbilical vein were frequently fixated by observers searching for the standardised abdominal plane in 3-D abdominal US volumes. It was hypothesised that these anatomical landmarks were harnessed as 3-D constraints and were distinguishable from artefacts with similar 2-D appearances based on their lengths across volume frames.

Based on the findings of these eye tracking experiments, a modified pictorial structures model is proposed for the automated detection of standardised abdominal planes in 3-D fetal abdominal US volumes (Table 6.9). It is hypothesised that the construction of probabilistic models for the optical flow profiles and lengths of fetal spines and spine-line artefacts in the z-direction will show significant differences between the optical flow and length distributions of spine and spine-like artefacts. On the basis of these findings, 3-D probabilistic optical flow and length based constraints are introduced to candidate spines, to mimic frequent fixations on the spine as demonstrated in Section 6.3.1. Similarly, probabilistic constraints are introduced to the lengths of fetal umbilical veins and vein-line structures in the z-direction.

The following sections describe an investigation into whether these 3-D optical flow and length constraints can be harnessed to distinguish the fetal spine and umbilical vein from artefacts, thus improving the accuracy of the 2-D pictorial structures model described in Chapter 5.

6.4 Standardised Abdominal Plane Selection Without 3-D Constraints

This section reports on the accuracy of the 2-D pictorial structures model, derived in Chapter 5, in detecting standardised abdominal planes from 3-D fetal abdominal US volumes without the introduction of additional 3-D optical flow and length based constraints. Section 6.5 will consider their inclusion.

Eye Tracking Result	Detector Design Decision
Two-phase search strategy displayed by observers	Two-phase detector consisting of sliding window detector followed by pictorial structures model
The spine was the most frequently fixated anatomical landmark	3-D optical flow and length based constraints on candidate spine bounding boxes
The umbilical vein was the second most frequently fixated anatomical landmark	3-D length based constraints on candidate umbilical vein bounding boxes

Table 6.9: Key findings of the eye tracking experiments outlined in Section 6.3.1, and corresponding decisions on the design of a detector for the automated selection of standardised abdominal planes.

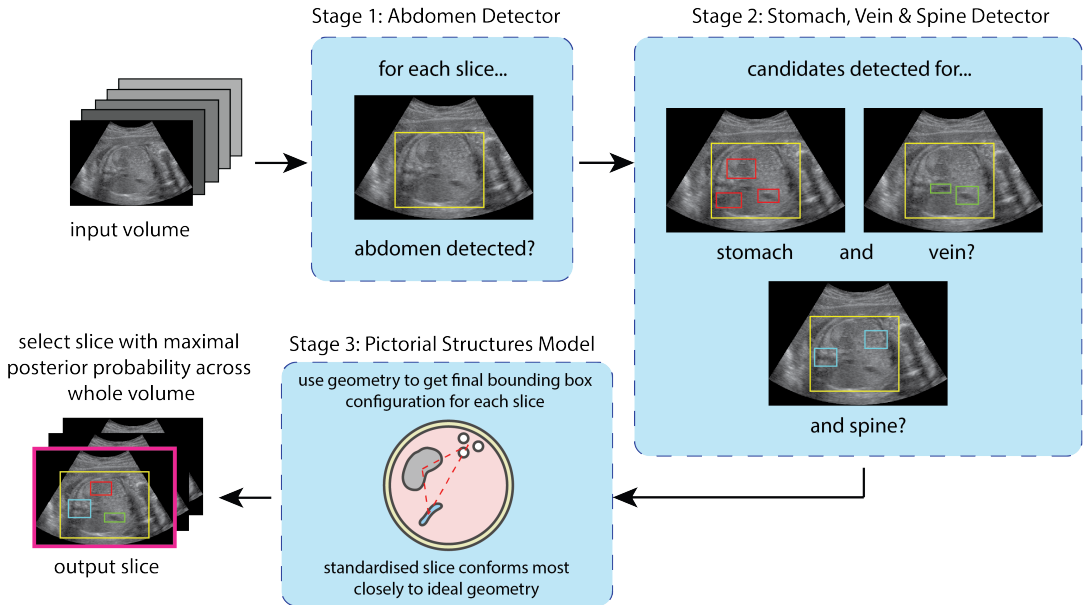


Figure 6.7: Schematic overview of a standardised abdominal plane selection framework with no 3-D constraints, where the first stage detects the abdominal wall, the second stage identifies candidate bounding boxes for the stomach bubble, umbilical vein, and spine, the third stage finds the optimal set of bounding boxes which maximise a posterior probability distribution based on candidate confidence scores and positions, and the final stage selects the frame which maximises the posterior probability over the whole volume as the standardised abdominal plane.

6.4.1 Methods

Testing Data

The 2-D pictorial structures model trained in Chapter 5 was applied to each frame of the testing set $V_{B\,n}$. Specifically, for each frame, a sliding window detector consisting of a boosted set of decision stumps detected possible candidates for the abdomen, stomach bubble, umbilical vein and spine. A pictorial structures model

determined the optimal configuration of candidate landmarks which maximised the posterior probability arg $\max_f p(L|V_{B\,n,f},\theta) \propto \Pi_{i=1}^{n} p(I|l_i,u_i)\Pi_{v_i,v_j} p(l_i,l_j|c_{ij})$ of a particular configuration of bounding boxes matching the ground truth for a given frame of a given volume, as a function of confidence scores and how closely the relative locations of the parts adhered to the geometric model. Here, L is a particular configuration of parts, $V_{B\,n,f}$ is a volume frame, θ are model parameters, l_i is a part at position i with a corresponding match score u_i, and c_{ij} is a cost function based on the distance between two parts at positions i and j. The resulting posterior probabilities $p(L|V_{B\,n,f},\theta)$ were normalised within each volume $V_{B\,n}$ and the frame $V_{B\,max\,n}$ which maximised the posterior probability across all frames f of a volume was selected as the standardised abdominal plane.

Individual detection accuracies for the abdomen, stomach bubble, umbilical vein and spine were defined as the mean proportion of detected bounding boxes across each volume for which the Dice coefficient between the geometrically optimal bounding box and the ground truth bounding box was greater than 0.75.

A correct standardised abdominal plane detection was defined as one for which the volume frame selected by the 2-D pictorial structures model corresponded to one of the manually labelled ground truth frames such that $V_{B\,max\,n} \in (V^{*}_{B\,\text{plane}\,n,f} = 1)$.

6.4.2 Results

The standardised plane selection accuracy across $V_{B\,n,f}$ was found to be 83.3%, lower than Rahmatullah's benchmark accuracy of 91.29%[7]. The framework displayed a non-optimised mean run-time of $20.1s$ per volume, on a MacBook Pro $2.8GHz$ Intel Core i7 processor, with $16GB$ $1600MHz$ DDR3 RAM.

Standalone mean detection accuracies for the abdomen, stomach bubble, umbilical vein and spine are shown in Table 6.10 prior to abdominal masking, and after the application of 2-D geometric constraints via the pictorial structures model. The confusion matrix in Figure 6.8 demonstrates that 11.88% and 28.36% of structures classified as the umbilical vein and spine respectively were misclassified artefacts. Figure 6.9 shows the ROC curve produced by applying a varying posterior

Landmark	Benchmark	Standalone	Abdomen Masking	2-D Pictorial Structures Model
Abdomen	-	91.30	-	-
Stomach bubble	78.94	74.79	82.71	87.43
Umbilical vein	62.80	62.93	73.98	82.40
Spine	-	59.48	68.38	71.64

Table 6.10: The accuracy in stomach bubble, umbilical vein and spine detection at each stage in the development of the anatomical landmark detector described above. Standalone detector accuracies with no geometric priors are given alongside standalone detector accuracies after abdominal masking, and after the application of geometric constraints via the pictorial structures model. Benchmark stomach bubble and umbilical vein accuracies as reported by Rahmatullah et al.[6] are also shown for reference.

probability threshold between 0 and 1 across $V_{B\,n}$ and computing the sensitivity and specificity in the detection of planes for which $V^*_{B\,\text{plane}\,n,f} = 1$ at each threshold value, resulting in a mean AUC of 0.78.

Correctly identified standardised abdominal planes fell into two categories; those where the stomach bubble, umbilical vein and spine were correctly localised as shown in Figure 6.10 (92% of correctly selected planes), and those where some or all of these landmarks were incorrectly localised (8% of correctly selected planes) as shown in Figure 6.11. All instances of incorrect plane selection were attributable to incorrect anatomical landmark localisation.

6.4.3 Discussion

This framework displayed a lower standardised abdominal plane selection accuracy than the method of Rahmatullah et al.[66].

The first limitation of this approach, as shown in Figures 6.11 and 6.8, is the misclassification of artefacts as the umbilical vein and spine. From observation, these artefacts typically do not propagate across consecutive volume frames in the z-direction to the same extent as the umbilical vein and spine. It is therefore hypothesised that spine and umbilical vein detection accuracies may be improved

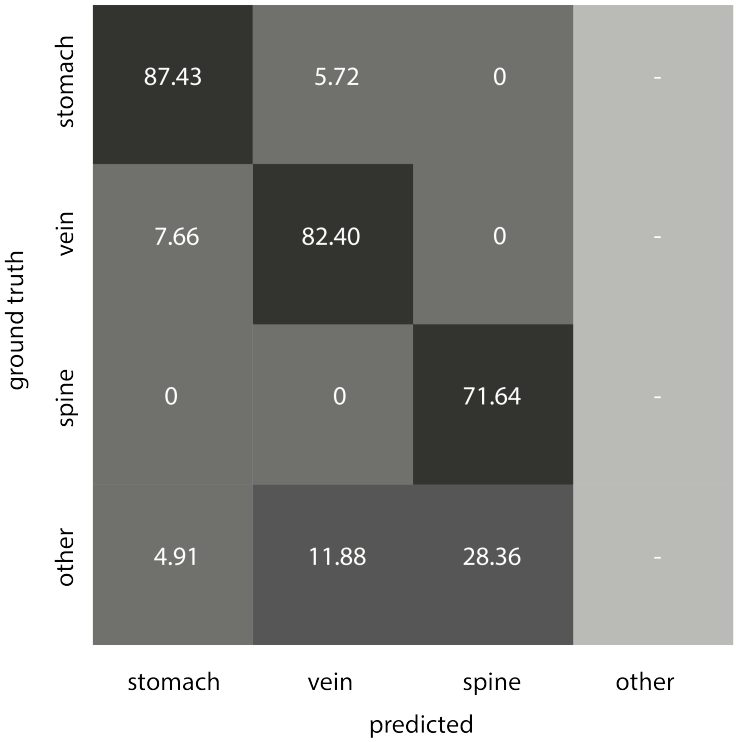


Figure 6.8: A confusion matrix showing percentage misclassifications of the 2-D pictorial structures model between the stomach bubble, umbilical vein, spine and other structures including artefacts or the fetal ribs.

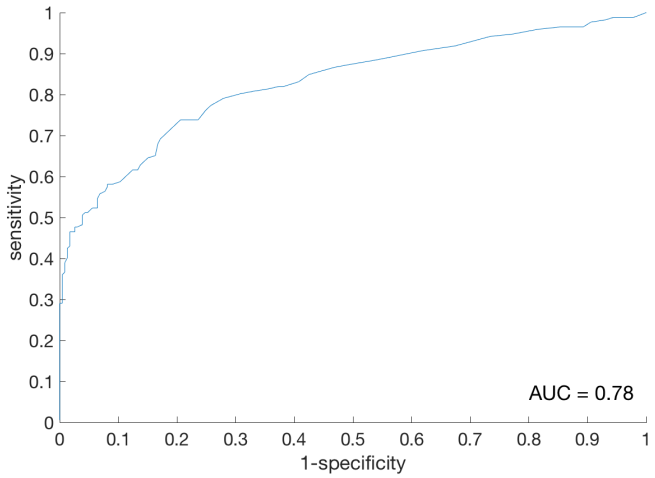


Figure 6.9: ROC curve showing the sensitivity, specificity and accuracy (AUC) of a 2-D pictorial structures model in standardised abdominal plane selection, in the absence of 3-D geometric priors.

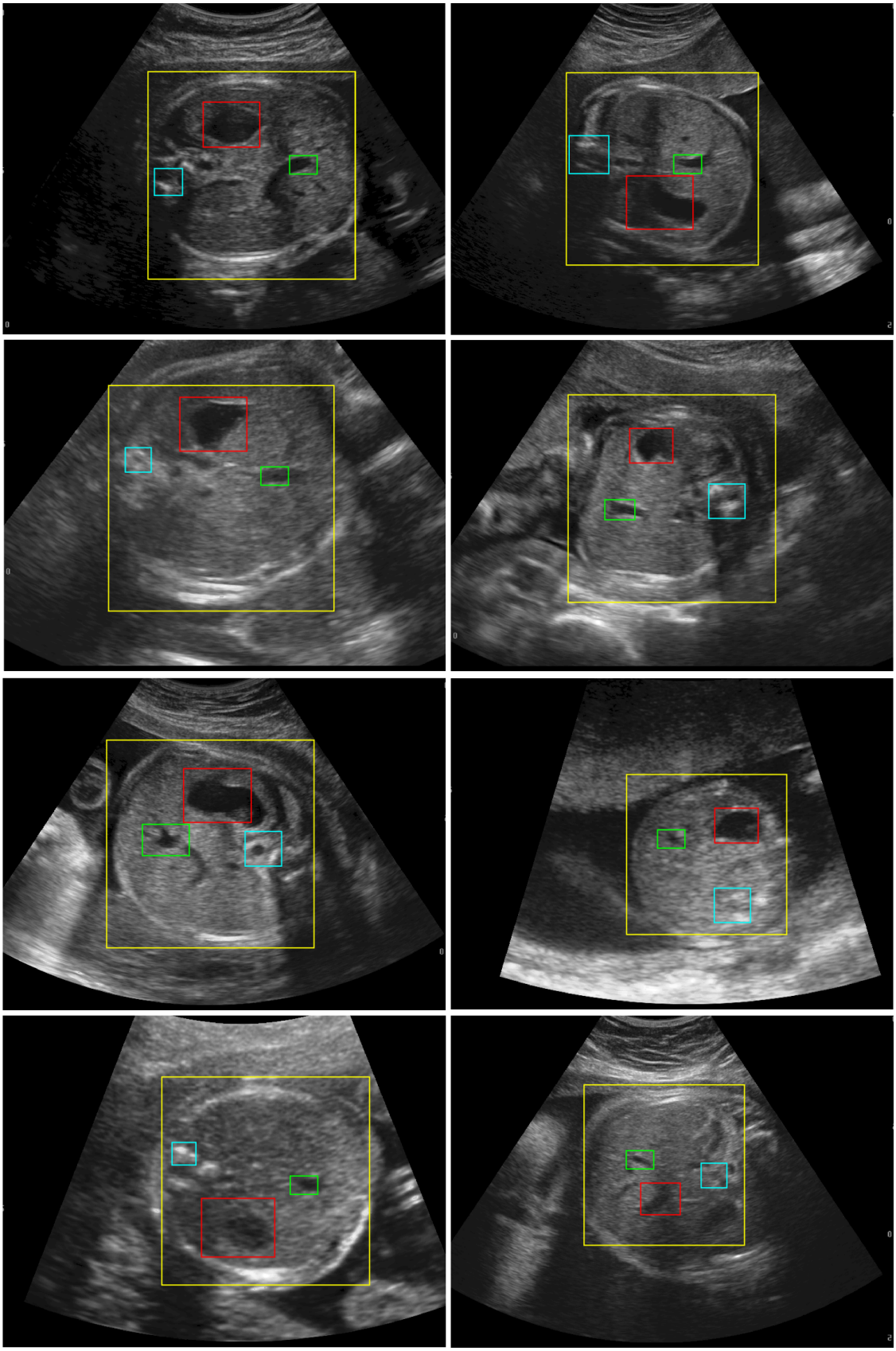


Figure 6.10: Correctly detected standardised abdominal planes using a 2-D pictorial structures model and no 3-D constraints, with correct anatomical landmark detections. Detected bounding boxes (red, green and blue for the stomach bubble, umbilical vein and spine respectively) have a Dice overlap coefficient greater than 0.75 with ground truth bounding boxes.

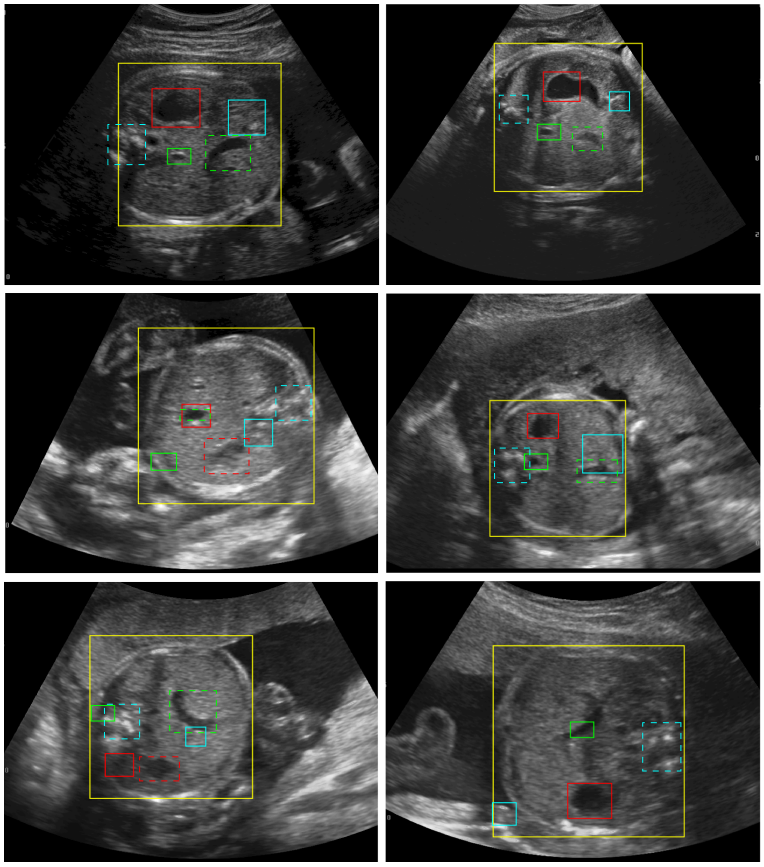


Figure 6.11: Correctly detected standardised abdominal planes using a 2-D pictorial structures model and no 3-D constraints, with incorrect anatomical landmark detections. Ground truth bounding boxes are shown in red dashed, green dashed, and blue dashed for the stomach bubble, umbilical vein and spine respectively where they do not have a Dice overlap coefficient greater than 0.75 with the bounding boxes detected by the pictorial structures model, shown in red solid, green solid and blue solid for the stomach bubble, umbilical vein and spine respectively.

by making use of 3-D constraints on the appearance and length of the spine, and the length of the umbilical vein.

Secondly, the process of computing the posterior probability $p(L|V_{B\,n,f}, \theta)$ of each possible bounding box configuration across all volume frames could be improved through implementing a dynamic programming scheme. This approach is described next.

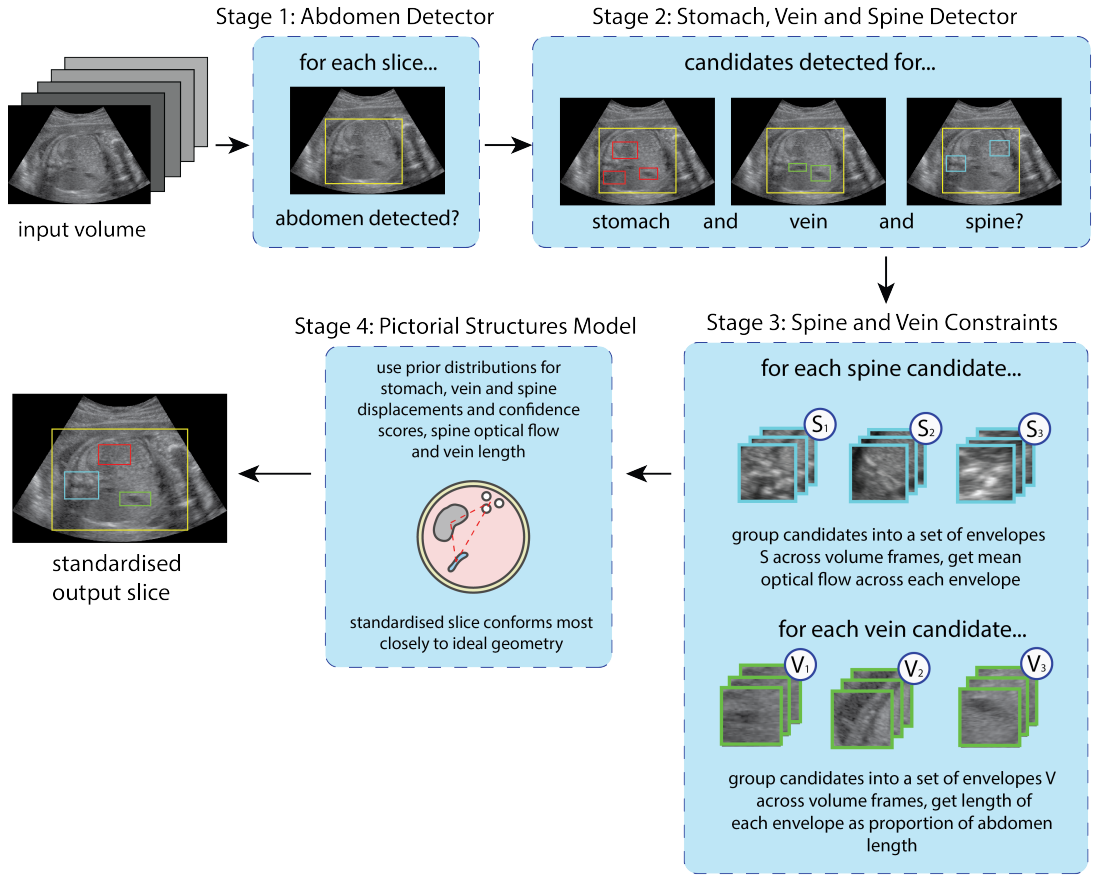


Figure 6.12: Schematic overview of a standardised abdominal plane selection framework, where the first stage detects the abdominal wall, the second stage identifies candidate bounding boxes for the stomach bubble, umbilical vein, and spine, the third stage finds the optimal set of bounding boxes which maximise a posterior probability distribution based on candidate confidence scores and positions, and the final stage selects the slice which maximises the posterior probability over the whole volume as the standardised abdominal plane.

6.5 Standardised Abdominal Plane Selection With 3-D Constraints

The findings in Section 6.3.1 were harnessed to implement an extended pictorial structures model for the automated selection of standardised abdominal planes with the inclusion of 3-D constraints on the appearance and length of the spine, and the length of the umbilical vein (Figure 6.12). The specific design decisions resulting from the eye tracking experiments are summarised in Table 6.9.

6.5.1 Methods

Training Data

Probability distributions for optical flow and length were constructed using dataset $V_{A\,n}$ as described in Chapter 5. The corresponding ground truth annotations for the f^{th} frame of the n^{th} volume were $V^*_{A\,m,n,f}$ for 2-D bounding boxes, $V^{*3D}_{A\,m,n}$ for 3-D bounding envelopes, and $V^*_{A\,\text{plane}\,n,f}$ for standardised planes, where $m = 1, ...4$ for the abdominal wall, stomach bubble, umbilical vein and spine.

Bounding boxes for umbilical vein and spine-like artefacts were described by $V^*_{A\,\text{uv}\,n,f}$ and $V^*_{A\,\text{sp}\,n,f}$, and bounding envelopes by $V^{*3D}_{A\,\text{uv}\,n,p}$ and $V^{*3D}_{A\,\text{sp}\,n,p}$.

Constraining Spine Length and Position

The distributions of lengths, in the z-direction along the longitudinal axis of the fetal abdomen, of ground truth 3-D fetal spine envelopes $V^{*3D}_{A\,4,n}$ were compared against those of 3-D spine-like artefact envelopes $V^{*3D}_{A\,\text{sp}\,n,p}$ as proportions of the lengths of the corresponding 3-D ground truth fetal abdomen envelopes $V^{*3D}_{A\,1,n}$.

The distributions of the lengths of the 3-D spine and spine-like artefact envelopes as proportions of abdominal envelope lengths were modelled by Gaussian distributions with parameters shown in Table 6.11. A two-sample t-test ($t(298) = 84.30, p < 0.01$) demonstrated that the lengths of ground truth spine envelopes were significantly larger than those of artefacts displaying a similar appearance to the spine. These findings suggest that the introduction of an additional geometric constraint on the length of candidate spine bounding envelopes may decrease the false positive detection of bright, spine-like artefacts.

Similarly the distributions of optical flow, in the z-direction along the longitudinal axis of the fetal abdomen, of ground truth 3-D fetal spine envelopes were compared against those of 3-D spine-like artefacts.

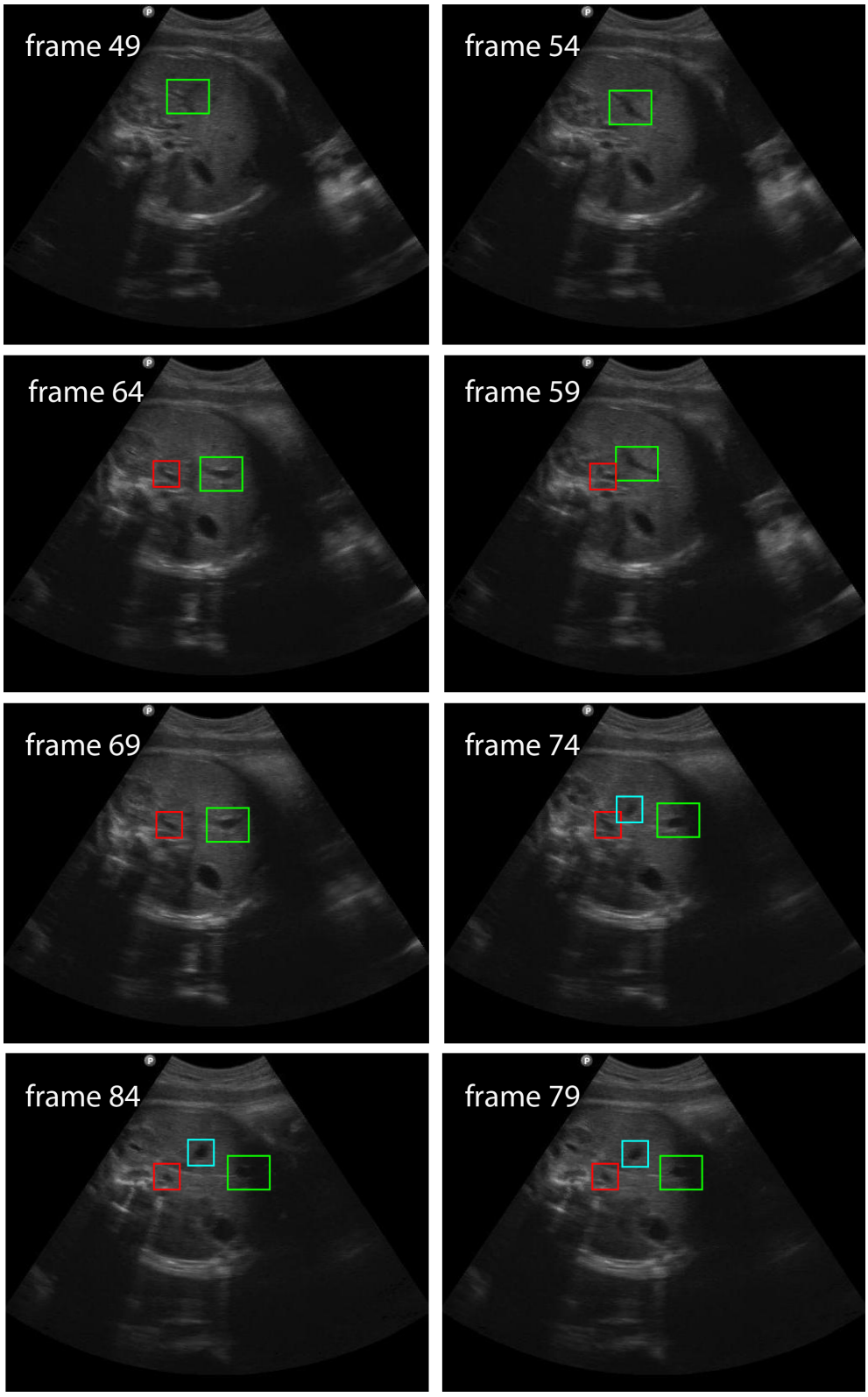


Figure 6.13: The annotation of ground truth umbilical vein (green) and umbilical vein-like artefact (red, blue) envelopes across a single 3-D fetal abdominal US volume. Here, the 3-D umbilical vein envelope extends for 8 of 76 volume frames and the two umbilical vein-like artefacts envelope extends for 3 and 6 of 76 volume frames.

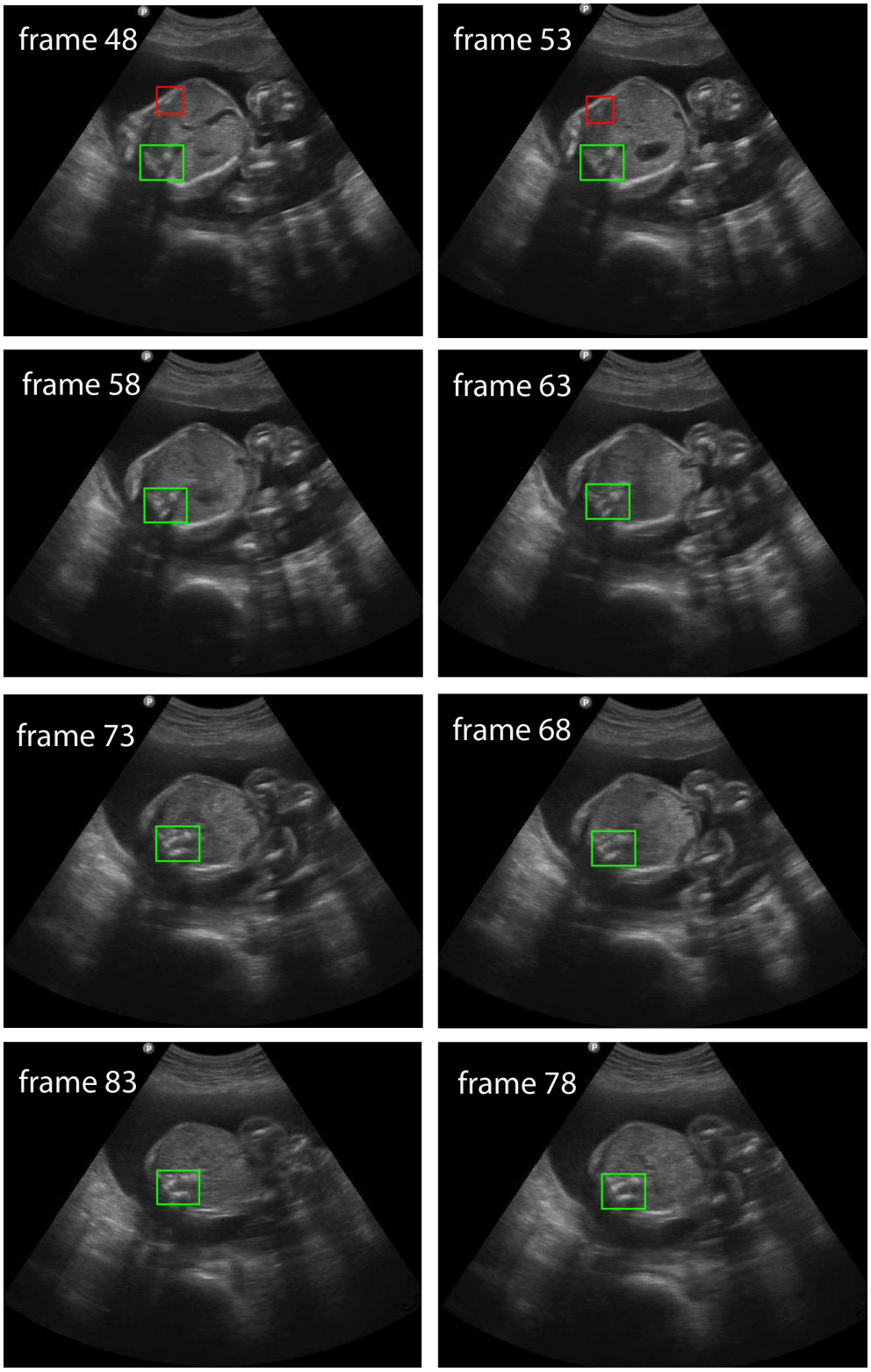


Figure 6.14: The annotation of ground truth spine (green) and spine-like artefact (red) envelopes across a single 3-D fetal abdominal US volume. Here, the 3-D spine envelope extends for 35 of 80 volume frames and the spine-like artefact envelope extends for 2 of 80 volume frames.

Landmark	Length	Optical Flow
Spines	0.89 ± 0.14	0.055 ± 0.008
Artefacts	0.09 ± 0.03	0.063 ± 0.006

Table 6.11: The mean and standard deviations of the lengths and mean optical flow of ground truth spine envelopes and spine-like artefact envelopes.

The optical flow vectors $[u, v]^T$ across the envelopes $V^{*3D}_{A\,4,n}$ and $V^{*3D}_{A\,\mathrm{sp}\,n,p}$ were computed through the Lucas-Kanade method[120] to characterise the motion and variations in appearance of ground truth spine bounding boxes and spine-like artefact bounding boxes between consecutive volume frames.

The distributions of mean optical flow $\sqrt{(u^2 + v^2)}$ magnitude across the 3-D spine and spine-like artefact envelopes were modelled by Gaussian distributions with parameters shown in Table 6.11. A two-sample t-test $t(298) = 10.63, p < 0.01$ demonstrated that the mean optical flow profiles of ground truth spine envelopes were significantly lower than those of artefacts displaying a similar appearance to the spine, confirming the hypothesis discussed in Section 6.4.2. These findings suggest that the introduction of an additional geometric constraint on the mean optical flow across candidate spine bounding envelopes may also increase the proportion of correct fetal spine detections and hence standardised abdominal plane selection accuracy.

Constraining Umbilical Vein Length

The distributions of lengths, in the z-direction along the longitudinal axis of the fetal abdomen, of ground truth 3-D fetal umbilical vein envelopes $V^{*3D}_{A\,3,n}$ were compared against those of 3-D umbilical vein-like artefacts $V^{*3D}_{A\,\mathrm{uv}n,p}$ as proportions of the lengths of the corresponding 3-D ground truth fetal abdomen envelopes $V^{*3D}_{A\,1,n}$.

The distributions of the lengths of the 3-D umbilical vein and umbilical vein-like artefact envelopes as proportions of abdominal envelope lengths were modelled by Gaussian distributions with parameters shown in Table 6.12. A two-sample t-test ($t(298) = 16.22, p < 0.05$) demonstrated that the lengths of ground truth umbilical vein envelopes were significantly larger than those of artefacts displaying

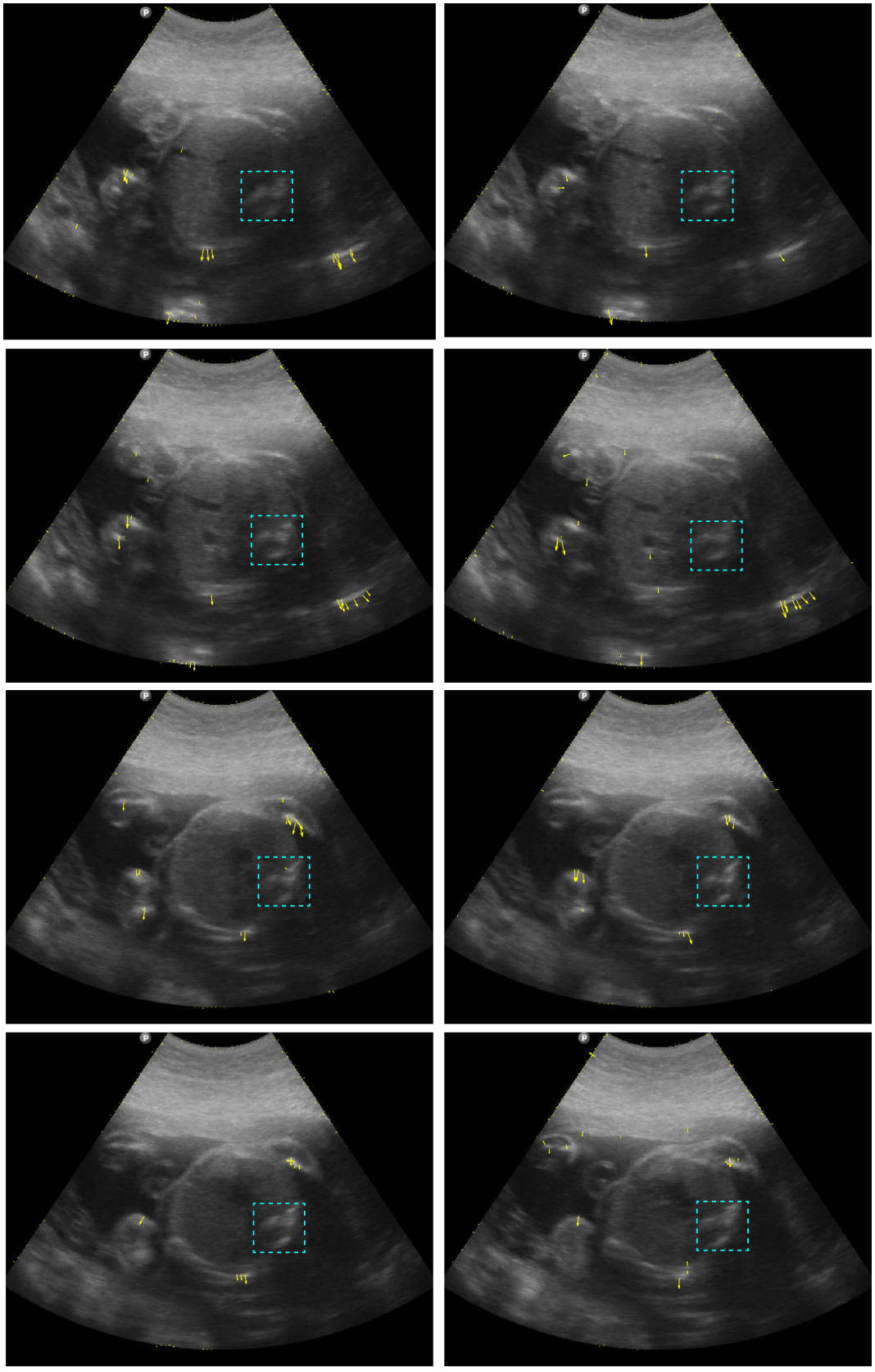


Figure 6.15: Optical flow (yellow) of bright spine-like artefacts between consecutive frames of a 3-D fetal abdominal US volume, where the ground truth spine envelope (blue, dashed) across frames is of constant position and displays no optical flow.

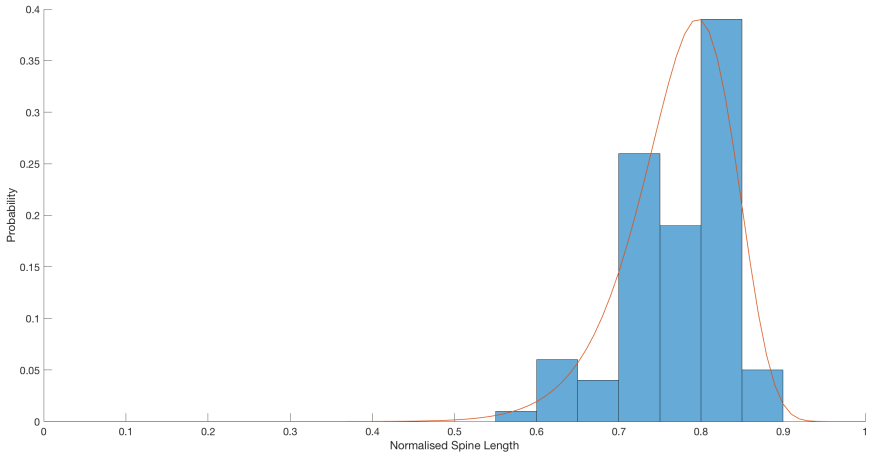


Figure 6.16: The distribution of the lengths of spine envelopes as a proportion of abdominal length (blue), modelled by a Gaussian distribution (red).

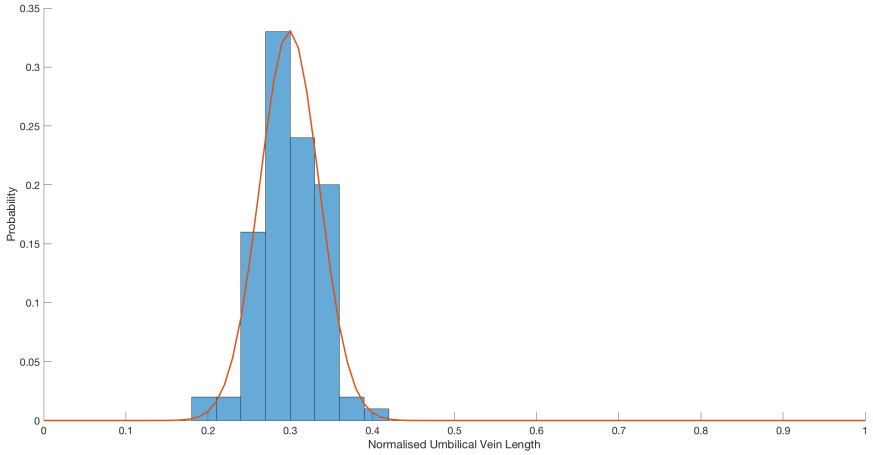


Figure 6.17: The distribution of the lengths of umbilical vein envelopes as a proportion of abdominal length (blue), modelled by a Gaussian distribution (red).

a similar appearance to the umbilical vein, confirming the hypothesis and design decisions discussed in Section 6.4.2. These findings suggest that the introduction of an additional geometric constraint on the length of candidate umbilical vein bounding envelopes may decrease the false positive detection of dark, umbilical vein-like artefacts.

Similarly the distributions of optical flow, in the z-direction along the longitudinal axis of the fetal abdomen, of ground truth 3-D fetal umbilical vein envelopes were compared against those of 3-D umbilical vein-like structures.

Landmark	Length	Optical Flow
Umbilical veins	0.29 ± 0.11	0.058 ± 0.009
Artefacts	0.12 ± 0.07	0.061 ± 0.012

Table 6.12: The mean and standard deviations of the lengths and mean optical flow of ground truth umbilical vein envelopes and umbilical vein-like artefact envelopes.

The Gaussian distribution parameters of mean optical flow magnitude $\sqrt{(u^2+v^2)}$ across the 3-D umbilical vein and umbilical vein-like artefact envelopes respectively are shown in Table 6.12. A two-sample t-test ($t(298) = 0.89, p = 0.38$ demonstrated that the mean optical flow profiles of ground truth spine envelopes did not significantly differ from those of artefacts displaying a similar appearance to the umbilical vein.

Dynamic Programming

Having established the feasibility of 3-D optical flow and length based constraints across candidate spine and umbilical vein envelopes, a modified 3-D framework for standardised abdominal plane selection was implemented.

Candidate bounding boxes $DT_{\mathrm{ab}\,r,f,n}$, $DT_{\mathrm{st}\,r,f,n}$, $DT_{\mathrm{uv}\,r,f,n}$, $DT_{\mathrm{sp}\,r,f,n}$ were detected for the abdominal wall, stomach bubble, umbilical vein and spine respectively for all volumes in the testing set $V_{B\,n}$.

The highest confidence abdominal bounding boxes on each frame were grouped to form the 3-D abdominal bounding envelope $DT^{*3D}_{\mathrm{ab}\,n}$. Candidate stomach bubble, umbilical vein and spine bounding boxes with centre co-ordinates not falling within $DT^{*3D}_{\mathrm{ab}\,n}$ were discarded. Remaining candidate umbilical vein and spine bounding boxes on consecutive frames displaying a Dice overlap coefficient greater than 0.75 were treated as representing the same anatomical structures and were indexed as members of the same 3-D bounding envelope, resulting in a set of envelopes $DT^{*3D}_{\mathrm{uv}\,n,p}$ and $DT^{*3D}_{\mathrm{sp}\,n,p}$ for the umbilical vein and spine respectively.

Each candidate umbilical vein bounding box $DT_{\mathrm{uv}\,r,f,n}$ and candidate spine bounding box $DT_{\mathrm{sp}\,r,f,n}$ was therefore associated with a prior probability based

on the length of its corresponding 3-D envelope as a proportion of detected abdominal envelope length. Similarly each candidate spine bounding box was associated with a prior probability based on the mean optical flow across its corresponding 3-D envelope.

The standardised abdominal plane was then selected as that which maximised the posterior probability ($p(L|V_{B\,n,f},\theta) \propto \Pi_{i=1}^{n} p(V_{B\,n,f}|l_i,u_i)\Pi_{v_i,v_j} p(l_i,l_j|cij)$) according to prior distributions for bounding box confidence scores, displacements between pairs of parts, spine and umbilical vein envelope lengths as a proportion of abdominal envelope length, and spine envelope optical flow where $p(V_{B\,n,f}|l_i,u_i)$ is now the prior distribution of bounding box confidence scores and the prior distributions of spine envelope optical flow, and spine length and umbilical vein length as proportions of abdominal length, and $p(l_i,l_j|c_{ij})$ are the prior joint distributions of displacements between pairs of parts given a particular volume frame $V_{B\,n,f}$.

An exhaustive search algorithm to find the optimal configuration of parts based on the maximum posterior probability runs with $O(h^n)$ complexity, where $h = 3$ and corresponds to the number of parts, and n is the number of bounding boxes for each part. This 3-D pictorial structures model was optimised for speed through the implementation of a dynamic programming algorithm.

Treating this model as a graph $G = (V, E)$ with vertices $V = v_1, ..., v_n$ corresponding to parts and connections $(v_i, v_j) \in E$ between parts, and a particular configuration of parts denoted by L, it is possible to express the prior distributions of part displacements as $L^* = argmax_L P(L|V_{B\,n,f})$ or, using Bayes' Rule, as $L^* = argmax_L P(V_{B\,n,f}|L)P(L)$.

$P(L)$ is equivalent to the prior joint distributions of displacements between pairs of parts described above, and may be written as $P(L) = \frac{1}{K} e^{-\Sigma_{(v_i,v_j)\in E} d_{ij}(l_i,l_j)}$.

Here, d_{ij} is the deformation cost between two parts (l_i, l_j) and K is a constant. $P(V_{B\,n,f}|L)$ is equivalent to the prior distribution of bounding box confidence scores, and may be written as $P(V_{B\,n,f}|L) \propto \Pi_{i=1}^{n} g_i(V_{B\,n,f}, l_i)$, where $g_i(V_{B\,n,f}, l_i)$ is the confidence score for a particular part at position l_i.

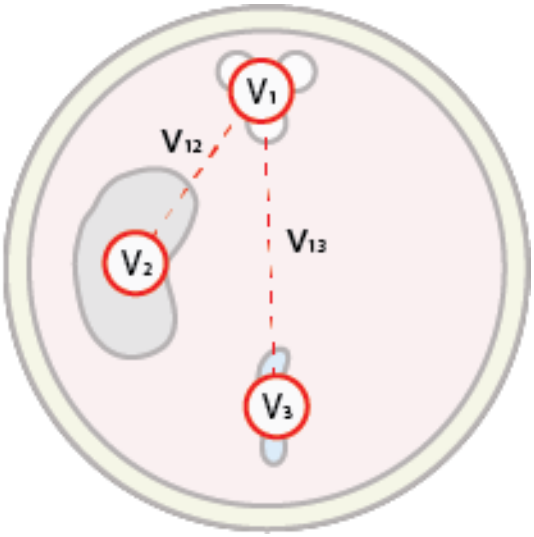


Figure 6.18: The pictorial structures model represented as a graph, specifically a depth-one tree with root vertex V_1 (spine) and leaf vertices V_2 (umbilical vein) and V_3 (stomach bubble) joined by edges $v_1 2$ and $v_1 3$.

Substituting Equations $P(L)$ and $P(V_{B\,n,f}|L)$ into L^*, the energy minimisation function is $L^* = argmin_L(\Sigma_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) - \Sigma_{v_i \in V} g_i(V_{B\,n,f}, l_i))$.

Treating this particular configuration of parts as a depth-one tree structure (Figure 6.18) with three vertices and two edges, it is then possible to find the optimal configuration of parts through a dynamic programming algorithm which runs with $O(nh^2)$ complexity, where $h = 3$ and corresponds to the number of parts, and n is the number of candidate bounding boxes for each part.

Taking the spine as the root vertex $v_r \in V$ at depth 0, some candidate spine bounding boxes are selected and the stomach and umbilical vein bounding boxes are found which maximise the posterior probability. The optimal locations of these leaf vertices can be computed as functions of the root vertex alone ($B_j(l_i) = \min_{l_j}(d_{ij}(l_i, l_j) + g_j(V_{B\,n,f}, l_j))$ where B_j gives the best location for a vertex v_j, given the best location for the root vertex v_i is l_i, and is dependent only on the displacement term $d_{ij}(l_i, l_j)$ and the appearance term $g_j(V_{B\,n,f}, l_j))$. This process is repeated for all candidate spine bounding boxes to find the optimal configuration of bounding boxes L^*. The frame $V_{B\,max\,n}$ which maximised the posterior probability across all frames f of a volume was selected as the standardised abdominal plane.

Testing Data

The model was tested on dataset $V_{B\,n}$ as described in Chapter 4. As previously, a correct standardised abdominal plane detection was defined as one for which the volume frame selected by the 2-D pictorial structures model corresponded to one of the manually labelled ground truth frames such that $V_{B\,max\,n} \in (V^*_{B\,\text{plane}\,n,f} = 1)$.

Individual detection accuracies for the abdomen, stomach bubble, umbilical vein and spine were defined as the mean proportion of detected bounding boxes across each volume for which the Dice overlap coefficient with the corresponding ground truth bounding box was greater than 0.75.

6.5.2 Results

The standardised plane selection accuracy across $V_{B\,n}$ was 92.7%, improving on Rahmatullah's benchmark accuracy of 91.29%[7] and the 2-D pictorial structures model accuracy of 83.3% as reported in Section 6.4.2. The framework displayed a non-optimised mean run-time of 15.4s per volume, on a MacBook Pro 2.8GHz Intel Core i7 processor, with 16GB 1600MHz DDR3 RAM.

Standalone mean detection accuracies for the abdomen, stomach bubble, umbilical vein and spine are shown in Table 6.13 following the application of the 3-D optical flow and length based geometric constraints. The confusion matrix in Figure 6.19 demonstrates that 8.61% and 23.7% of structures classified as the umbilical vein and spine respectively were misclassified artefacts, an improvement on the misclassification rates reported in Section 6.4.2 prior to the introduction of 3-D constraints.

Figure 6.20 shows the ROC curve produced by applying a varying posterior probability threshold between 0 and 1 across $V_{B\,n}$ and computing the sensitivity and specificity at each threshold value, resulting in a mean AUC of 0.82. Correctly identified standardised abdominal planes fell into two categories; those where the stomach bubble, umbilical vein and spine were correctly localised as shown in Figure 6.10 (95% of correctly selected planes), and those where some or all of these landmarks were incorrectly localised (5% of correctly selected planes) as

Landmark	Benchmark	2-D Pictorial Structures Model	3-D Pictorial Structures Model
Stomach bubble	78.94	87.43	88.70
Umbilical vein	62.80	82.40	85.67
Spine	-	71.64	76.30

Table 6.13: The accuracy in stomach bubble, umbilical vein and spine detection via a 3-D pictorial structures model including optical flow and length based constraints on the spine and umbilical vein. Benchmark stomach bubble and umbilical vein accuracies as reported by Rahmatullah et al.[6], and 2-D pictorial structure accuracies as reported in Section 6.4.2, are also shown for reference.

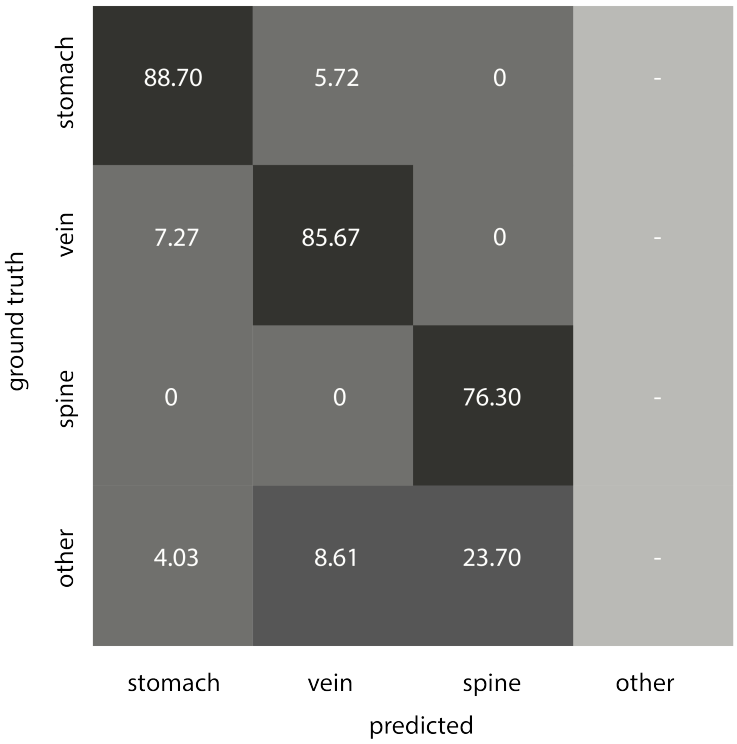


Figure 6.19: A confusion matrix showing percentage misclassifications of the 3-D pictorial structures model between the stomach bubble, umbilical vein, spine and other structures including artefacts or the fetal ribs.

shown in Figure 6.11. All instances of incorrect plane selection were attributable to incorrect anatomical landmark localisation.

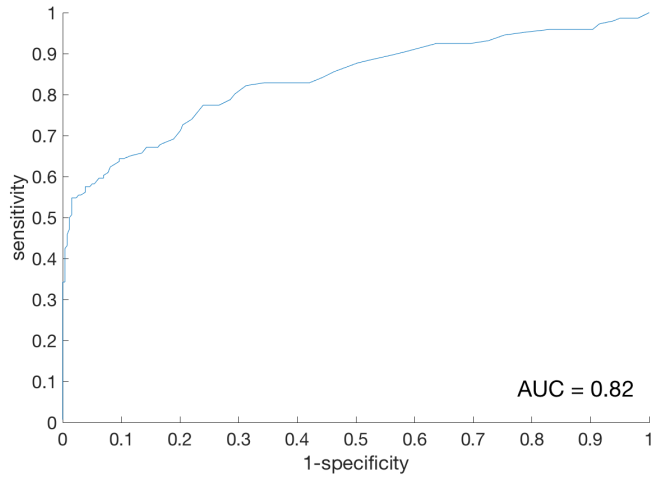


Figure 6.20: ROC curve showing the sensitivity, specificity and accuracy (AUC) of a 2-D pictorial structures model in standardised abdominal plane selection, with 3-D geometric priors.

6.5.3 Discussion

It has been demonstrated that a pictorial structures model incorporating 3-D constraints on the length and position of candidate spine bounding boxes, and the length of candidate umbilical vein bounding boxes, produces higher accuracies in standardised plane selection from 3-D fetal abdominal US volumes than a 2-D pictorial structures model incorporating only 2-D geometric constraints.

In particular it has been shown that the introduction of 3-D constraints leads to a reduction in the misclassification of bright spine-like artefacts and structures as the spine, and the misclassification of dark artefacts and structures as the umbilical vein. This results in a standardised abdominal plane selection framework which performs with a higher accuracy than the benchmark accuracies of Rahmatullah [7].

I, for one, welcome our new machine learning overlords

— Empire of the Ants, H. G. Wells, (1905)

7

Learning a Spatio-Temporal Interest Point Operator from Fixations for Video Classification

Contents

7.1 Introduction

In clinical practise, sonographers perform routine biometric US scans through the manipulation and movement of the US probe until standardised views of the fetus are visible[121]. Therefore the automated identification of structures of interest, for example the fetal abdomen and head, in 2-D+t video clips of fetal US scans is a crucial step in fully automating the acquisition of standardised views of the fetus. Maraci's[51] approach to this problem employed a bag-of-visual-words (BoVW) model and SIFT descriptors, modelling the temporal evolution of anatomical structures of interest in US videos as a linear dynamical system. However as discussed in Chapter 2, points identified by SIFT descriptors and local maxima of other hand crafted features may not necessarily correspond to anatomically salient regions[25–29].

In this chapter, this challenge is approached from an eye-tracking perspective, developing a biologically inspired spatio-temporal interest point operator and a BoVW framework for 2-D+t fetal US video clip classification.

Specifically, a series of eye tracking experiments were conducted whereby ten observers viewed a set of 60 2-D+t fetal US video clips, each showing one portion of the fetal anatomy, and verbally classified the clip as 'head', 'abdomen' or 'other'. A low correspondence was found between the fixations of observers on these video clips and the points identified by the Harris[13] and Periodic[14] spatio-temporal interest point operators, suggesting that the regions fixated by observers may indeed be more anatomically meaningful than those identified by conventional spatio-temporal operators.

Building on the work of Kienzle[12], a learned saliency function was developed by fitting the weights of a series of spatial and temporal filters to the fixations of observers. When used to classify likely fixation points on unseen video clips, this saliency function acted as a biologically inspired spatio-temporal interest point operator and was used within a BoVW framework for automated video clip classification.

This work constitutes the development of the first interest point operator learned directly from human fixations on US images, or on medical images more broadly.

7.2 Originality and Individual Role

Independently, I designed and wrote Python applications to interface with eye tracking hardware, acquired and post-processed eye movements, analysed the resulting data, and computed Harris and Periodic spatio-temporal interest points. I independently trained and tested a feed-forward neural network to produce a perception inspired spatio-temporal operator. BoVW frameworks were implemented using helper functions from an open source Matlab library[115].

7.3 Eye Tracking

This section describes a series of eye tracking experiments conducted to determine which regions of 2-D+t fetal US video clips are most visually salient, and whether spatio-temporal interest points identified by the Harris and Periodic operators correspond to visually salient regions.

7.3.1 Methods

Stimuli

The eye tracking stimuli consisted of dataset $C_{A\,n}$ as described in Chapter 4. Stimuli were presented to observers as described in Chapter 5.

Hardware

Eye movements were recorded using an EyeTribe eye tracker and a custom GUI as described in Chapter 5.

Participants

Ten observers comprised of three clinical medical undergraduate students, three Biomedical Engineers with experience in US image analysis, and four Mathematical Physical and Life Sciences undergraduate students, all with normal acuity, participated in this study.

Experimental Procedure

Calibration was carried out for each observer prior to stimulus presentation, as described in Chapter 5.

Each observer was then presented with each US video clip, played once at $14fps$ to ensure a sufficient number of gaze co-ordinates would be recorded for each frame. The videos were presented in a pre-determined randomised sequence, and observers were instructed to verbally report whether the video showed the fetal head, abdomen, or other part of the fetal anatomy before proceeding to the next video using the mouse button. This experiment was designed to replicate the task, routinely performed by sonographers, of identifying different portions of the fetal anatomy in real-time during an US scanning session in order to obtain standardised views of the fetus. The experiment deviated from clinical practise as observers were not required to perform a fetal US scan or manipulate a US probe whilst intepreting video footage. This may have affected observers' visual search strategies, however the video clip stimuli were deemed to be a sufficient analogue for the US footage viewed by sonographers during a clinical US scan. A Gaussian white noise image was displayed for $5s$ between each video to de-focus the observer's gaze as described in Chapter 5. To avoid fatigue, observers were given a ten minute break between each group of ten videos. Ethics approval for this procedure was obtained via the University of Oxford Central University Research Ethics Committee (Reference: MS-IDREC-C1-2015-166).

This resulted in a set of raw gaze co-ordinates, where the j^{th} observer's gaze data on the n^{th} video was described by $R_{n,j}$ and each set of gaze co-ordinates $R(x, y, t_{vid}, d_{\text{os}}, t)$ consisted of the mean x and y co-ordinates across the left and right pupils, t_{vid}, the index of the video frame currently displayed on the screen, the observer-to-screen distance d_{os} and a timestamp t.

Fixation Filtering

Raw gaze co-ordinates for each observer and each volume frame were filtered into fixations and saccades using the I-VT algorithm described in Chapter 5[97].

7.3.2 Regions of Interest

The proportions of fixations falling within the ground truth bounding boxes described in Chapter 4 ($C^*_{\mathrm{ROI}A\,m,n}$ with $m \in 1, 2$ corresponding to ground truth bounding boxes for the head cavity and abdominal cavity respectively) were computed for each video where present.

7.3.3 Interest Point Operator Comparison

The fixated locations of observers on these video clips were compared to the points identified by two established methods for spatio-temporal interest point detection[12], the Harris and Periodic operators, to establish whether fixations fall on more anatomically meaningful regions than conventional interest point operators.

Spatio-Temporal Harris Operator

The spatio-temporal Harris operator[13] is an extension of the 2-D Harris corner operator[122], identifying salient points in videos based on intensity gradients in spatial and temporal directions. Each video was treated as a function of time $I(x, y, t)$. At each pixel, derivatives in spatial and temporal directions are obtained by convolution with the derivative of a Gaussian kernel at spatial scale σ and temporal scale τ, giving the local gradient distribution around each pixel ($I_x = G_x(\sigma) * I\ I_y = G_y(\sigma) * I\ I_t = G_t(\tau) * I$).

The 3×3 second moment matrix, or structure tensor, is then given at each pixel by:

$$M = \begin{bmatrix} I_x^2 & I_x I_y & I_x I_t \\ I_y I_x & I_y^2 & I_y I_t \\ I_t I_x & I_t I_y & I_t^2 \end{bmatrix} \tag{7.1}$$

The second moment matrix is then integrated at each pixel over a spatio-temporal window at larger scales of $\sigma' = 2\sigma$, $\tau' = 2\tau$ and $\sigma = 15px$ in accordance with the parameters used by Kienzle[12] ($S_x = G(x, \sigma') * I_x\, S_y = G(y, \sigma') * I_y\, S_t = G(t, \tau') * I_t$) and the Hessian matrix is then given by:

$$H = \begin{bmatrix} S_x^2 & S_x S_y & S_x S_t \\ S_y S_x & S_y^2 & S_y S_t \\ S_t S_x & S_t S_y & S_t^2 \end{bmatrix} \tag{7.2}$$

Interest points are found by computing the local maxima of the saliency function $S = det(H) - k(trace(H))^3$ where $k = 0.005$ is an empirical constant.

Non-maximal suppression was used to select pixels with an operator response greater than or equal to all other pixels in a surrounding window of diameter $30px$. This radius corresponds to a visual angle of 1.5°, chosen to match the foveated area of an observer seated $0.5m$ from a 25-inch 1920×1080 pixel LCD monitor[9]. The 13 strongest interest points in each frame were selected, corresponding to the mean number of fixations on each stimulus video frame across all observers.

Spatio-Temporal Periodic Operator

The spatio-temporal Periodic interest point operator proposed by Dollar[14] detects video regions with periodic intensity variations in the temporal dimension, which may correspond to the appearance and occlusion of anatomical landmarks with time in the US video dataset. The response of the operator is defined by $R = (I * G(x, y, \sigma) * h_{even})^2 + (I * G(x, y, \sigma) * h_{odd})^2$, where $G(x, y, \sigma)$ is a 2-D Gaussian kernel applied spatially, and a quadrature pair of 1-D Gabor filters $h_{even} = (t; \tau, \omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{odd} = (t; \tau, \omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2}$ are applied temporally where $\sigma = 15px$, $\tau = 3px$ and $\omega = 0.5/\tau$ in accordance with the parameters set by Kienzle[12].

Non-maximal suppression was used to select pixels with a response value greater than or equal to all pixels in a surrounding window of diameter $30px$ and the 13 strongest interest points in each frame were selected.

Similarity Score

Spatio-temporal Harris and Periodic interest points were compared with fixation points. For each video frame in each video, a binary fixation map was generated for the fixations of all observers on that frame; initial pixel values were set to 0, and pixel values at fixation points were incremented by 1. As in Chapter 5, the binary maps

were convolved with a 2-D Gaussian kernel of mean 0 and standard deviation of $15px$ (corresponding to a visual angle of 1.5°), resulting in a set of maps representing the visual attention of all observers on each frame of each stimulus video.

The similarity between fixation points and Harris and Periodic points was found by computing the accuracy of each Harris and Periodic map in predicting the corresponding fixation map for each video frame. A varying threshold between 0 and 1 was applied to binarize each fixation map; at each threshold level, the sensitivities and specificities of the Harris map and the Periodic map in predicting the current fixation map were recorded, resulting in two ROC curves with the two AUCs taken as the similarity scores. This process was repeated across all video frames and all videos to compute a mean fixation-Harris similarity score and a mean fixation-Periodic similarity score.

A cross-image control measure was calculated to assess how accurately the summed fixation maps of all observers on one randomly selected video frame could be predicted by Harris and Periodic interest points on a different randomly selected frame. Repeated for 50 randomly generated pairs of frames, the mean cross-image similarity scores for each of the Harris and Periodic operators acted as baseline similarity scores. Due to biases in the stimuli, such as the tendency for anatomically significant structures to appear in the centre of video frames, and the tendency to initially fixate on the centre of images, the control similarity score was expected to be greater than random chance (50%).

7.3.4 Results

The proportions of fixations falling within anatomical ROIs are shown in Figures 7.2 and 7.3 and Table 7.1. The majority of fixations and Periodic interest points on head and abdomen video frames fell within the fetal skull and abdominal cavity respectively, whereas the majority of Harris interest points fell outside these anatomical ROIs. The calculation of similarity scores between fixations and spatio-temporal Harris and Periodic operators are shown in Figures 7.4 and 7.5 respectively,

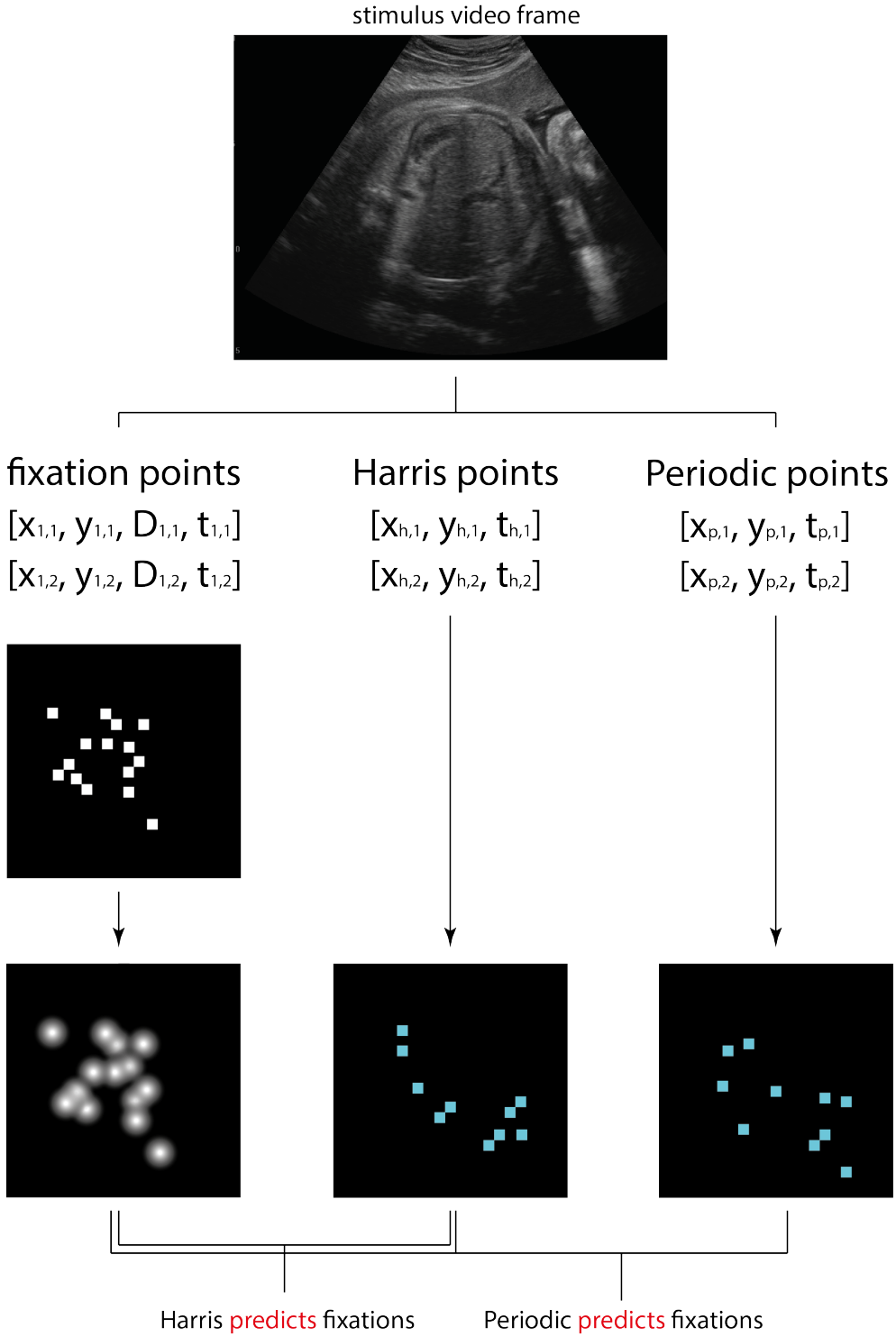


Figure 7.1: Schematic illustrating the comparison of the fixation points of all observers and detected spatio-temporal Harris and Periodic interest points on individual video frames.

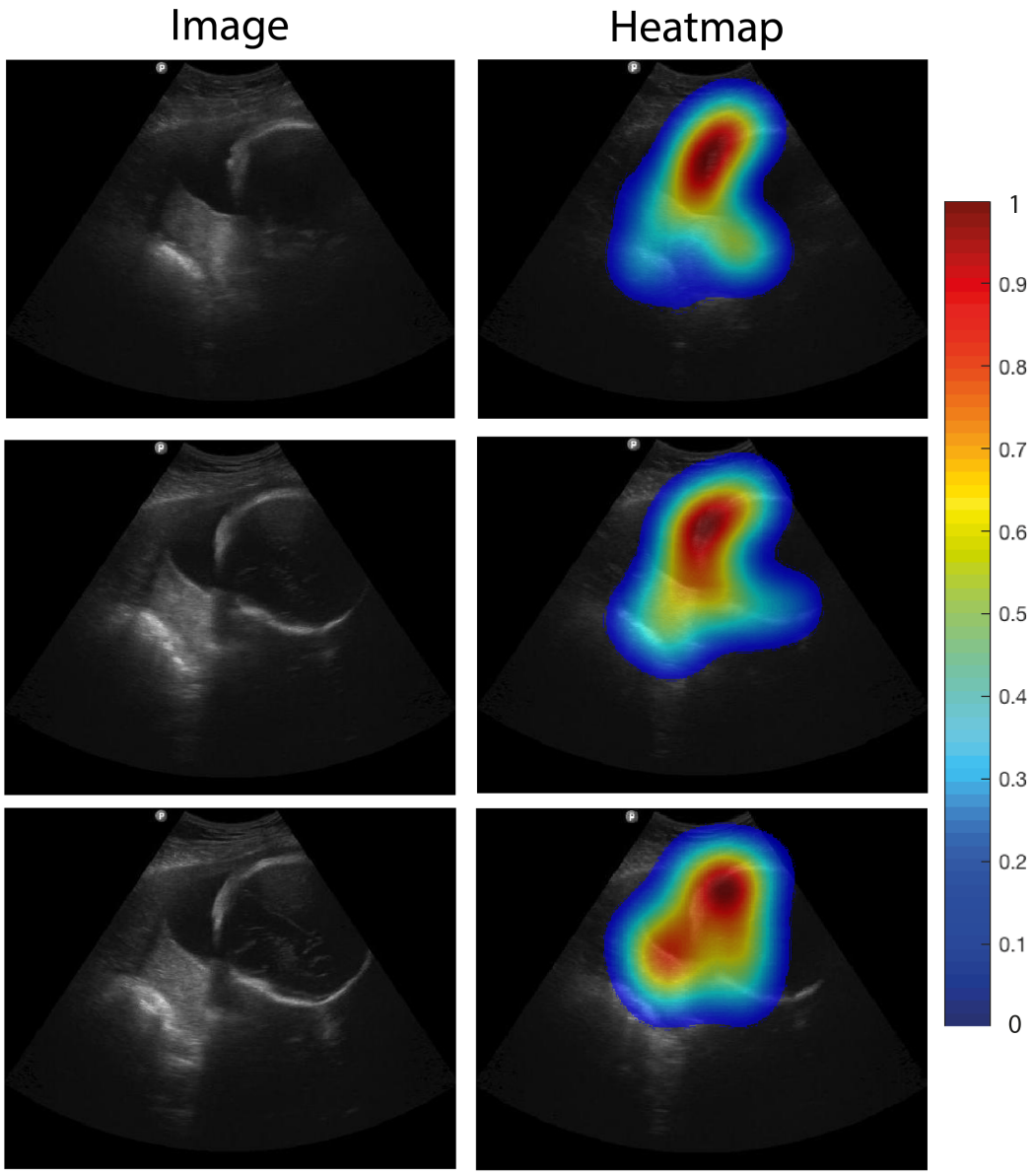


Figure 7.2: Examples of heat maps (right) showing fixations on anatomical ROIs on 2-D+t US video frames (left) of the fetal head.

.

and in Table 7.2. The Harris operator was the weaker predictor of fixation points across all frame types, with fixations on head frames the least accurately predicted by both interest point operators, and fixations on abdomen frames most accurately predicted by both interest point operators. In all cases, mean similarity scores were greater than the random cross-image control.

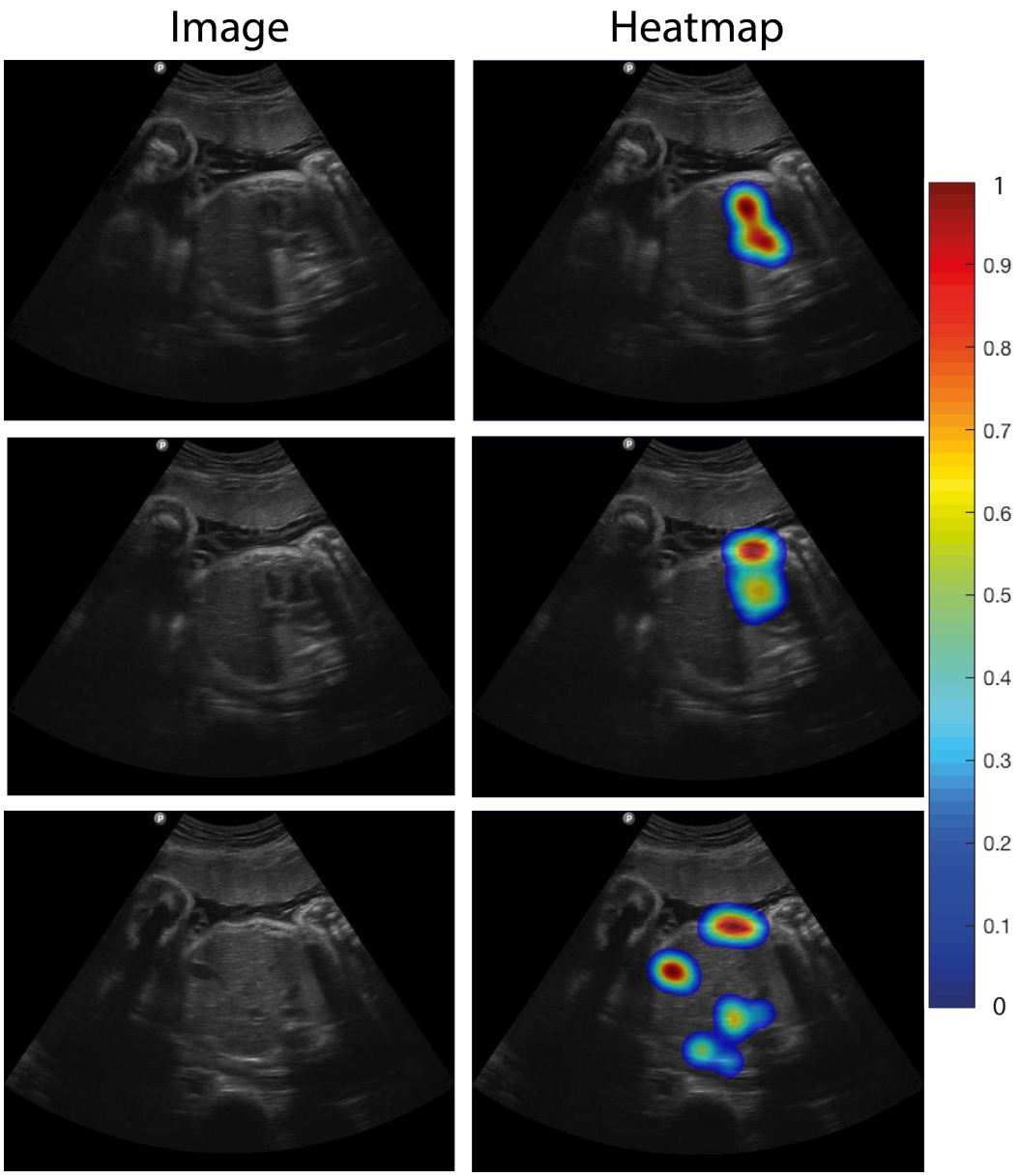


Figure 7.3: Examples of heat maps (right) showing fixations on anatomical ROIs on 2-D+t US video frames (left) of the fetal abdomen.

.

	Head (%)	**Abdomen (%)**
Fixations	78.3	74.8
Periodic	67.0	72.1
Harris	31.5	24.2

Table 7.1: The percentages of fixations, Harris and Periodic interest points falling within anatomical ROIs in video frames showing the fetal head and abdomen.

	Harris	**Periodic**
Head	0.51	0.53
Abdomen	0.63	0.64
Other	0.52	0.54
Control	0.51	0.53

Table 7.2: The mean similarity scores of the Harris and Periodic spatio-temporal interest point operators with respect to fixation points. The random cross-image baseline measure is also shown.

7.3.5 Discussion

It has been demonstrated, through ROI analysis, that fixations tend to fall on more anatomically significant regions than both Harris and Periodic interest points. A low correspondence has also been found between visually salient points in 2-D+t US video footage and points identified by spatio-temporal Harris and Periodic operators.

The lower mean AUC of the Harris operator in predicting fixations may be attributable to the identification of bright and dark artefacts as spatio-temporal corners by the Harris saliency function, despite these regions not being sufficiently anatomically significant to attract the gaze of observers. The other limitation of the spatio-temporal Harris operator is its treatment of the temporal dimension as equivalent to another spatial dimension. Whilst this may be a robust approach

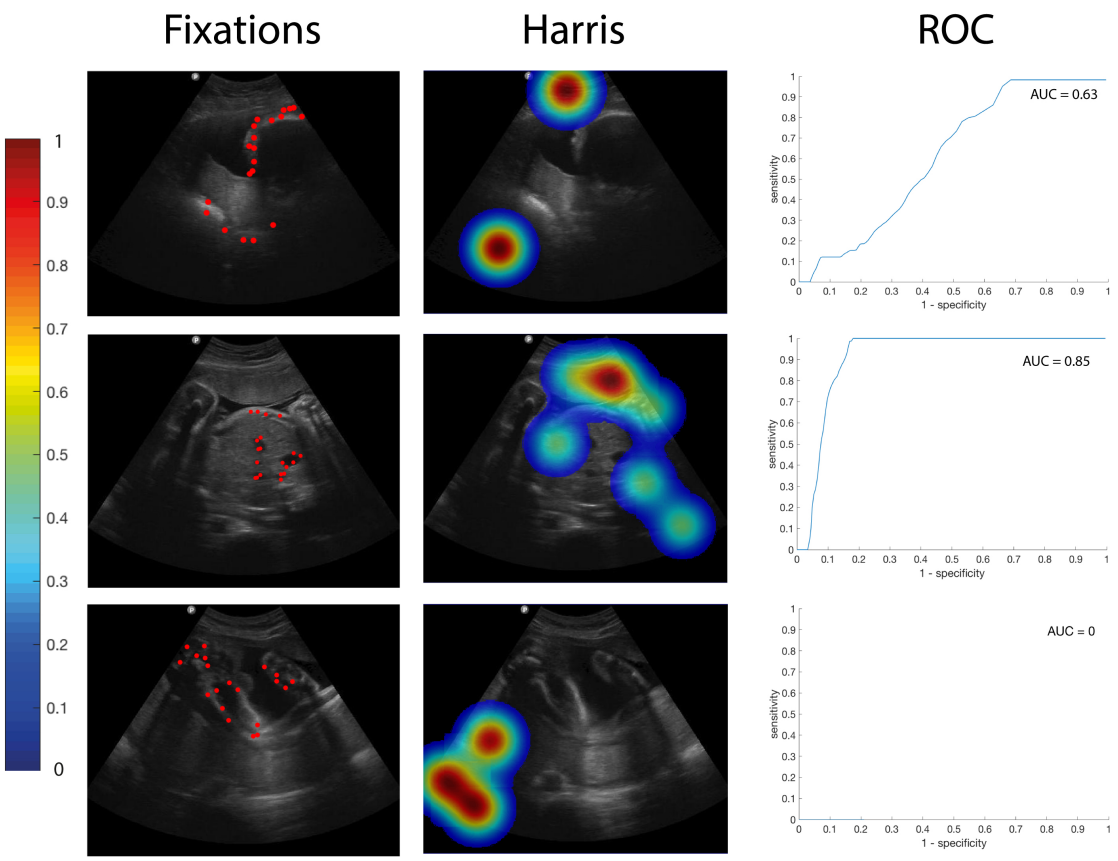


Figure 7.4: ROC curves and AUCs calculated for the accuracy of the spatio-temporal Harris operator in predicting fixation points on a (top) head (centre) abdomen (bottom) other video frame.

.

for identifying interest points in volume data and 3-D meshes[13], the temporal evolution of a structure is not equivalent to the spatial variations of a 3-D structure.

As demonstrated in Figure 7.4, in 41% of frames overall the Harris operator acted as a negative predictor of fixations. However for all frames, the Periodic operator was a positive, albeit weak, predictor of fixations, suggesting that observers have a tendency to fixate on regions which are explicitly not spatio-temporal corners but rather show periodic variations in intensity with time.

The higher mean AUC of the Periodic operator in predicting fixations is perhaps expected, as the representation of images by the human visual cortex has been shown to include both spatial and frequency components, and it has been demonstrated that visually salient image regions are well described by 1-D temporal Gabor filters[123,124].

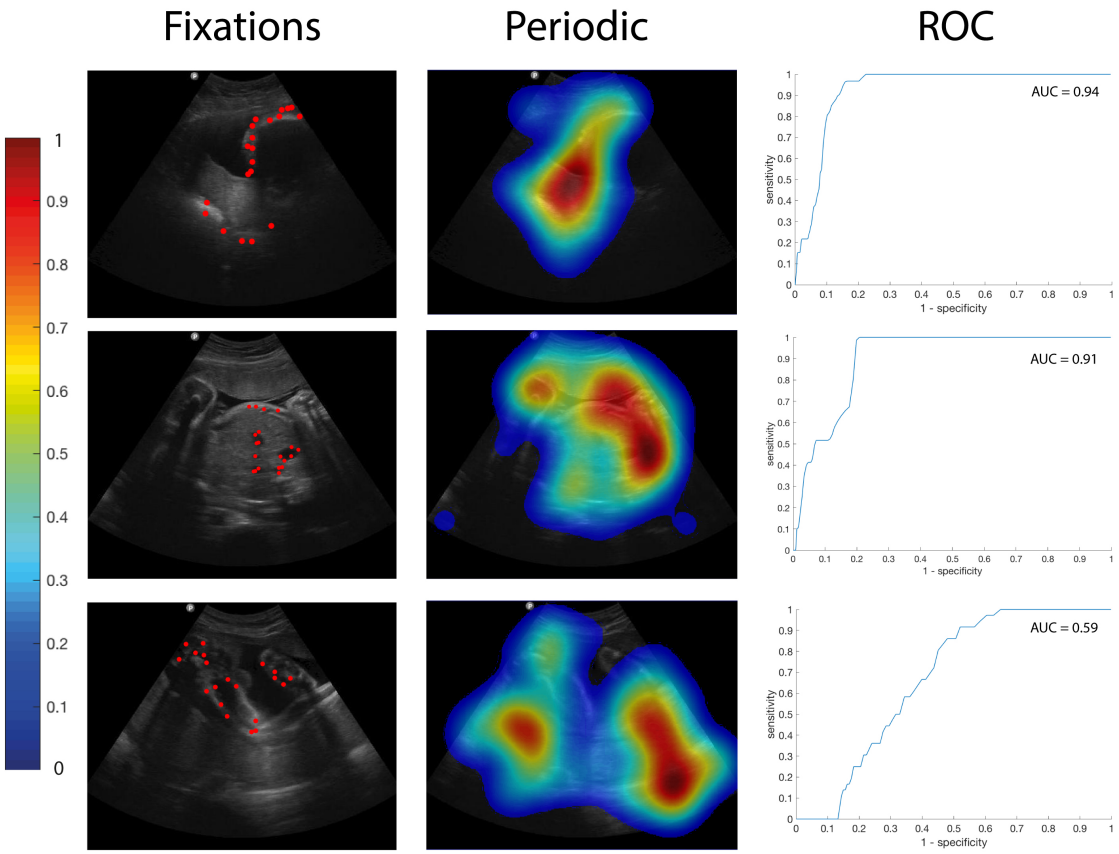


Figure 7.5: ROC curves and AUCs calculated for the accuracy of the spatio-temporal Periodic operator in predicting fixation points on a (top) head (centre) abdomen (bottom) other video frame.

.

7.4 Learning an Interest Point Operator

This section describes a generalised variant of the Periodic operator whereby the weights of a series of 1-D sigmoid basis functions, applied temporally to video stimuli, were learned to fit the fixation points of observers. It is hypothesised, based on the work of Kienzle[12], that fixation prediction accuracies may be improved by fitting the parameters of a saliency function, based on the Periodic operator, to the fixation points of observers. This approach is described next.

7.4.1 Methods

Training Data

The learned operator was trained using the dataset $C_{A1\,n}$ described in Chapter 4. The fixations of observers acted as labels. For each video frame, a fixation map was generated by summing the fixations of all observers, resulting in a binary map $B(x, y, t)$ for each video clip where fixated pixels had value 1 and all others had value 0. To obtain a ground truth map of visual saliency, or visual attention, the binary maps were convolved with Gaussian kernels of standard deviation $\sigma = 15px$ to mimic the foveated area of an observer seated $0.5m$ from a 25-inch 1920×1080 pixel LCD monitor[9].

Learned Saliency Function Parameters

The architecture of the learned saliency function was that of a feed-forward neural network with a sigmoid activation function ($S_L = b_0 + \Sigma_{j=1}^{m} tanh(\Sigma_{i=1}^{n} p_i * W_{i,j} + b_j)$) as described by Kienzle[12] and shown in Figure 7.6. This method differed significantly from that of the Periodic operator as convolution in the temporal direction is performed with a series of arbitrarily shaped temporal filters learned to fit the fixations of observers, rather than employing a pair of 1-D Gabor filters.

The saliency function was modelled as a multi-layer perceptron with n inputs, m hidden units, and 1 output as shown in Figure 7.6. Input training videos I were first convolved in the spatial x and y dimensions with a 2-D Gaussian kernel of standard deviation σ. Saliency values were then computed on a pixel-wise basis across input videos; for a given pixel of interest in the video $I(x, y, t)$, the values of 17 pixels preceding the pixel of interest in the temporal direction t such that the inputs to the neural network were defined by $p = I(x, y, t - n), I(x, y, t - n - 1), ..., I(x, y, t)$ and subsequently convolved with an arbitrarily shaped temporal filter, or weights,

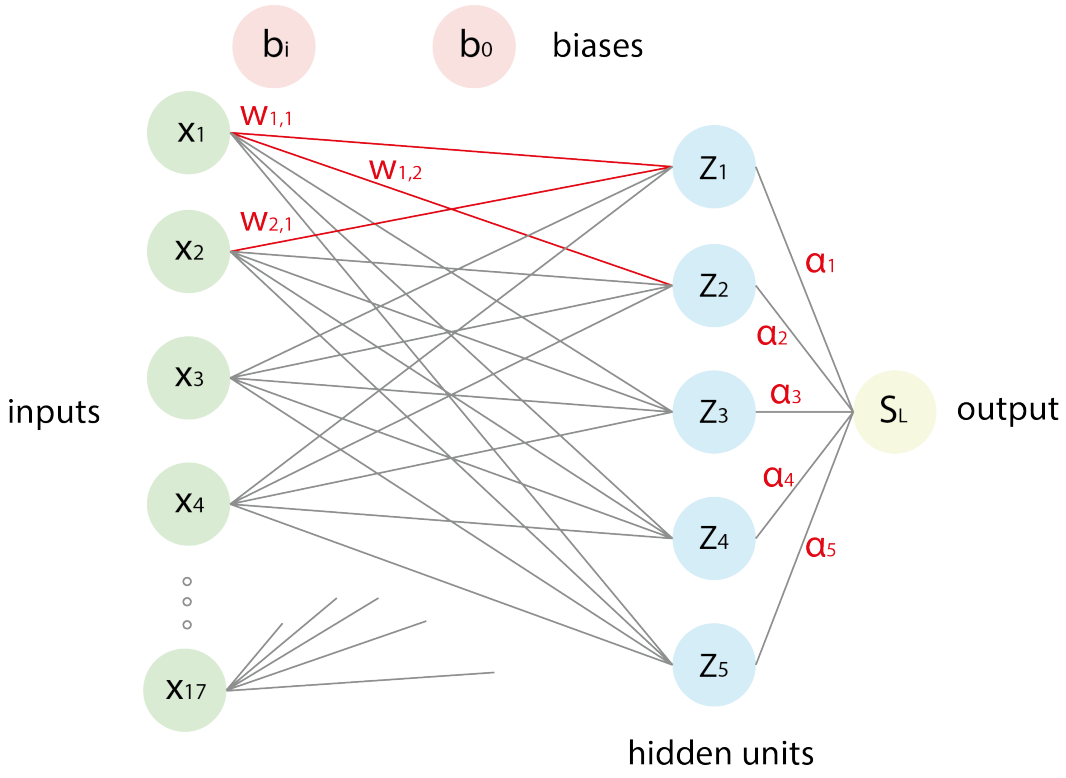


Figure 7.6: Schematic diagram of the feed-forward neural network classifier designed to act as a learned interest point operator. 17 input (green) intensities taken from consecutive video frames feed into 5 hidden units acting as temporal filters (blue) with learned weights (red) and a sigmoid activation function is applied to each unit. Filter outputs are summed and give a saliency value for a given pixel of interest.

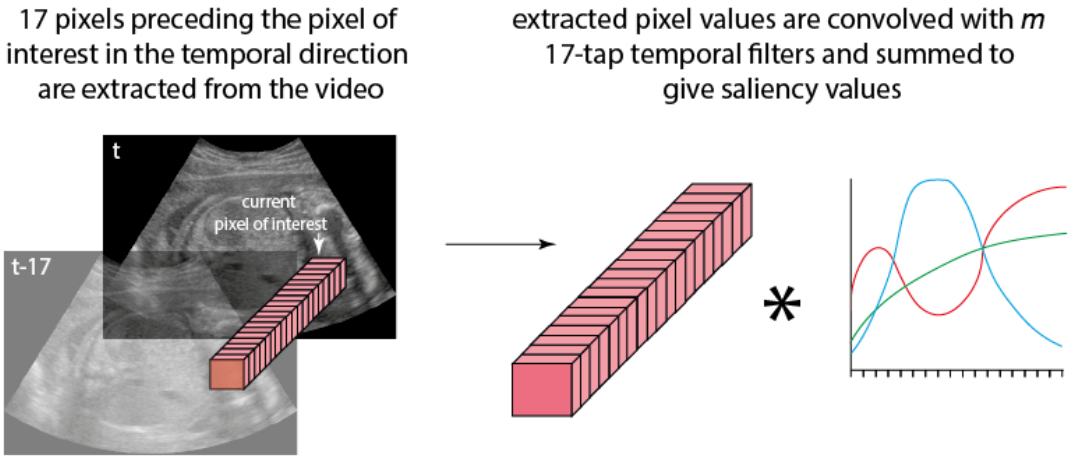


Figure 7.7: Schematic diagram demonstrating the application of temporal filters to consecutive pixels extracted from videos in the temporal direction to compute the saliency value of a given pixel of interest.

Number of Hidden Units	AUC
5	0.66
10	0.59
15	0.55
20	0.53

Table 7.3: The results of cross-validation to determine the optimal number of neural network hidden units for visual saliency prediction. The mean accuracies of predictive visual saliency maps in predicting ground truth visual saliency maps are shown for each number of hidden units.

W in the temporal direction with bias b_j. This was repeated over m summed components, and a global bias b_0 was added.

The size of the temporal filters, or number of inputs n, was set to 17 frames. This is equivalent to the number of frames viewed during a $607ms$ fixation, which was the mean fixation length across all observers during the experimental process outlined in Section 7.3.1. The number of hidden units, or summed temporal filters m, was optimised over a search space of 5 to 20 with a step size of 5 via 4-fold cross-validation on the training set to maximise fixation point prediction accuracy. As shown in Table 7.3, $m = 5$ produced predictive saliency maps with the greatest mean AUC in predicting ground truth saliency maps.

For a given pixel of interest the corresponding training label was taken from the ground truth visual saliency map $B(x, y, t)$. Backpropogation was used to optimise the weights α, temporal filters W and biases b_0 and b_j to fit the training labels through minimising a mean squared error function ($E_n = \frac{1}{2}\Sigma_k(S_{L,k}(x_n, W) - t_{nk})^2$ where the mean squared error for a particular set of inputs n, is the error between the network output S_L, dependent on input values x and weights W, and the corresponding targets $t_{n,k}$, summed across all k hidden units) via a gradient descent algorithm. The weights W were initially set to random values, and iteratively updated in the direction of the negative gradient of the error function until this gradient fell below 1×10^{-5} in accordance with the training threshold used by Kienzle[12].

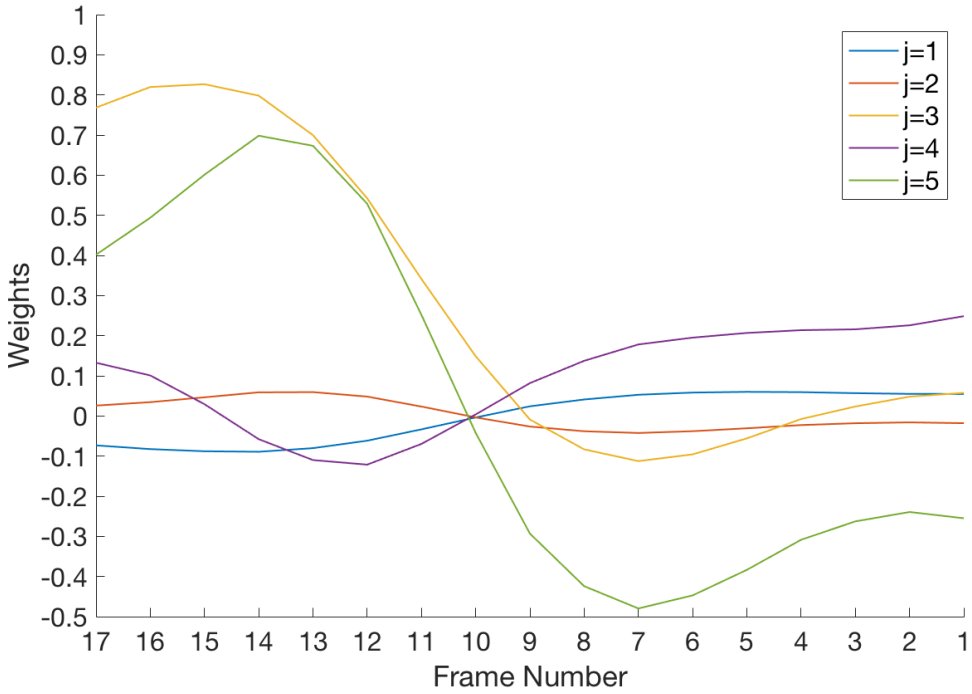


Figure 7.8: Learned weights for a feed-forward neural network trained to classify fixation points on unseen video frames. Weights are given for 17 inputs and a series of 5 hidden units (denoted by $j = 1, 2, ..., 5$), equating to 5 temporal filters applied and summed across 17 pixels in consecutive video frames.

.

Non-maximal suppression was used to select pixels with a response value greater than or equal to all pixels in a surrounding window of radius $15px$, and the 13 strongest interest points in each frame were selected as the final spatio-temporal interest points.

The learned weights for the 5 temporal filters applied to a particular pixel across the 17 frames preceding the pixel of interest are shown in Figure 7.8.

Testing Data

The learned operator was tested on the dataset $C_{A2\,n}$ as described in Chapter 4. The fixations of observers on each video frame acted as ground truth labels, with fixated pixels labelled with value 1 and all others with value 0.

Regions of Interest

The proportion of learned interest points falling within the ROIs described in Section 7.3.2, with respect to the total number of fixations on that frame, was recorded. The mean proportions of learned interest points on abdomen and head ROIs across all video frames were then computed.

Similarity Score

Interest points detected by the learned operator were compared with fixation points. For each video frame in the testing set, a binary fixation map was generated for the fixations of all observers on that frame; initial pixel values were set to 0, and pixel values at fixation points were incremented by 1. As in Section 7.3.3, the binary maps were convolved with a 2-D Gaussian kernel of mean 0 and standard deviation of $15px$ (corresponding to a visual angle of 1.5°), resulting in a set of maps representing the visual attention of all observers on each frame of each stimulus video.

A varying threshold between 0 and 1 was applied to binarize each fixation map, and the sensitivities and specificities of the learned operator in predicting the current fixation map were recorded, resulting in an ROC curves with the AUC taken as the similarity scores. This process was repeated across all testing frames to compute a mean similarity score, or accuracy, for the learned operator.

7.4.2 Results

The proportions of learned interest points falling within anatomical ROIs are shown in Figures 7.2 and 7.3 and Table 7.1. Learned detector response strengths compared to the Harris and Periodic detector response strengths across abdomen and head images are shown in Figures 7.9 and 7.10.

The calculation of similarity scores between fixations the learned operator is shown in Figure 7.4 and Table 7.2. The learned operator outperforms both the Harris and Periodic detectors on all frame types in predicting fixation points, with fixations on head frames the most accurately predicted and fixations on abdomen

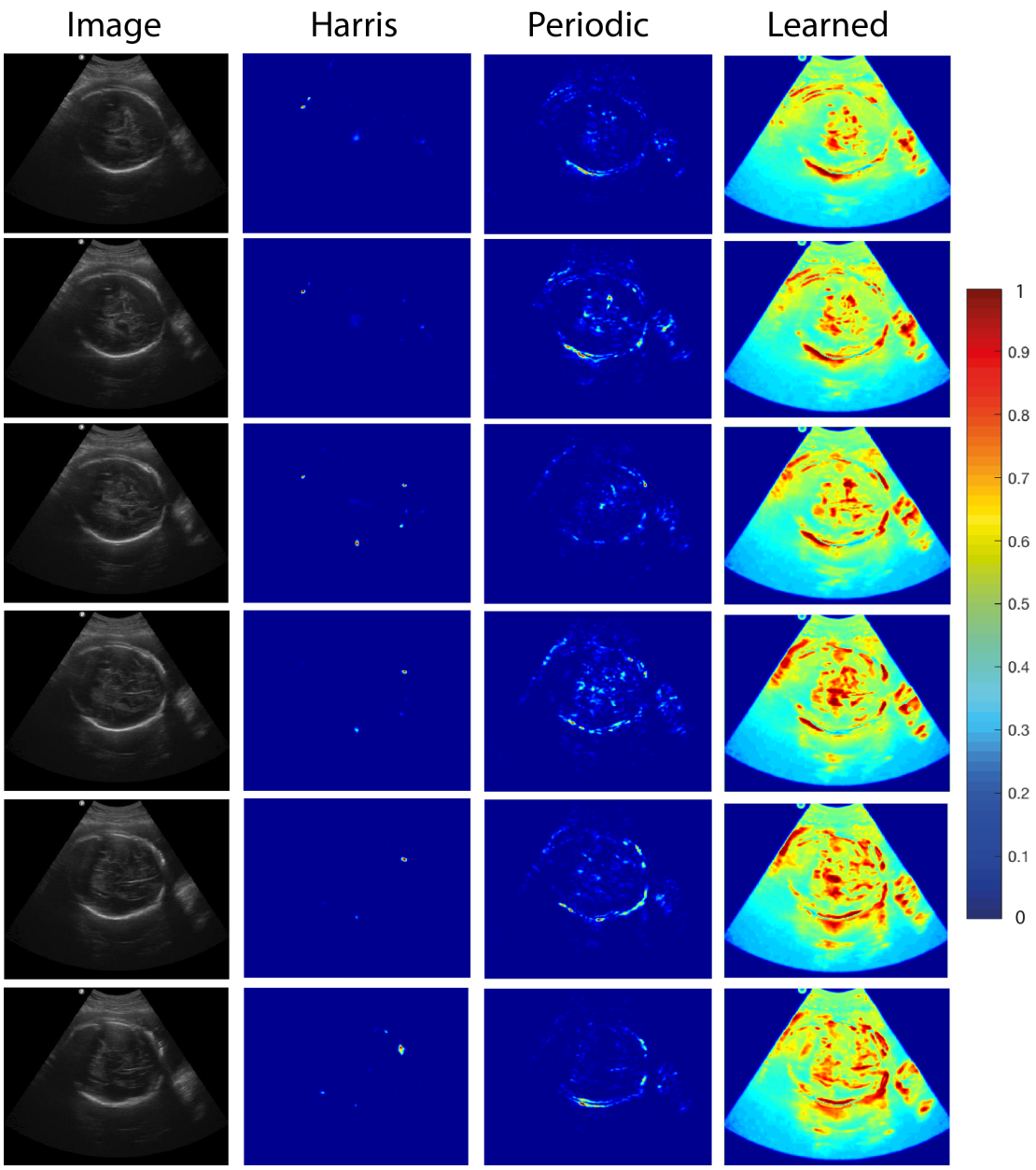


Figure 7.9: The response strength of the learned interest point operator across a sequence of images showing the fetal head.

.

frames most accurately predicted. Across all frame types, mean similarity scores were greater than the random cross-image control.

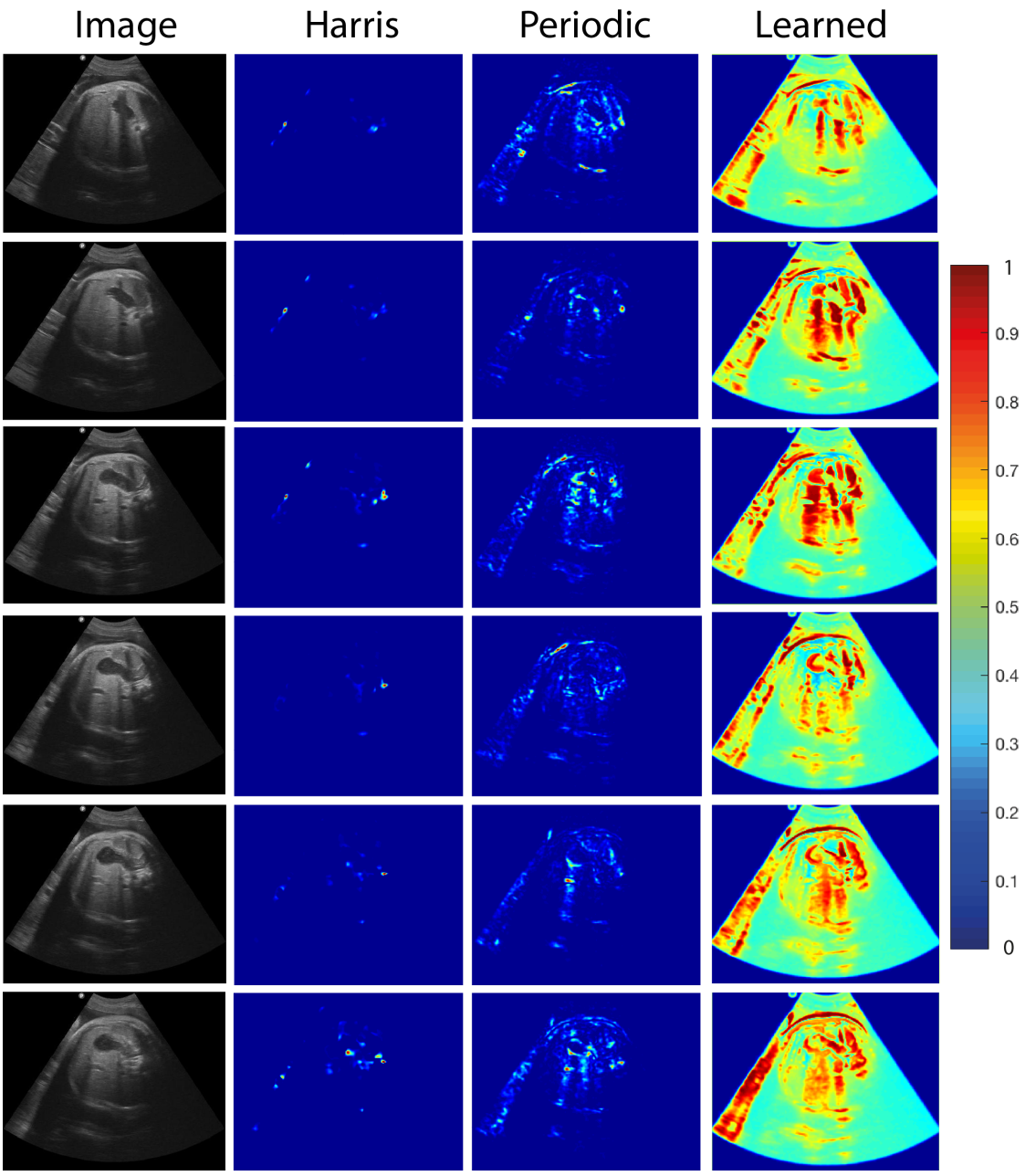


Figure 7.10: The response strength of the learned interest point operator across a sequence of images showing the fetal abdomen.

.

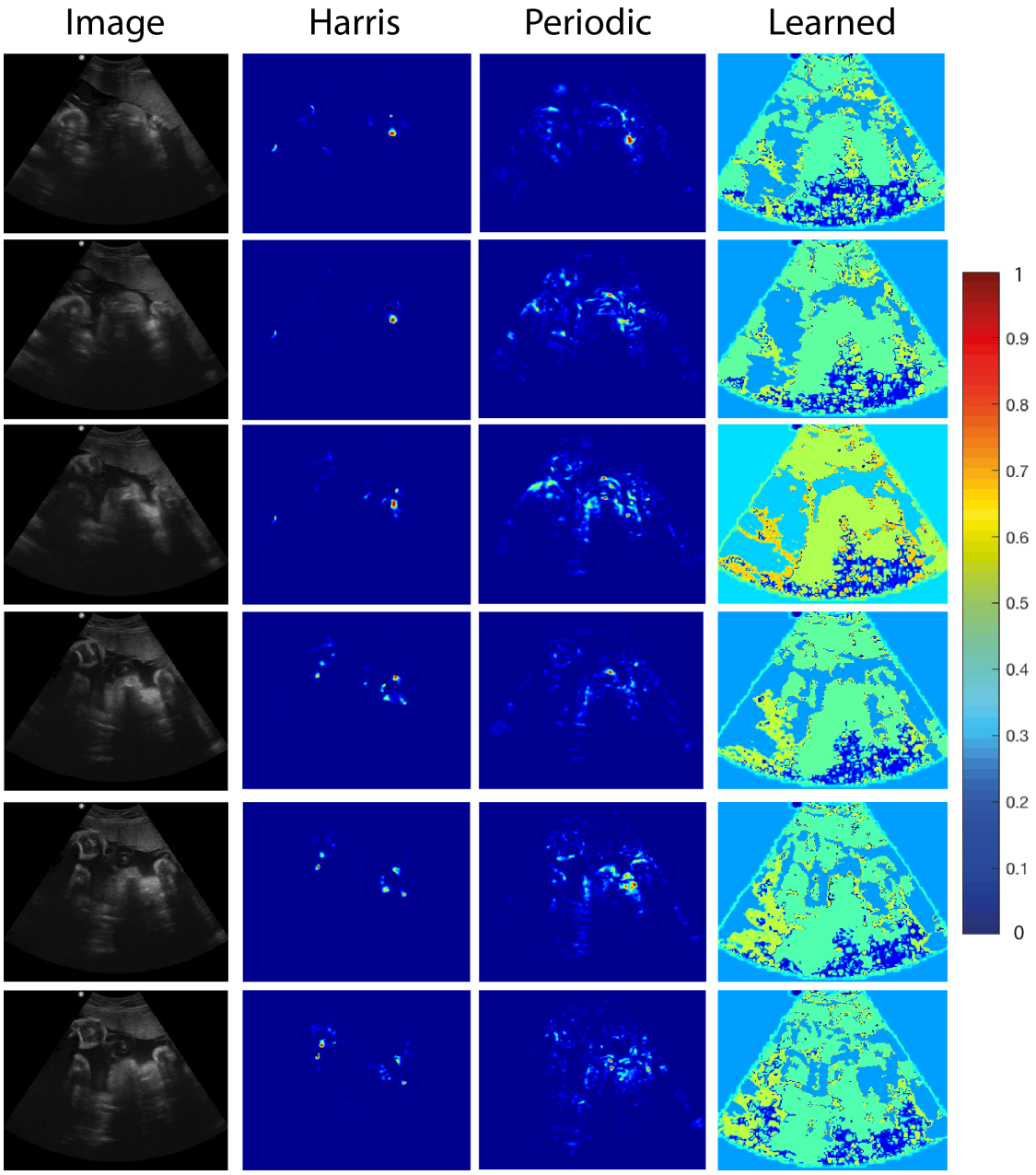


Figure 7.11: The response strength of the learned interest point operator across a sequence of images showing the fetal femur

.

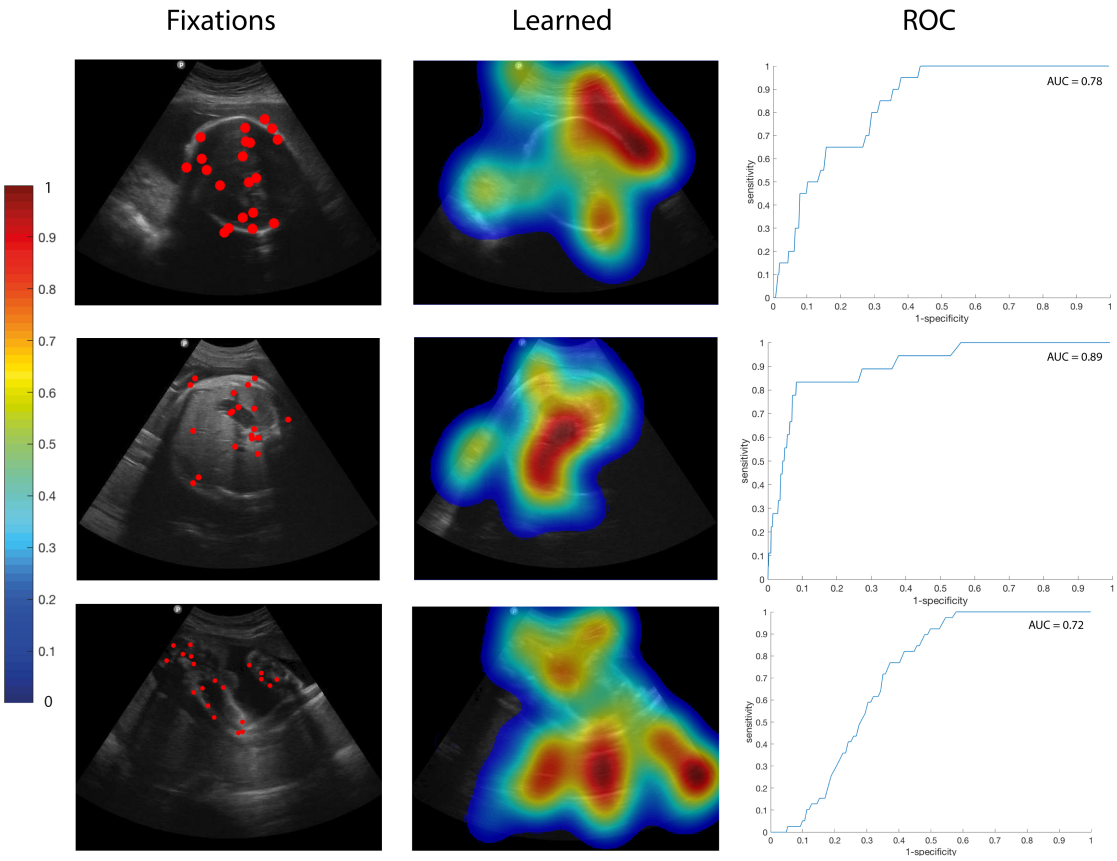


Figure 7.12: ROC curves and AUCs calculated for the accuracy of the learned spatio-temporal operator in predicting fixation points on a (top) head (centre) abdomen (bottom) other video frame.

.

	Learned AUC
Head	0.69
Abdomen	0.66
Other	0.62
Control	0.53

Table 7.4: The mean similarity scores of the learned spatio-temporal interest point operator with respect to fixation points. The random cross-image baseline measure is also shown.

7.4.3 Discussion

The ROC curves shown in Figure 7.12 and similarity scores in Table 7.4 demonstrate that the learned spatio-temporal interest point operator is more accurate in predicting fixations than either the Harris or Periodic operators.

This is to be expected as the learned operator was trained specifically for the purpose of predicting fixations. More interestingly, a greater proportion of interest points identified by the learned operator fall within manually labelled ROIs, suggesting that this operator identifies visually and anatomically salient image regions more effectively than the Harris or the Periodic operators.

This is further demonstrated in Figures 7.9 and 7.10. Here, the Harris operator saliency function fails to identify anatomically salient regions across head and abdominal video frames whereas the Periodic operator identifies portions of the stomach bubble, abdominal wall, skull, and thalami but fails to consistently identify these structures between frames. The learned operator saliency function produces a consistently high response across frames for the stomach bubble, umbilical vein, skull, and thalami but in abdomen images this operator also produces a strong response in shadow regions. This may either be due to a misclassification of shadows by the learned saliency function as likely fixation targets, or due to a tendency for observers to fixate on shadows as regions showing high levels of movement between video frames.

Figure 7.11 shows that on video frames showing other parts of the fetal anatomy, the learned operator fails to identify portions of the fetal anatomy such as the femur whereas the Harris and Periodic operators have stronger responses at these locations. This may be because, in the absence of specific anatomical landmarks to guide visual search (as with the stomach bubble and umbilical vein in abdomen videos, and the skull and thalami in head videos), fixated regions of these video frames are more disperse and may not exhibit the same variations in intensities between frames that characterise fixated locations in abdomen and head videos.

7.5 Bag of Visual Words Model

Having established that the learned spatio-temporal interest point detector outperforms both Harris and Periodic detectors in predicting fixations on 2-D+t fetal US video clips, this detector is then harnessed within a BoVW framework (Figure 7.13) for the automated classification of 2-D+t fetal US video clips into three categories: 'head', 'abdomen', and 'other'.

Spatio-temporal interest points were computed across each video in the training set by the learned operator and a cuboid centred on each interest point was extracted.

Gradient histograms were extracted from each cuboid and the resulting feature vectors describing each cuboid were clustered using k-means clustering, where each final cluster centre was designated as a single visual word. This resulted in a vocabulary, or bag, of visual words. Therefore each video in the training set could be represented by a histogram of visual words, or a single feature vector, as well as a ground-truth class label of 'head', 'abdomen' or 'other'.

A multi-class support vector machine (SVM) was trained to classify the visual word histograms representing unseen video clips as 'head', 'abdomen' or 'other'. Equivalent BoVW models based on the Periodic and Harris operators were trained for comparison with the learned operator.

7.5.1 Model Training

Training Data

The model was trained using dataset $C_{B\,n}$ as described in Chapter 4.

Interest Point Detection

The learned spatio-temporal interest point operator was used to detect interest points $P(x, y, t)$ across the set of training video clips. As in Section 7.3.3, non-maximal suppression was used to select pixels with a response value greater than or equal to all pixels in a surrounding window of diameter $15px$, again chosen to

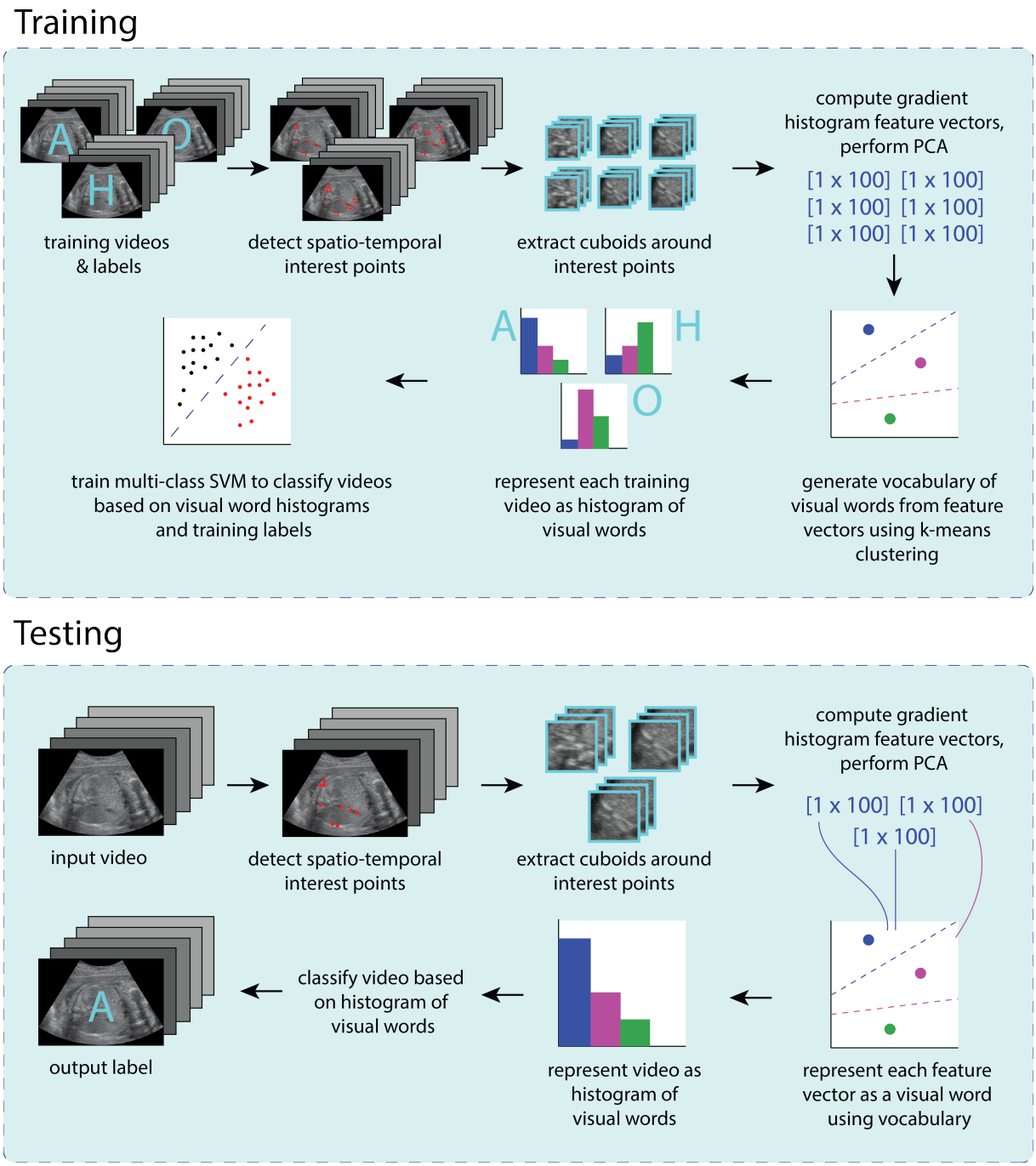


Figure 7.13: A schematic diagram of the bag-of-visual-words model for 2-D+t fetal US video clip classification.

match the foveated area of an observer seated $0.5m$ from a 25-inch 1920×1080 pixel LCD monitor[9]. The 13 strongest interest points in each frame were selected as the final spatio-temporal interest points; this number was chosen to match the mean number of fixations on each stimulus video frame across all observers during the experimental process described in Section 7.3.1, as the purpose of the learned spatio-temporal operator is to mimic the visual search behaviour of human observers. This resulted in a set of interest points for each video where $P_{i,j,n}$ describes the i^{th} interest point on the j^{th} frame of the n^{th} video clip in the training set. For comparison, corresponding sets of spatio-temporal interest points detected using the Harris and the Periodic operators were also computed across the set of training video clips.

Cuboid Extraction

A cuboid $C(x, y, t)$ of dimensions 30×30 pixels in the spatial direction and 34 pixels in the temporal direction was extracted centred around each spatio-temporal interest point (Figure 7.14). It was hypothesised that the frames preceding a fixation would be visually salient in order to attract the gaze to the region of the fixation, therefore the temporal cuboid dimension of 34 was chosen as twice the number of frames viewed during a $679ms$ fixation, which was the mean fixation length across all observers. The spatial cuboid dimensions were chosen to match the foveated area of an observer seated $0.5m$ from a 25-inch 1920×1080 pixel LCD monitor. This resulted in a set of cuboids where $P_{i,j,n}$ describes a cuboid extracted around the i^{th} interest point on the j^{th} frame of the n^{th} video clip in the training set.

Feature Extraction

Spatio-temporal features were then computed from each cuboid. It has been shown that gradient and optical flow[120] descriptors produce similar accuracies when used within BoVW pipelines for human action classification in 2-D+t video clips[125]. Given the significant differences between human action video datasets and fetal US video datasets, both gradient and optical flow descriptors were computed

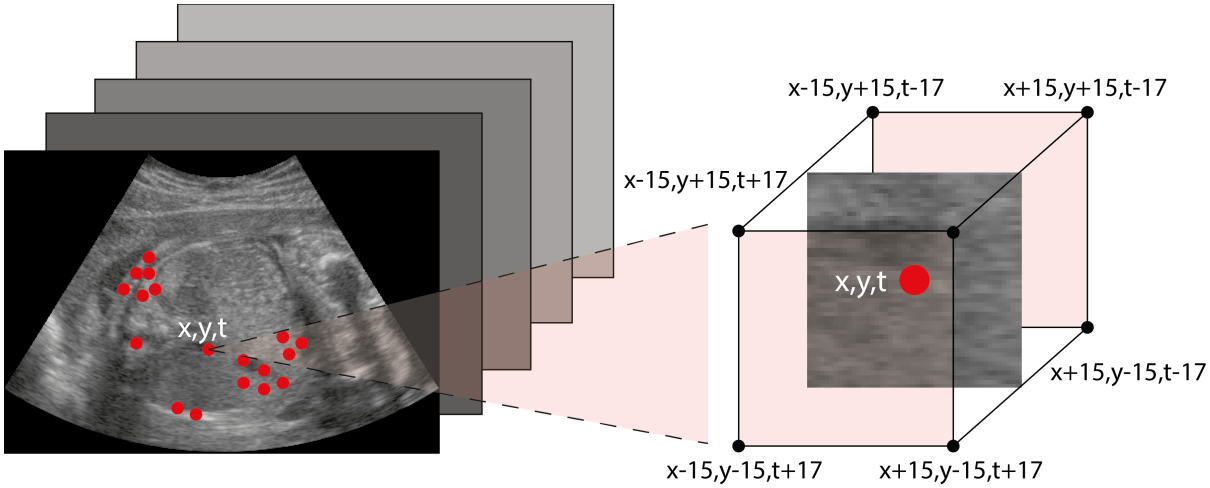


Figure 7.14: A schematic diagram showing the extraction of cuboids around spatio-temporal interest points.

to ascertain which produced higher accuracies within the BoVW framework for fetal US video classification.

Gradient descriptors were obtained by convolving each cuboid with 3-D Gaussians ($G(\sigma, \sigma, \tau)$ where σ and τ are the standard deviations in the spatial and temporal directions respectively) at two scales $\sigma = 1, \tau = 0.5$ and $\sigma = 2, \tau = 0.5$, chosen in accordance with the parameters used by Dollar[125]. The numerical gradients $\partial C_s/\partial x$, $\partial C_s/\partial y$, $\partial C_s/\partial t$ of the resulting smoothed cuboids C_s were then computed on a pixel-wise basis.

Optical flow descriptors were computed across each cuboid C in the temporal direction, between subsequent frames of the original training video clips. Each cuboid was convolved in the spatial directions with a 2-D Gaussian ($G(\sigma)$ where σ is the standard deviation in the spatial x and y directions) at the scale $\sigma = 2$. The velocity vectors (V_x, V_y) were then computed via the Lucas-Kanade method described in Chapter 6, across a radius of $4px$ in accordance with the parameters used by Dollar[125].

Principal component analysis was used to reduce the dimensionality of the gradient and optical flow descriptors via singular value decomposition. The dimensionality of gradient descriptors was reduced from 1×544 to 1×66 per cuboid. The number of principal components was chosen as 90% of the data was

Clusters	50	100	150	200	250	300	350	400	450	500
Accuracy (%)	51.85	55.56	55.56	66.67	74.07	70.37	62.96	66.67	59.26	62.96

Table 7.5: The results of cross-validation to determine the optimal number of clusters, or visual words, in the vocabulary of visual words. Mean classification accuracies of the overall BoVW pipeline in classifying $C_{A2\,n}$ are shown for each visual vocabulary size.

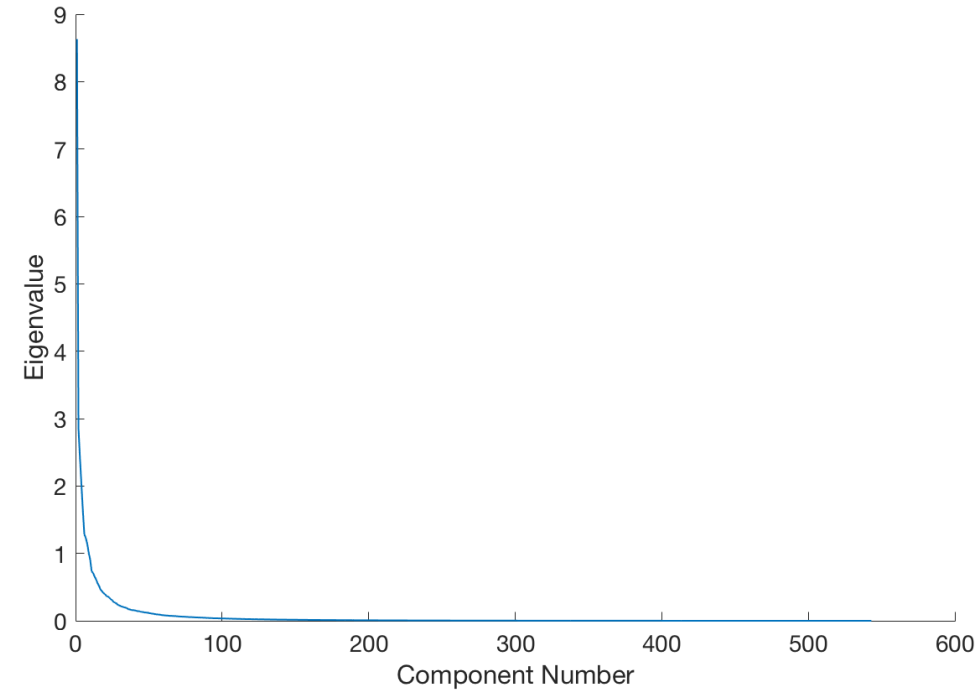
explained by 66 eigenvalues. This is shown in Figure 7.15, where it is apparent that the majority of variation in the data is described by less than 100 principal components. Similarly the dimensionality of the optical flow descriptors was reduced from 1×544 to 1×61 per cuboid, again capturing 90% of the variance in the data.

7.5.2 Vocabulary Construction

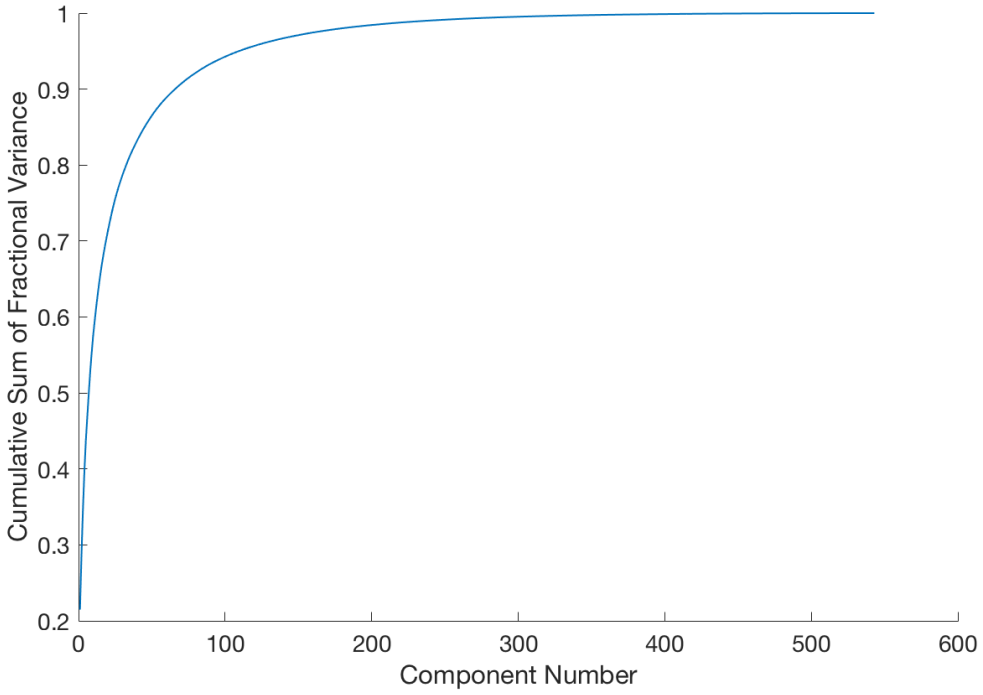
Gradient and optical flow descriptors were then clustered to form a vocabulary of visual words where each cluster represented a single visual word. Clustering of the 1×66 and 1×61 gradient and optical flow descriptors was carried out in 66-D and 61-D space respectively using a k-means algorithm.

As an initialisation step, k cluster centroids were seeded at random locations in the clustering space. Point-to-cluster-centroid distances were computed, giving the Euclidean distances from each point, or feature vector, and each cluster centroid. Each point was subsequently assigned to its closest cluster centroid, and k updated cluster centroids were computed as the mean of all points within a cluster. This process was repeated until cluster assignments did not change between consecutive iterations.

The number of clusters, or visual words in the vocabulary, was set through 10-fold cross-validation across a search space from 50 to 500 with a step size of 50. As shown in Table 7.5, $k = 250$ maximised the mean classification accuracy of the overall BoVW pipeline on $C_{A2\,n}$ and was selected as the size of the visual word vocabulary. Each training video clip could then be represented as a histogram of visual words, or a single 250-D feature descriptor, accompanied by a ground truth classification label of 'head', 'abdomen' or 'other'.



(a)



(b)

Figure 7.15: (a) Eigenvalues plotted against principal component numbers for PCA carried out on cuboid gradient descriptors (b) The cumulative sum of eigenvalues (as a fraction of the total sum of eigenvalues) plotted against principal component numbers.

7.5.3 Multi-Class Support Vector Machine

To obtain a m-class SVM for the classification of visual word histograms, m hard margin radial basis function (RBF) SVMs were combined using a 'one-vs-all' scheme where $m = 3$ for the US video dataset. The final trained SVMs were then combined to form a 3-class classifier. Here j_f, the final classification of a new visual word histogram x describing a new video clip corresponded to the j_f^{th} classifier which produced the maximal confidence score or positive value of $y(x)$:

$$arg\ max_{j=1...m} \Sigma_n \alpha_n^j K(x_n, x) + b^j \tag{7.3}$$

7.5.4 Model Testing

The trained BoVW model was tested across dataset $C_{C\,n}$ as described in Chapter 4. A given video clip in the testing set was correctly classified by the BoVW model if the ground truth class label $C^*_{C\,n}$ was equal to the class label assigned by the multi-class SVM.

7.5.5 Results

Classification accuracies for the BoVW pipeline across the testing set of 2-D+t US video clips are shown in Table 7.6 for both gradient and optical flow descriptors and learned, Harris and Periodic descriptors. The learned operator produces the highest classification accuracies for both types of descriptors across all video clip types, and a mean classification accuracy of 80.00% across all video clip types for gradient descriptors. For the learned operator, video clips with ground truth labels of 'head' were classified with the greatest accuracy, and those with ground truth labels of 'other' with the lowest accuracy.

A confusion matrix showing percentage misclassifications between video classes for the learned operator and gradient descriptor based BoVW model is shown in Figure 7.16 and demonstrates that the greatest number of misclassifications occurred when 'abdomen' videos were incorrectly classified as 'other'.

Video Type	Accuracy (%) Harris		Periodic		Learned	
	Gradient	Flow	Gradient	Flow	Gradient	Flow
Head	62.50	60.00	77.50	77.50	85.00	82.50
Abdomen	57.50	57.50	72.50	75.00	80.00	82.50
Other	52.50	52.50	65.00	62.50	72.50	75.00
Mean	57.50	56.67	71.67	71.67	79.17	80.00

Table 7.6: Classification accuracies of the BoVW pipeline across the testing set of 2-D+t fetal video clips showing the fetal head, abdomen, and other anatomical structures, for gradient and optical flow descriptors.

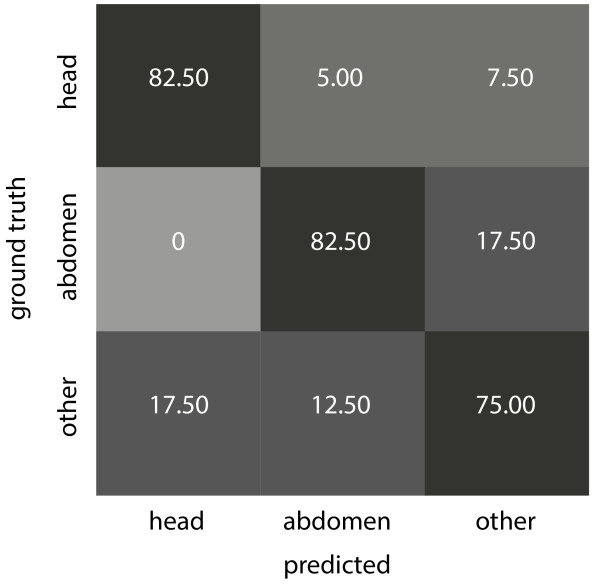


Figure 7.16: Confusion matrix showing correct and incorrect classifications of the learned operator and gradient descriptor based BoVW pipeline for 2-D+t fetal video clips showing the fetal head, abdomen, and other anatomical structures.

7.5.6 Discussion

It has been demonstrated that a biologically inspired spatio-temporal interest point operator produces higher classification accuracies for fetal US video clips within a BoVW pipeline compared to equivalent pipelines based on Harris or Periodic spatio-temporal operators. This is expected, as Section 7.4 demonstrated that the learned spatio-temporal operator is better able to identify anatomically and

visually salient video regions than conventional operators. However these findings are further confirmed here, as frequently occuring visual words identified by the learned operator appear to show anatomically meaningful structures, with 'head' videos characterised by the motion bright skull-like structures and 'abdomen' video by umbilical vein and stomach bubble-like structures. The prevalence of bright and dark structures in 'head' and 'abdomen' videos respectively perhaps accounts for the higher classification accuracies for these video types. In contrast, 'other' video clips showed a wider range of anatomical structures including the fetal femur, heart, arms, and feet; these contain regions which may appear similar to the skull, stomach bubble or umbilical vein and as such are more likely to be misclassified. Sub-dividing the 'other' class into more specific anatomical regions and repeating the experimental process to develop a BoVW pipeline capable of classifying, for example, 'femur' and 'heart' video clips in addition to 'head' and 'abdomen' may result in higher classification accuracies. It has also been demonstrated that mean classification accuracies across all classes are similar regardless of whether gradient or optical flow descriptors are used, in accordance with the findings of Kienzle[12] and Dollar[125]. This constitutes the first investigation into the efficacy of a spatio-temporal interest point operator learned directly from fixations in both predicting visually salient video regions and in automating the classification of 2-D+t fetal US video clips. This also presents an alternative to the use of CNNs for the analysis of spatio-temporal video data.

Young fool, only now, at the end, do you understand.

— The Emperor, Star Wars: Return of the Jedi (1983)

8 Conclusions and Future Work

Contents

8.1 Contributions

The main contributions of this thesis are as follows:

1. **An eye tracking inspired framework for anatomical landmark localisation in 2-D fetal US abdominal images** The first significant contribution of this thesis consisted of an eye tracking inspired framework for the automated localisation of the stomach bubble and umbilical vein in 2-D fetal abdominal US images. The first eye tracking experiments using US images as stimuli were reported on; subsequent analysis established high spatial and temporal similarities[9] between the eye movements of expert and novice observers. It was found that the fetal spine was frequently fixated on by observers, and was likely harnessed as a reference point against which

the positions of the fetal stomach bubble and umbilical vein were confirmed, despite not featuring in clinical guidelines[1] on the manual interpretation of fetal abdominal US images. The two-phase 'global-focal' model[10] of visual search was validated with respect to the eye tracking data, forming the first investigation into the high level constraints and anatomical structures which guide visual search in observers viewing US images. A pictorial structures model[11] was then derived to mimic the visual search behavior of observers. Candidate anatomical landmarks were identified by a sliding window detector trained on a boosted set of decision stumps, and their optimal configuration was determined by a probabilistic model trained on the relative positions of anatomical landmarks. The stomach bubble and umbilical vein detection accuracies of this framework showed an improvement on existing methods[6].

2. **An eye tracking inspired framework for standardised plane selection in 3-D fetal US abdominal volumes** The second significant contribution of this thesis was the development of an eye tracking inspired framework for the automated selection of standardised abdominal planes from 3-D fetal abdominal US volumes. The first eye tracking experiments using US volumes as stimuli were reported on; further analysis demonstrated high spatial and temporal similarities between the eye movements of observers, and that the spine and umbilical vein were the most frequently fixated anatomical landmarks. Significant differences were found between the length distributions and optical flow profiles of fetal spines, and artefacts with a similar appearance to the spine, in 2-D volume frames alone. Similarly, significant differences were found between the length distributions of fetal umbilical veins, and artefacts with a similar appearance to the umbilical vein, in 2-D volume frames alone. Informed by these findings, 3-D constraints were incorporated into the 2-D pictorial structures model derived in Chapter 3. The position and length of the fetal spine was constrained through the application of optical flow and length priors, and the length of the umbilical vein was similarly constrained. A dynamic programming algorithm was implemented to increase standardised

abdominal plane selection speeds, with the final framework improving on existing benchmark methods.[66].

3. **A predictive model of visual saliency in 2-D+t fetal US videos** The third contribution of this thesis was a model of visual saliency in 2-D+t fetal US video clips showing the fetal abdomen, head, and other parts of the fetal anatomy. The first eye tracking experiments using US video clips as stimuli were reported on, and demonstrated a low correspondence between the fixated locations of observers and the spatio-temporal interest points identified by the Harris and Periodic operators, indicating that the regions fixated on by observers may be more anatomically meaningful than those identified by conventional spatio-temporal operators. A multi-layer perceptron was trained directly on fixated locations to classify visually salient regions in unseen video clips, resulting in a superior model of visual saliency for US videos compared with the spatio-temporal Harris and Periodic operators.

4. **A framework for the classification of 2-D+t fetal US videos** Lastly, this thesis contributed a BoVW model for the classification of 2-D+t fetal US video clips into three categories: 'head', 'abdomen' and 'other'. The model was trained on gradient histograms from cuboids extracted around spatio-temporal interest points computed by the learned spatio-temporal operator derived previously. Feature vectors describing the extracted cuboids were clustered to produce a vocabulary of visual words, and each video was encoded as a histogram of visual words alongside a ground truth label. A multi-class SVM was trained to classify the visual word histograms into the categories of 'head', 'abdomen' and 'other'. The trained model produced higher classification accuracies across all three video clip classes compared to equivalent BoVW pipelines tested with the Harris and Periodic operators used in place of the learned spatio-temporal operator.

8.2 Future Work

There are a number of potential areas for future work based on the contributions presented in this thesis, namely the generalisation of this approach to other fetal imaging planes, alternative experimental designs, the development of a feature based model of visual saliency in fetal US images, and the potential use of CNNs to develop more accurate learned models of visual saliency.

8.2.1 Generalisation to Other Imaging Planes

The eye tracking experiments and automated US analysis frameworks presented in Chapters 5 & 6 of this thesis have been concerned with analysis of the standardised abdominal imaging plane. This plane was chosen due to the visibility of anatomical structures, indicating its suitability for eye tracking experiments, its significance for fetal growth monitoring, and the limitations of existing automated abdominal plane analysis methods. However, the generalisation of these methods to other standardised imaging planes warrants further investigation.

The eye tracking experiments described in this thesis would generalise best to other standardised imaging planes showing distinct anatomical structures which can be manually labelled as anatomical ROIs. This would enable analysis of which ROIs are most frequently fixated by observers, further validation of the global-focal model of visual search[10], and the discovery of high level constraints employed by observers. For this reason, it is hypothesised that these methods would be applicable to standardised views of the fetal head due to the visibility of anatomical landmarks including the thalami, cavum septum pellucidum, and midline and the strong geometric relationship between these structures. Standardised imaging planes of the femur would perhaps not be well suited to eye tracking experiments due to the lack of distinct anatomical landmarks in these images.

8.2.2 Experimental Design

There is considerable scope to modify the designs of the experiments presented in this thesis in order to mitigate the effects of increased observer proficiency during the experimental process.

In particular, the proficiency of novice observers may have improved during eye tracking experiments through increased familiarity with the shapes and orientations of anatomical landmarks in 2D image stimuli, and through learning that standardised abdominal planes tend to fall within the central portions of 3D volume stimuli. This may have altered observers' visual search strategies as they progressed through the stimuli, and encouraged repetitive learned behaviour whereby observers navigated to the central portion of each stimulus volume by default, rather than actively engaging in visual search to determine the position of the standardised plane. This could be addressed through varying the locations of standardised abdominal planes within 3D volume stimuli. Observers would therefore be unable to assume that standardised abdominal planes would become visible in the central portion of the volume, ensuring the eye tracking process was recording active visual search on unfamiliar stimuli rather than learned behaviour. Similarly including US images, volumes and videos showing growth restricted or otherwise abnormal fetuses in the set of eye tracking stimuli would perhaps prevent expert observers assuming standardised landmark orientations and appearances.

Equally, analysing changes in novice observers' search strategies with time may provide insights into the mechanisms via which expertise in US image interpretation is developed. The number of viewings required for novices to reach similar interpretation times, static and dynamic consistencies, and search strategies as experts woud also be a potential avenue for future work.

8.2.3 Visual Saliency in 2-D Fetal Abdominal US Images

Chapters 5 & 6 of this thesis primarily investigate which high level constraints guide visual search behaviour in observers analysing US images and volumes. Although a learned model of visual saliency is presented in Chapter 7, there has to date

been no investigation of which low level image descriptors attract the gaze in US images. As discussed in Chapter 3, a number of models of visual saliency in natural images have been constructed using low level features including gradient and intensity based features[8,9,106].

An initial step in constructing an equivalent model of visual saliency in US images would involve determining which features, at which spatial scales, are the best discriminators between fixation and non-fixation points. As outlined in Chapter 3, Rajashekar et al.[108] computed local descriptors at varying spatial scales around fixation points and non-fixation points in natural images, and found that regions surrounding fixation points were characterised by higher bandpass contrast values than non-fixations.

Further work could be conducted to determine which local descriptors computed at varying spatial scales are the best predictors of visual saliency, based on the approach of Rajashekar et al.[108]. Local descriptors would be extracted at varying spatial scales centred on fixation and non-fixation points. The distributions of feature values around fixation and non-fixation points would then be comparable using relative entropy to determine which features at which spatial scales were the strongest discriminators between fixations and non-fixations. Descriptors to be investigated may include gradient and intensity based features (such as bandpass intensity[108], steerable pyramids[82], Gabor pyramids[106], SURF, and SIFT) and geometric features (such as the distance from the image centre[108]) at varying spatial scales. The efficacy of classifiers including SVMs[82] in predicting visually salient image regions when trained on these local descriptors would also warrant investigation. In keeping with the results of Rajashekar et al. and with the tendency for the human visual system to fixate on high contrast, 'blob' like structures[126], it is expected that Haar-like features would be amongst the strongest predictors of fixations.

8.2.4 Visual Saliency with Convolutional Neural Networks

One advantage of the learned model of visual saliency presented in Chapter 7, compared to feature based models, was the use of a feed-forward neural network to avoid the need for feature selection or optimisation.

CNN frameworks are becoming a standard machine learning tool in medical image analysis. Therefore a natural question is whether the work presented in this thesis might be extended using such frameworks. It has been suggested by Li et al.[127] that CNNs may be even better adapted to modelling visual saliency than feed-forward neural networks due to their architecture; convolutional layers resemble individual cells within the human visual system, and fully connected layers are analogous to higher level decision making processes. Li et al. have proposed a CNN architecture to identify visually salient regions of natural images; however, in their study, ground truth annotations for visual saliency were provided by manual segmentations of salient objects within the images, rather than fixations recorded during eye tracking experiments.

There is therefore potential for CNNs to be used to generate visual saliency maps for fetal US images, which in turn could be harnessed to aid anatomical landmark localisation. Baumgartner et al.[59,60] have developed a CNN based framework for standardised plane detection and anatomical landmark localisation during freehand fetal US scanning. It could be possible to adapt a similar or related model to identify salient regions in 2-D fetal abdominal US images, and compare these regions to areas fixated on by human observers. This would constitute an initial step in building a CNN based model of visual saliency. Ground truth annotations consisting of manually segmented anatomical landmarks could then be substituted with fixations obtained through eye tracking experiments. The accuracy of the newly trained model in predicting visual saliency maps and identifying standardised imaging planes and localising anatomical landmarks could then be re-assessed. This idea is currently being explored by Yifan Cai, a doctoral research student within the Biomedical Image Analysis Group at the University of Oxford Institute of Biomedical Engineering, building directly on the work presented in this thesis.

References

[1] L J Salomon and J Bernard, "Feasibility and Reproducibility of an Image-Scoring Method for Quality Control of Fetal Biometry in the Second Trimester," *Ultrasound in Obstetrics and Gynaecology*, vol. 27, no. 1, pp. 34–40, 2006.

[2] K Haram, E Softeland, and R Bukowski, "Intrauterine Growth Restriction," *International Journal of Gynaecology and Obstetrics*, vol. 93, no. 1, pp. 5–12, 2006.

[3] J Kurmanavicius, E Wright, and P Royston, "Fetal Ultrasound Biometry: 2. Abdomen and Femur Length Reference Values," *British Journal of Obstetrics and Gynaecology*, vol. 106, pp. 136–143, 1999.

[4] Royal College of Obstetricians and Gynaecologists, "The Investigation and Management of the Small for Gestational Age Fetus," Tech. Rep., 2013.

[5] I Sarris, C Ioannoi, and P Chamberlain, "Interobserver Variability in Fetal Ultrasound Measurements," *Ultrasound in Obstetrics and Gynaecology*, vol. 39, no. 3, pp. 266–273, 2012.

[6] B Rahmatullah, I Sarris, J A Noble, and A Papageorghiou, "Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using adaboost," *Proceedings of the 15th Conference on Medical Image Understanding and Analysis*, pp. 6–9, 2011.

[7] B Rahmatullah, I Sarris, J A Noble, and A Papageorghiou, "Automated standard plane selection from fetal abdominal ultrasound volumes using a machine learning algorithm," *Abstracts of the 22nd World Congress on Ultrasound in Obstetrics and Gynecology*, pp. 134–5, 2012.

[8] Krista Ehinger, B Hidalgo-Sotelo, A Torralba, and A Oliva, "Modeling Search for People in 900 Scenes: A Combined Source Model of Eye Guidance," *Visual Cognition*, vol. 17, no. 6-7, pp. 945–978, 2009.

[9] S Mathe and C Sminchisescu, "Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 842–856.

[10] H L Kundel, C F Nodine, E Krupinski, and C Mello-Thoms, "Using Gaze-Tracking Data and Mixture Distribution Analysis to Support a Holistic

Model for the Detection of Cancers on Mammograms," *Academic Radiology*, vol. 15, no. 7, pp. 881–6, 2008.

[11] M Fischler and R Elschlager, "The Representation and Matching of Pictorial Structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, 1973.

[12] W Kienzle and B Scholkopf, "How to Find Interesting Locations in Video: a Spatiotemporal Interest Point Detector Learned from Human Eye Movements," in *Pattern Recognition: Lecture Notes in Computer Science*, pp. 405–414. 2007.

[13] I Sipiran and B Bustos, "Harris 3D: a robust extenstion of the Harris operator for interest point detection on 3D meshes," *Image and Vision Computing*, vol. 27, pp. 963–976, 2011.

[14] P Dollar, V Rabaud, G Cottrell, and S Belongie, "Behavior Recognition via Sparse Spatio-temporal Features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[15] J E Lawn, S Cousens, and J Zupan, "4 Million Neonatal Deaths: When? Where? Why?," *Lancet*, vol. 365, no. 9462, pp. 9–18, 2000.

[16] L J Salomon, Alfirevic Z, and V Berghella, "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *Ultrasound in Obstetrics and Gynaecology*, vol. 37, no. 1, pp. 116–126, 2011.

[17] P Loughna, L Chitty, R Evans, and R Chudleigh, "Fetal Size and Dating: Charts Recommended for Clinical Obstetric Practice," *Ultrasound*, vol. 17, no. 3, pp. 160–166, 2009.

[18] E Platz and R Newman, "Diagnosis of IUGR: Traditional Biometry," *Seminars in Perinatology*, vol. 32, no. 3, pp. 140–147, 2008.

[19] I Sarris, C Ioannou, M Dighe, A Mitidieri, M Oberto, W Qingqing, J Shah, S Sohoni, W Al Zidjali, and L Hoch, "Standardization of Fetal Ultrasound Biometry Measurements: Improving the Quality and Consistency of Measurements," *Ultrasound in Obstetrics & Gynecology*, vol. 38, no. 6, pp. 681–687, 2011.

[20] S Akmal, E Tsoi, and K Nicolaides, "Intrapartum Sonography to Determine Fetal Occipital Position: Interobserver Agreement," *Ultrasound in Obstetrics & Gynecology*, vol. 24, no. 4, pp. 421–424, 2004.

[21] T Rackham, S Rueda, C Knight, and J A Noble, "Ultrasound Image Segmentation Using Feature Asymmetry and Shape Guided Live Wire," in *International Society for Optics and Photonics, Medical Imaging*, 2013.

[22] M Yaqub, B Kelly, A T Papageorghiou, and J A Noble, "Guided Random Forests for Identification of Key Fetal Anatomy and Image Categorization in Ultrasound Scans," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 687–694, 2015.

[23] B Rahmatullah, I Sarris, A Papageorghiou, and J A Noble, "Quality Control of Fetal Ultrasound Images: Detection of Abdomen Anatomical Landmarks Using Adaboost," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, 2011, pp. 6–9.

[24] B Rahmatullah, A Papageorghiou, and J A Noble, "Image Analysis Using Machine Learning: Anatomical Landmarks Detection in Fetal Ultrasound Images," in *Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual*. IEEE, 2012, pp. 354–355.

[25] S Liao, Y Gao, A Oto, and D Shen, "Representation Learning: a Unified Deep Learning Framework for Automatic Prostate MR Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 254–261.

[26] M Kim, G Wu, and D Shen, "Unsupervised Deep Learning for Hippocampus Segmentation in 7.0 Tesla MR Images," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2013, pp. 1–8.

[27] G Wu, D Shen, and M Sabuncu, *Machine Learning and Medical Imaging*, Elsevier Science, 2016.

[28] M R Avendi, A Kheradvar, and H Jafarkhani, "A combined deep-learning and deformable- model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Medical Image Analysis*, vol. 30, pp. 108–119, 2016.

[29] A Krizhevsky, I Sutskever, and G E Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[30] NHS, "Ultrasound Scans in Pregnancy" [website]. Accessed 3 Feb 2016: `http://www.nhs.uk/conditions/pregnancy-and-baby-care.aspx`

[31] R Mikolajczyk, J Zhang, A Betran, J Souza, R Mori, M Gülmezoglu, and M Merialdi, "A Global Reference for Fetal-Weight and Birthweight Percentiles," *The Lancet*, vol. 377, no. 9780, pp. 1855–1861, 2011.

[32] K Butt, K Lim, S Bly, Y Cargill, G Davies, N Denis, G Hazlitt, L Morin, A Ouellet, and S Salem, "Determination of gestationa age by ultrasound," *Journal of Obstetrics and Gynaecology Canada*, vol. 36, no. 2, pp. 171–181, 2014.

[33] L Chitty and P Pandya, "Ultrasound Screening for Fetal Abnormalities in the First Trimester," *Prenatal Diagnosis*, vol. 17, no. 13, pp. 1269–1281, 1998.

[34] L Hui and D Challis, "Diagnosis and Management of Fetal Growth Restriction: the Role of Fetal Therapy," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 22, no. 1, pp. 139–158, 2008.

[35] T Van Mieghem, S Giusca, P DeKoninck, L Gucciardo, E Doné, A Hindryckx, J D'hooge, and J Deprest, "Prospective Assessment of Fetal Cardiac Function with Speckle Tracking in Healthy Fetuses and Recipient Fetuses of Twin-to-Twin Transfusion Syndrome," *Journal of the American Society of Echocardiography*, vol. 23, no. 3, pp. 301–308, 2010.

[36] I Germanakis and H Gardiner, "Assessment of Fetal Myocardial Deformation using Speckle Tracking Techniques," *Fetal diagnosis and therapy*, vol. 32, no. 1-2, pp. 39–46, 2012.

[37] T Szabo, *Diagnostic Ultrasound Imaging: Inside Out*, Academic Press, 2004.

[38] V Chalana, T Winter, and D Cyr, "Automatic Fetal Head Measurements from Sonographic Images," *Academic Radiology*, vol. 3, no. 8, pp. 628–635, 1996.

[39] S Pathak, V Chalana, and Y Kim, "Interactive Automatic Fetal Head Measurements from Ultrasound Images Using Multimedia Computer Technology," *Ultrasound in Medicine & Biology*, vol. 23, no. 5, pp. 665–673, 1997.

[40] S Jardim and M Figueiredo, "Automatic Contour Estimation in Fetal Ultrasound Images," in *Proceedings of the IEEE International Conference on Image Processing*, 2003, pp. 1065–1068.

[41] S Jardim and M Figueiredo, "Segmentation of Fetal Ultrasound Images," *Ultrasound in Medicine & Biology*, vol. 31, no. 2, pp. 243–250, 2005.

[42] G Carneiro, "Detection and Measurement of Fetal Anatomies from Ultrasound Images using a Constrained Probabilistic Boosting Tree," *Medical Imaging, IEEE Transactions*, vol. 27, no. 9, pp. 1342–1355, 2008.

[43] J Yu, Y Wang, P Chen, and Y Shen, "Fetal Abdominal Contour Extraction and Measurement in Ultrasound Images," *Ultrasound in Medicine & Biology*, vol. 34, no. 2, pp. 169–182, 2008.

[44] J Yu, Y Wang, and P Chen, "Fetal Ultrasound Image Segmentation System and its use in Fetal Weight Estimation," *Medical & Biological Engineering & Computing*, vol. 46, no. 12, pp. 1227–1237, 2008.

[45] J Nithya and M Madheswaran, "Detection of Intrauterine Growth Retardation using Fetal Abdominal Circumference," in *International Conference on Computer Technology and Development*. IEEE, 2009, vol. 2, pp. 371–375.

[46] M Yaqub, P Mahon, M Javaid, C Cooper, and J A Noble, "Weighted voting in 3D random forest segmentation," in *Medical Image Understanding and Analysis*, 2010, pp. 261–266.

[47] B Rahmatullah, A Papageorghiou, and J A Noble, "Integration of Local and Global Features for Anatomical Object Detection in Ultrasound," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pp. 402–409. Springer, 2012.

[48] S Rueda, C Knight, A Papageorghiou, and J A Noble, "Regularised Feature-Based Fuzzy Connectedness Segmentation of Ultrasound Images for Fetal Soft Tissue Quantification Across Gestation," in *International Symposium on Biomedical Imaging*, 2012, pp. 1323–1326.

[49] A Foi, M Maggioni, A Pepe, S Rueda, J A Noble, A T Papageorghiou, and J Tohka, "Difference of Gaussians Revolved Along Elliptical Paths for Ultrasound Fetal Head Segmentation," *Computerized Medical Imaging and Graphics*, vol. 38, no. 8, pp. 774–784, 2014.

[50] *Object Classification in an Ultrasound Video using lp-SIFT Features*, 2014.

[51] M A Maraci, R Napolitano, A T Papageorghiou, and J A Noble, "Searching for Structures of Interest in an Ultrasound Video Sequence," *International Workshop on Machine Learning in Medical Imaging*, pp. 133–140, 2014.

[52] Ana IL Namburete, Richard V Stebbing, Bryn Kemp, Mohammad Yaqub, Aris T Papageorghiou, and J Alison Noble, "Learning-Based Prediction of Gestational Age from Ultrasound Images of the Fetal Brain," *Medical Image Analysis*, vol. 21, no. 1, pp. 72–86, 2015.

[53] C P Bridge and J A Noble, "Object Localisation in Fetal Ultrasound Images using Invariant Features," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 156–159.

[54] H Ryou, M Yaqub, A Cavallaro, F Roseman, A Papageorghiou, and J A Noble, "Automated 3D Ultrasound Biometry Planes Extraction for First Trimester Fetal Assessment," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2016, pp. 196–204.

[55] M Yaqub, S Rueda, A Kopuri, P Melo, A T Papageorghiou, P B Sullivan, K McCormick, and J A Noble, "Plane Localization in 3-D Fetal Neurosonography for Longitudinal Analysis of the Developing Brain," *IEEE journal of biomedical and health informatics*, vol. 20, no. 4, pp. 1120–1128, 2016.

[56] Y Gao, M A Maraci, and J A Noble, "Describing Ultrasound Video Content using Deep Convolutional Neural Networks," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 787–790.

[57] M A Maraci, C P Bridge, R Napolitano, A Papageorghiou, and J A Noble, "A Framework for Analysis of Linear Ultrasound Videos to Detect Fetal Presentation and Heartbeat," *Medical image analysis*, vol. 37, pp. 22–36, 2017.

[58] W Huang, C P Bridge, J A Noble, and A Zisserman, "Temporal HeartNet: Towards Human-Level Automatic Analysis of Fetal Cardiac Screening Video," *arXiv preprint arXiv:1707.00665*, 2017.

[59] C F Baumgartner, K Kamnitsas, J Matthew, T P Fletcher, S Smith, L M Koch, B Kainz, and D Rueckert, "Real-Time Detection and Localisation of Fetal Standard Scan Planes in 2D Freehand Ultrasound," *arXiv preprint arXiv:1612.05601*, 2016.

[60] C F Baumgartner, K Kamnitsas, J Matthew, T P Fletcher, S Smith, L M Koch, B Kainz, and D Rueckert, "SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound," *IEEE Transactions on Medical Imaging*, 2017.

[61] J A Noble and D Boukerroui, "Ultrasound Image Segmentation: A Survey," *IEEE Transactions on medical imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[62] J Udupa and S Samarasekera, "Fuzzy Connectedness and Object Definition: Theory, Algorithms, and Applications in Image Segmentation," *Graphical Models and Image Processing*, vol. 58, no. 3, pp. 246–261, 1996.

[63] S Rueda, S Fathima, C L Knight, M Yaqub, A T Papageorghiou, B Rahmatullah, A Foi, M Maggioni, A Pepe, J Tohka, et al., "Evaluation and Comparison of Current Fetal Ultrasound Image Segmentation Methods for Biometric Measurements: a Grand Challenge," *IEEE Transactions on medical imaging*, vol. 33, no. 4, pp. 797–813, 2014.

[64] A Ciurte, X Bresson, and M B Cuadra, "A Semi-Supervised Patch Based Approach for Segmentation of Fetal Ultrasound Imaging," *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, pp. 5–7, 2012.

[65] R V Stebbing and J E McManigle, "A Boundary Fragment Model for Head Segmentation in Fetal Ultrasound," *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, pp. 9–11, 2012.

[66] B Rahmatullah, A Papageorghiou, and J A Noble, "Automated Selection of Standardized Planes from Ultrasound Volume," in *Machine Learning in Medical Imaging*, pp. 35–42. Springer, 2011.

[67] G Carneiro, B Georgescu, and S Good, "Knowledge Based Automated Fetal Biometrics," Tech. Rep., Siemens Medical Solutions, 2008.

[68] G Doretto, A Chiuso, Y N Wu, and S Soatto, "Dynamic Textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[69] R M Haralick and K Shanmugam, "Textural Features for Image Classification," *IEEE Transactions on systems, man, and cybernetics*, , no. 6, pp. 610–621, 1973.

[70] *Plane Selection for Corpus Callosum Visualization in 3D Ultrasound Images*, 2015.

[71] P A Viola and M J Jones, "Robust Real-Time Face Detection," in *ICCV*. Citeseer, 2001, vol. 2, p. 747.

[72] J Matas, O Chum, M Urban, and T Pajdla, "Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[73] N Dalal and B Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[74] D G Low, "Object Recognition from Local Scale-Invariant Features," in *Proceedings of the International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[75] H Bay, T Tuytelaars, and L Van G, "Surf: Speeded Up Robust Features," *Computer vision–ECCV 2006*, pp. 404–417, 2006.

[76] *Fisher Vector Encoding for Detecting Objects of Interest in Ultrasound Videos*, 2015.

[77] *Object Localisation in Fetal Ultrasound Images Using Invariant Features*, 2015.

[78] M Felsberg and G Sommer, "The Monogenic Signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[79] J Matas, O Chum, M Urban, and T Pajdla, "Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[80] P Viola and M Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Conference on Computer Vision and Pattern Recognition*, 2001.

[81] R Kwitt, N Vasconcelos, S Razzaque, and S Aylward, "Localizing target structures in ultrasound video–a phantom study," *Medical Image Analysis*, vol. 17, no. 7, pp. 712–722, 2013.

[82] A Torralba, A Oliva, K S Castelhano, and J M Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, no. 4, p. 766, 2006.

[83] S Frintrop and E Rome, "Simulating Visual Attention for Object Recognition," in *Proceedings of the Workshop on Early Cognitive Vision, Isle of Skye, Scotland*, 2004.

[84] L Itti and C Koch, "Computational Modelling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[85] L Itti, C Koch, and E Neibur, "A Model of Saliency Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, p. 1254, 1998.

[86] H L Kundel and C F Nodine, "Interpreting Chest Radiographs without Visual Search," *Radiology*, vol. 116, no. 3, pp. 527–532, 1975.

[87] H L Kundel, C F Nodine, E F Conant, and S P Weinstein, "Holistic Component of Image Perception in Mammogram Interpretation: Gaze-Tracking Study," *Radiology*, vol. 242, no. 2, pp. 396–402, 2007.

[88] Robert J. K. Jacob, "What You Look at is What you Get: Eye Movement Based Interaction Techniques," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People - CHI '90*, pp. 11–18, 1990.

[89] P Majaranta and R Kari-Jouko, "Twenty Years of Eye Typing: Systems and Design Issues," in *Eye Tracking Research & Applications*, 2002, pp. 15–22.

[90] J D Smith and T C N Graham, "Use of Eye Movements for Video Game Control," 2006, p. 20, ACM Press.

[91] O Spakov and D Miniotas, "Gaze-Based Selection of Standard-Size Menu Items," in *7th International Conference on Multimodal Interfaces*. 2005, p. 124, ACM Press.

[92] A Stellmach and R Dachselt, "Designing Gaze-Based User Interfaces for Steering in Virtual Environments," in *Eye Tracking Research & Applications*, 2010, pp. 131–138.

[93] V Tanriverdi and R J K Jacob, "Interacting with Eye Movements in Virtual Environments," in *SIGCHI Conference on Human Factors in Computing Systems*. 2000, pp. 265–272, ACM Press.

[94] C Ware, "An Evaluation of an Eye Tracker as a Device for Computer Input," in *SIGCHI conference on Human factors in Computing Systems Empowering People*, 1987, pp. 183–188.

[95] Tobii, "Tobii Pro X3-120" [website]. Accessed 17 Jun 2016: `http://www.tobiipro.com/product-listing/tobii-pro-x3-120/#Specifications`

[96] Tobii, "Tobii Technology," .

[97] A Olsen, "The Tobii I-VT Fixation Filter: Algorithm Description," Tech. Rep., 2012.

[98] Harold L Kundel, "History of Research in Medical Image Perception," *Journal of the American College of Radiology*, vol. 3, no. 6, pp. 402–8, 2006.

[99] A L Yarbus, *Eye movements during perception of complex objects*, Springer, 1967.

[100] C Beam, E Krupinski, H L Kundel, E Sickles, and R F Wagner, "The Place of Medical Image Perception in 21st-Century Health Care," *Journal of the American College of Radiology*, vol. 3, no. 6, pp. 409–12, 2006.

[101] C F Nodine and H L Kundel, "Using Eye Movements to Study Visual Search and to Improve Tumor Detection," *Radiographics: a Review Publication of the Radiological Society of North America*, vol. 7, no. 6, pp. 1241–50, 1987.

[102] J J H Leong, M Nicolaou, R J Emery, W Darzi, and G Z Yang, "Visual Search Behaviour in Skeletal Radiographs: a Cross-Specialty Study," *Clinical Radiology*, vol. 62, no. 11, pp. 1069–77, 2007.

[103] A Mallett, P Phillips, T R Fanshawe, E Helbren, D Boone, A Gale, S A Taylor, D Manning, D G Altman, and S Halligan, "Tracking Eye Gaze During Interpretation of Endoluminal Three-Dimensional CT Colonography: Visual Perception of Experienced and Inexperienced Readers," *Radiology*, vol. 273, no. 3, pp. 783–792, 2014.

[104] R Bertram, L Helle, J K Kaakinen, and E Svedstr, "The Effect of Expertise on Eye Movement Behaviour in Medical Image Perception," *PloS one*, vol. 8, no. 6, p. e66169, 2013.

[105] M Antonelli and G Z Yang, "Lung Nodule Detection Using Eye-Tracking," in *IEEE International Conference on Image Processing*, 2007, pp. 457–460.

[106] L Itti, C Koch, and E Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[107] Z Bylinskii, T Judd, A Borji, L Itti, F Durand, A Oliva, and A Torralba, "Mit saliency benchmark," .

[108] U Rajashekar, I Van der Linde, A Bovik, and L Cormack, "Foveated Analysis of Image Features at Fixations," *Vision Research*, vol. 47, no. 25, pp. 3160–3172, 2007.

[109] J Harel, C Koch, and P Perona, "Graph-based Visual Saliency," in *NIPS*, 2006, vol. 1, p. 5.

[110] W Kienzle and F Wichmann, "Learning an Interest Point Operator from Human Eye Movements," *IEEE Computer Vision and Pattern Recognition Workshop*, p. 24, 2006.

[111] I Laptev, "On Space Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.

[112] Intergrowth 21^{st}, "Intergrowth 21^{st}" [website]. Accessed 14 Aug 2016: `http://www.intergrowth21.org.uk`

[113] SC Perni, FA Chervenak, RB Kalish, S Magherini-Rothe, M Predanic, J Streltzoff, and DW Skupski, "Intraobserver and interobserver reproducibility of fetal biometry," *Ultrasound in obstetrics & gynecology*, vol. 24, no. 6, pp. 654–658, 2004.

[114] Bero O Verburg, Paul GH Mulder, Albert Hofman, Vincent WV Jaddoe, Jacqueline Witteman, and Eric AP Steegers, "Intra-and interobserver reproducibility study of early fetal growth parameters," *Prenatal diagnosis*, vol. 28, no. 4, pp. 323–331, 2008.

[115] Piotr Dollar, "Piotr's computer vision matlab toolbox (PMT)," https://github.com/pdollar/toolbox.

[116] C F Nodine, H L Kundel, S C Lauver, and L C Toto, "Nature of Expertise in Searching Mammograms for Breast Masses," *Academic Radiology*, vol. 3, no. 12, pp. 1000–1006, 1996.

[117] C F Nodine, H L Kundel, C Mello-Thoms, S P Weinstein, S G Orel, D C Sullivan, and E F Conant, "How Experience and Training Influence Mammography Expertise," *Academic Radiology*, vol. 6, no. 10, pp. 575–585, 1999.

[118] P Dollar, S Belongie, and P Perona, "The Fastest Pedestrian Detector in the West," 2010.

[119] Pedro Felzenszwalb and Daniel Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[120] B D Lucas and T Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," 1981.

[121] T Kiserud and S Johnsen, "Biometric Assessment," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 23, no. 6, pp. 819–831, 2009.

[122] C Harris and M Stephens, "A Combined Corner and Edge Detector," in *Alvey Vision Conference*. Citeseer, 1988, vol. 15, p. 50.

[123] J G Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[124] S Marcelja, "Mathematical description of the responses of simple cortical cells," *JOSA*, vol. 70, no. 11, pp. 1297–1300, 1980.

[125] P Dollar, V Rabaud, G Cottrell, and S Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[126] M Jahangiri and M Petrou, "An Attention Model for Extracting Components that Merit Identification," in *Proceedings of the IEEE International Conference on Image Processing*, 2009.

[127] G Li and Y Yu, "Visual Saliency Based on Multiscale Deep Features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.