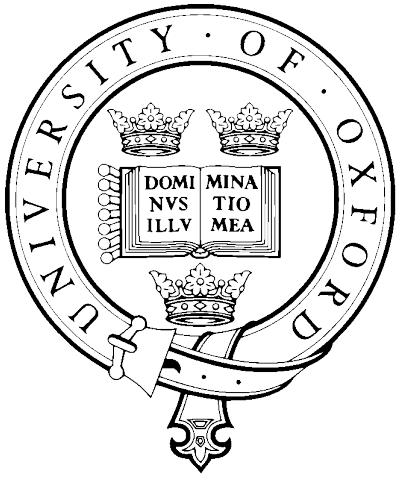
Assessment of Obstetric Ultrasound Images using Machine Learning

Bahbibi Rahmatullah

St Catherine's College



Supervisors: Prof J. Alison Noble and Dr Aris T. Papageorghiou

Trinity Term 2012

Institute of Biomedical Engineering

Department of Engineering Science

University of Oxford

Abstract

Ultrasound-based fetal biometry is used to derive important clinical information for identifying IUGR (intra-uterine growth restriction) and managing risk in pregnancy. Accurate and reproducible biometric measurement relies heavily on a good standard image plane. However, qualitative visual assessment, which includes the visual identification of certain anatomical landmarks in the image is prone to inter- and intra-reviewer variability and is also time-consuming to perform. Automated anatomical structure detection is the first step towards the development of a fast and reproducible quality assessment of fetal biometry images. This thesis deals specifically with abdominal scans in the development and evaluation of methods to automatically detect the stomach and the umbilical vein within them.

First, an original method for detecting the stomach and the umbilical vein in fetal abdominal scans was developed using a machine learning framework. A classifier solution was designed with AdaBoost learning algorithm with Haar features extracted from the intensity image. The performance of the new method was compared on different clinically relevant gestational age groups.

Speckle and the low contrast nature of ultrasound images motivated the idea of introducing features extracted from local phase images. Local phase is contrast invariant and has proven to be useful in other ultrasound image analysis application compared with intensity. Nevertheless, it has never been implemented in a machine learning environment before. In our second experiment, local phase features were proven to have higher discriminative power than intensity features which enabled them to be selected as the first weak classifiers with large classifier weight.

Third, a novel approach to improving the speed of the detection was developed using a global feature symmetry map based on local phase to select the candidate locations for the stomach and the umbilical vein. It was coupled with a local intensity-based classifier to form a "hybrid" detector. A nine-fold increase in the average computational speed was recorded along with higher accuracy in the detection of both the anatomical structures.

Quantitative and qualitative evaluations of all the algorithms were presented using 2384 fetal abdominal images retrieved from the image database study of the Oxford Ultrasound Quality Control Unit of the INTERGROWTH-21st project.

Finally, the "hybrid" detection method was evaluated in two potential application scenarios. The first application was clinical scoring in which both the computer algorithm and four experts were asked to record presence or absence of the stomach and the umbilical vein in 400 ultrasound images. The computer-experts agreement was found to be comparable with the inter-expert agreement. The second application concerned selecting the standard image plane from 3D abdominal ultrasound volume. The algorithm was successful in selecting 93.36% of the images plane defined by the expert in 30 ultrasound volumes.

Acknowledgements

I would like to thank my supervisors Prof. Alison Noble and Dr. Aris Papageorghiou for their help, support and guidance throughout this research degree. I am truly grateful for the opportunity to do this research project with the funding support of the Sultan Idris Education University (UPSI) Malaysia and Malaysian Ministry of Higher Education (MOHE). I would also like to thank Dr Ippokratis Sarris, Dr Christos Ioannau and Dr Caroline Knight for their collaboration and contribution in this work. I am grateful to all of my colleagues within the Biomedical Image Analysis (BioMedIA) Laboratory for thought-provoking discussions, support and kindness. A special thanks to the examiners who have kindly agreed to examine this thesis.

I would also like to thank my family and friends for all their support and encouragement throughout this research. Most importantly, I wish to thank my mother for her non-stop encouragements and prayers throughout my academic work.

Lastly I cannot express enough thanks to my husband Arfian and my three children, Abdullah, Bilal and Hasanah for their love, patience and support. This thesis is dedicated to them.

.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Small for gestational age (SGA) refers to the situation when a fetus is smaller than expected for the number of weeks of pregnancy. Newborn babies with SGA are often associated with having intrauterine growth restriction (IUGR), which is a more specific condition where the fetus fails to reach its growth potential. 18 million babies are born every year with low birth weight because of IUGR and/or prematurity, resulting in significant short- and long term morbidity and mortality (Lawn et al., 2005). Growth restricted fetuses have poorer neonatal outcomes and it is recognised that developmental delay associated with IUGR leads to significant health care and developmental problems during childhood and most likely in adult life (Barker, 2006). Recognition of the serious risks associated with IUGR has elevated its diagnostic importance among perinatologists. Thus, obtaining accurate assessment of fetal growth and gestational age from fetal biometry for identifying risks to the fetus/neonate is very important.

Historically, X-ray was used to measure fetal dimensions (e.g. fetal head, pelvic dimension) (Shenton, 1922) before the development of ultrasound. The development of two-dimensional (2D) ultrasound made it possible to measure the dimensions of bones and soft tissue structures of the fetus faster and more reliably than with x-rays. 2D ultrasound is currently considered to be the first choice for a safe, non-invasive, accurate and cost-effective investigation in the fetus. It has progressively become an indispensable obstetric tool and plays an important role in pregnancy management. Comprehensive ultrasound examination during pregnancy includes standard fetal biometric measurements, which are primarily used to estimate the gestational age of the fetus, to track fetal growth patterns, to estimate fetal

weight and to detect abnormalities. A detailed description of ultrasound-based fetal biometry used for age estimation and growth assessments is given in Section 2.2.

Fetal biometry is determined from standardized ultrasound planes taken from the fetal head, abdomen and thigh. The acquisition of optimal image planes from which these measurements are taken is crucial to allow for accurate and reproducible biometric measurements, and also to minimize inter- and intra-observer variability. Criteria and description of standard fetal biometric planes are presented in Section 2.5.

The importance of quality control for the scanning procedures and measurements has been emphasized (Dudley, 2006, Ville, 2008) and a quality control policy based on image scoring has been proposed (Salomon et al., 2006). To highlight challenges in scanning and acquiring the standard image plane, scans made by several different sonographers for finding the abdominal measurements after they had been briefed on the scanning protocol for a growth study known as INTERGROWTH-21st are shown in Figure 1.1. Even though all the scans shown in the figure are magnified satisfactorily, the appearance of the stomach and the umbilical vein are inadequate in some of the scans. According to Salomon's grading, scans in the first column are acceptable with the stomach and the umbilical vein clearly identified and in the correct position. However, elongated umbilical vein appearance in the second column's images indicates that the plane is too angled.

Figure 1.1: Scans acquired by different sonographers for finding abdominal measurement during image quality training session. The stomach and the umbilical vein (white arrows) are clearly visible in the images in the first column. However, elongated umbilical vein in images in the second column indicates that the plane is too angled.

There are several limitations of visual image quality assessment. The process of image review, which includes the visual identification of certain anatomical landmarks in the image, requires significant human resources. Extensive qualitative analysis is also time-consuming and costly. Furthermore, there is an issue of inter- and intra-reviewer variability and also bias imposed by a human reviewer. An automated image scoring system, which 1) can perform the evaluation quickly, 2) is robust to appearance variations of the visual object of interest, and 3) efficient and economical for any scale of implementation would be a valuable support to the quality control process.

This thesis deals specifically with the development of automated methods for the detection of two important landmarks (the stomach and the umbilical vein) in fetal abdominal ultrasound scans using machine learning. The plane containing these two landmarks is described in the early proposal for using the abdominal circumference measurement for fetal weight estimation (Campbell and Wilkin, 1975). The plane containing these two landmarks was adopted in constructing the widely used chart for abdominal circumference size (Chitty et al., 1994) and also proposed in the image quality scoring system (Salomon and Ville, 2005).

1.2 Contributions

The main contributions of this thesis are summarized below:

1. The development of an original method to detect the stomach and the umbilical vein in fetal abdominal scan using a machine learning technique (Chapter 3). Parts of this chapter have been published at peer-reviewed conferences:

    i. Quality Control of Fetal Ultrasound Images: Detection of Abdomen Anatomical Landmarks using Adaboost. *IEEE International Symposium on Biomedical Imaging (ISBI),* 2011.

    ii. Image Analysis Using Machine Learning: Anatomical Landmarks Detection in Fetal Ultrasound Image. *IEEE Signature Conference on Computers, Software, and Applications (COMPSAC),* 2012.

2. The investigation of introducing features extracted from the local phase image into the machine learning framework for the detection of the two anatomical landmarks (stomach and umbilical vein) (Chapter 4). Part of this chapter has been published at a peer-reviewed conference:

i. Multi-Scale Local Phase Features for Anatomical Object Detection in Fetal Ultrasound Images. *Medical Image Understanding and Analysis Conference (MIUA)*, 2012.

3. The development of a faster and more accurate detector using a hybrid approach for the detections of the stomach and the umbilical vein (Chapter 5). Part of this chapter has been published at a peer-reviewed conference:

i. Integration of Local and Global Features for Anatomical Object Detection in Ultrasound. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2012.

4. The evaluation of the proposed detection method in two clinical application scenarios (Chapter 6). Parts of this chapter has been published at a peer-reviewed conference:

i. Automated Selection of Standardized Planes from Ultrasound Volume. *MICCAI Workshop on Machine Learning in Medical Imaging (MLMI)*, 2011.

and as abstracts in the following clinical meetings:

ii. A Pilot Study of Automated Image Scoring for Quality Control Purposes in the Context of Multicentre Studies: Abdominal Circumference. *World Congress on Ultrasound in Obstetrics and Gynecology*, 2011.

iii. Automated Fetal Biometry Image Landmark Detection for Confirming Correct Image Planes: Abdominal Circumference. *World Congress on Ultrasound in Obstetrics and Gynecology*, 2012.

iv. Automated Standard Plane Selection from Fetal Abdominal Ultrasound Volumes using a Machine Learning Algorithm. *World Congress on Ultrasound in Obstetrics and Gynecology*, 2012.

1.3 Thesis Outline

Chapter 2 describes the background knowledge on fetal growth restriction and the current clinical practice which uses ultrasound for its assessment along with its challenges and the quality control process. The chapter also provides the review of related image analysis work in fetal ultrasound domain and the application of machine learning for detection purposes in medical imaging.

Chapter 3 describes the initial method used for the detection of important anatomical landmarks in fetal abdominal ultrasound images using machine learning framework.

Chapter 4 deals with utilizing features from multi-scale local phase images in the same detection framework. The efficiency of the new feature sets are compared to the performance intensity-based features used in Chapter 3.

Chapter 5 introduces a new hybrid approach for the enhancement of the performance and the speed of the detection. A multi-scale feature symmetry measure derived using local phase is combined with the local intensity-based detector (developed in Chapter 3) is utilized for fast object detection and its detection performance is analysed.

Chapter 6 evaluates the application of the proposed algorithm in two potentials scenarios: comparison with experts' agreements in recording the presence and absence of the anatomical structures in fetal abdominal scan and utilizing the algorithm for the selection of standard plane from 3D volumes.

Chapter 7 concludes the thesis and discusses directions for future work.

Chapter 2 Literature Review

This chapter presents a review of literature relevant to this thesis. The first part provides relevant clinical background dealing with the application of ultrasound imaging in the management of pregnancy and the detection of fetal growth abnormalities. Section 2.1 describes fetal growth restriction and its clinical implications. Various ultrasound-based fetal biometry measures used in clinical practice for growth assessment are explained in Section 2.2 and the description of an ongoing clinical study for optimal growth standard is presented in Section 2.3. The challenges in the fetal ultrasound application are defined in Section 2.4 and the quality assessment process to overcome such challenges is mentioned in Section 2.5. The second part of this chapter covers technical background with Section 2.6 presenting a review of related work on automatic image analysis in the fetal ultrasound domain. This is followed by a discussion of relevant medical image analysis techniques using machine learning in Section 2.7. A summary is presented in Section 2.8.

2.1 Introduction

Intra-Uterine Growth Restriction (IUGR) (formerly known as Intra-Uterine Growth Retardation) refers to a condition in which a fetus fails to reach its genetically determined potential size or appropriate growth potential (a specific biometric or estimated weight threshold by a specific gestational age) in the uterus. IUGR is part of a wider condition known as Small for Gestational Age (SGA) fetuses. This group includes fetuses that have a poor growth rate and fetuses that are genetically small but have reached their appropriate growth potential. Approximately 50–70% of fetuses with a birthweight below the tenth centile for gestational age are constitutionally small (Wilcox, 1983, Ott, 1988) and the lower

the centile for defining SGA, the higher the likelihood of IUGR (Royal College of Obstetricians and Gynaecologists, 2002).

A growth restricted fetus has much greater short-term morbidity and mortality compared with its normal counterpart. In 2004, WHO and UNICEF estimated that more than 20 million infants worldwide, representing 15.5 per cent of all births, were born with low birth weight (UNICEF and WHO, 2004). Low weight at birth is either the result of prematurity and/or IUGR, where both are the leading causes of perinatal morbidity and mortality (Gabbe et al., 2007). IUGR has been identified as a major factor in over 60% of the 4 million neonatal deaths worldwide and an equal number of stillbirths (Lawn et al 2005). Growth restricted fetuses are at greater risk of birth hypoxia (McIntire et al., 1999), neonatal complications (Fleischer et al., 1992, Bernstein et al., 2000) and impaired neurodevelopment (Roth et al., 1999). More recently, long-term consequences of this condition are being recognized. Among them are cardiovascular disease (Andersen et al., 2010, Risnes et al., 2011), obesity (Reinehr et al., 2009), type-2 diabetes and hypertension in adult life (Barker, 2006, Eriksson et al., 2006).

Recognition of the serious risks associated with IUGR has elevated its diagnostic importance. A correct estimation of gestational age (GA) and estimated fetal weight (EFW) are two important factors in the assessment of fetal growth for identifying IUGR risks to the fetus. Prior to the introduction of ultrasound, gestational age estimation was based on the woman's last menstrual period (LMP) or by tape measurement (known as MacDonald’s rule). Fetal biometric measurements from ultrasound images have enabled a more accurate estimation of gestational age and fetal weight, by using some specific formulas and growth charts (Loughna et al., 2009). The serial measurements of these anatomic parameters, consisting of length, diameter or circumferences of various body segments, have proven to be useful in assessing fetal growth and detecting any growth abnormalities.

2.2 Ultrasound-based Fetal Biometry

Real-time acquisition capability, low cost (relative to other modalities) and portability of ultrasound imaging makes it the most widely used imaging modality in the field of obstetrics. Currently, the standard clinical practice for fetal age and growth estimation is by using 2D ultrasound for taking measurements of fetal dimensions. In this section, the commonly used fetal biometric ultrasound measurements are presented.

2.2.1 Weeks 6 - 13 of gestation

Beginning Week 6 to Week 13 of gestation, fetal age can be estimated from the length of the fetus, also called the Crown-Rump Length (CRL) (Figure 2.1). The measurement usually made at very early gestational age is actually the longest longitudinal dimension (LLD) of the fetal pole, from the top of the head to the bottom of the rump (Goldstein, 1991). The CRL measurement is recommended throughout the first trimester since it is easy to measure and there is an excellent correlation between length and age in early pregnancy when growth is rapid and minimally affected by pathologic disorders (Fleischer et al., 1991).

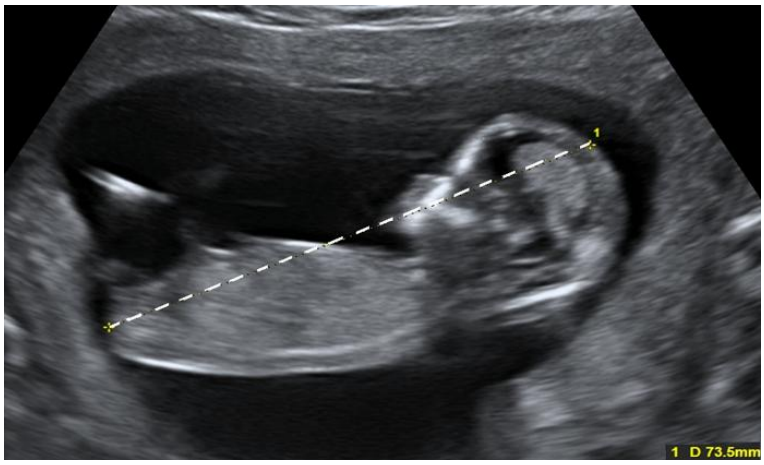


Figure 2.1: An ultrasound showing a fetus measured to have a crown-rump length (CRL) of 73.5mm

2.2.2 Weeks 13 - 25 of gestation

Standard fetal biometric ultrasound measurements are taken from three main body parts, namely: the head, the thigh and the abdomen. In United Kingdom, these measurements are typically taken during the routine scan around 20 weeks of gestation (Royal College of Obstetricians and Gynaecologists, 2000). The errors associated with these individual measurements are significantly greater than those in the first trimester (Evans, 2006).

There are three standard measurements (see Figure 2.2) taken from fetal head ultrasound image:

i. Biparietal Diameter (BPD): It is measured from the outer border of the parietal bones at the widest part of the skull.

ii. Occipito-Frontal Diameter (OFD): It is measured across the longest part of the skull, from the outer border of the occipital and frontal edges.

iii. Head Circumference (HC): It could either be calculated from the BPD and the OFD measurements using the formula $HC = \pi\,(BPD + OFD)/2$ or directly obtained from the circumference of the ellipse placement on the skull.

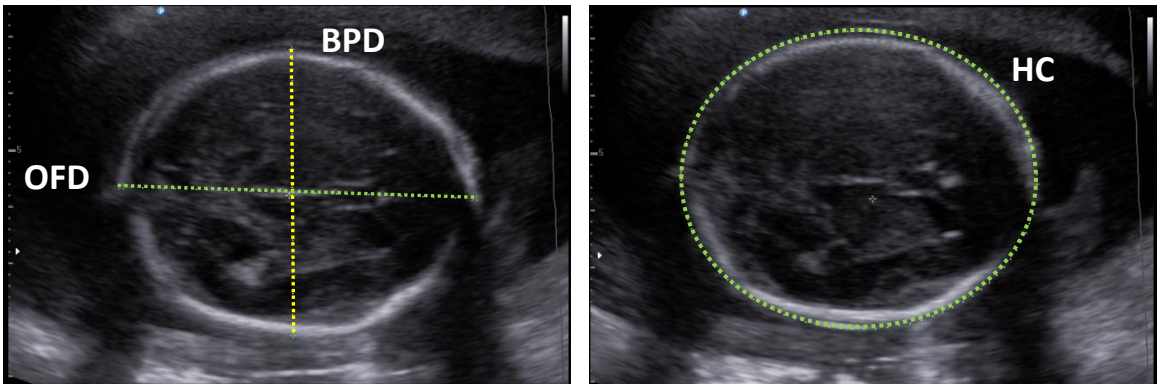


Figure 2.2: Fetal biometric measurements from fetal head ultrasound images showing the three standard measurements (BPD, OFD and HC).

Similarly, three standard measures (illustrated in Figure 2.3) are taken from the fetal abdomen image:

i. Anterio-Posterior Abdominal Diameter (APAD): It is measured from the outer borders of the body outline, from the edge of the skin covering the spine to the anterior abdominal wall.

ii. Transverse Abdominal Diameter (TAD): It is measured from the outer borders of the body outline across the abdomen at the widest point, taken perpendicular to the APAD.

iii. Abdominal Circumference (AC): It could either be calculated from the APAD and the TAD measurements using the formula $AC = \pi\,(APAD + TAD)/2$ or directly obtained from the circumference of the ellipse placement on the abdomen.

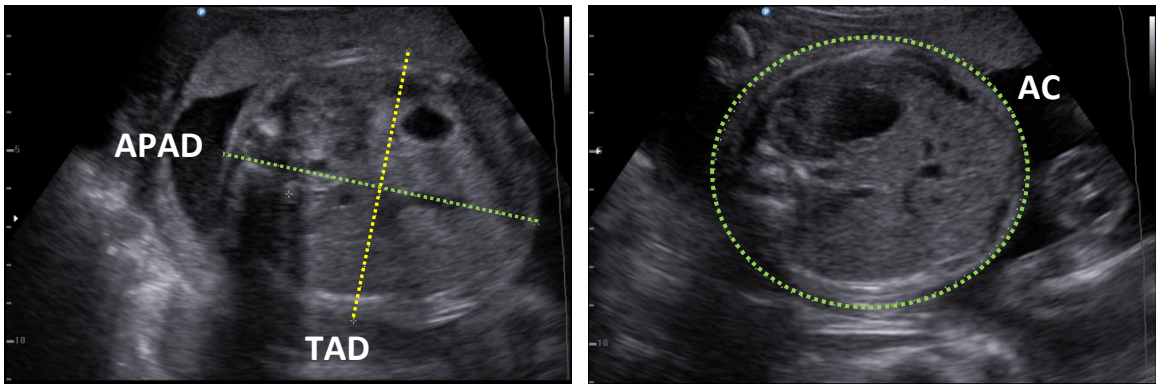


Figure 2.3: Fetal biometric measurements from fetal abdomen ultrasound images showing the three standard measurements (APAD, TAD and AC).

For the thigh area, the standard fetal biometry measure taken is the femur length (FL). FL is measured from the outer edges of the bone, without taking into account the trochanter[1] of the femur (refer to Figure 2.4).

[1] large, irregular, quadrilateral eminence, toward the near end of the femur.

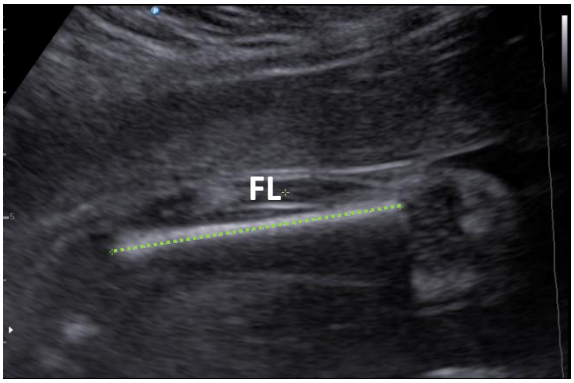


Figure 2.4: Fetal biometric measurement from fetal thigh ultrasound image showing the FL measurement.

2.2.3 Weeks 26 - 42 of gestation

Fetal age determination in the third trimester is similar to that in the second trimester with the same anatomical parameters used. However, errors in age estimates increase progressively toward term. The reason is that growth of different structures may vary considerably in the third trimester due to intrinsic factors (e.g genetic growth potential) and varying growth support received from the placenta and mother (Fleischer et al., 1991). Reduction in age estimation errors can be obtained by averaging the age estimates determined from the BPD, HC, AC and FL. This average age estimate, called the *composite menstrual age*, provides the best ultrasound estimate of fetal age in the second and third trimesters (Hadlock et al., 1984).

From all the presented parameters, the AC is recognised to be more correlated to birth weight and currently the most important measure in assessing fetal growth and to detect growth abnormalities (Warsof et al., 1986). One of the reasons is because the abdomen contains both lean mass (liver) and fat mass (subcutaneous and visceral) and is not based on rigid bone structures, making it more sensitive to environmental conditions. However, the variability in AC measurement is also broader compared to the other two fetal biometry measures (Hadlock et al., 1982).

2.2.4 Other Anatomical Measurements

Ultrasound measurements of other anatomical structures have been used to detect different types of abnormalities including the width of the nuchal translucency (Nicolaides et al., 1992), various dimensions of the fetal heart (Souka and Nicolaides, 1997), the anterior-posterior diameter of the renal pelvis (Gloor et al., 1997), and the diameter of the atrium of the fetal brain (Pilu et al., 1989). Hata and Deter (1992) gives a review of a large number of specific anatomical measurements made on individual organs such as brain, heart, lung, thymus, liver, spleen, pancreas, stomach, gallbladder, kidney, adrenal glands, intestine, and bladder. However, these measurements have not been adopted into clinical practice since their clinical value is yet to be fully established.

2.2.5 3D Ultrasound

The introduction of 3D ultrasound has the potential to transform the management of pregnancy as 3D measurements can be made. While 2D measurement is accurate for structures of simple shape, it can be less accurate for structures without obvious symmetrical plane such as the lung or liver. The 2D ultrasound image represents a thin plane at some arbitrary angle in the body and it can be difficult to localize the image plane and reproduce it at a later time for follow-up studies. Acquisition of structures with complex volumetric shapes has become much easier with 3D ultrasound. Hence analysis is now possible though it has not yet been well established in clinical practice.

Publications from studies deriving fetal growth volume curves using 3D ultrasound are increasing. The focus are mainly the organs such as the liver (Kuno et al., 2002, Chang et al., 2006, Dos Santos Rizzi et al., 2010), kidney (Hsieh et al., 2000, Yu et al., 2000, Darahem

Tedesco et al., 2009), lung (Sabogal et al., 2004, Britto et al., 2009, Prendergast et al., 2011), and brain (Chang et al., 2003, Roelfsema et al., 2004, Rutten et al., 2009).

However, there are wide discrepancies in the reported volumes. Inconsistencies in volumetric methodology such as the definitions of imaging planes and the anatomical landmarks used for measurement, are the common factors for poor agreement in volumetric measurements as shown by Ioannou et al. (2011). They proposed for the standardization of these factors along with the validation of the method in vitro and in vivo, that would improve intra- and inter-researcher agreement for reported volumetric measures.

2.3 INTERGROWTH-21st

A set of international fetal and newborn growth standards (fetal growth, birth weight for gestational age and postnatal growth of preterm infants) is being developed by The International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). INTERGROWTH-21st aims to develop prescriptive standards which describe optimal growth in well-nourished, healthy fetuses and how such growth relates to neonatal outcome (Intergrowth 21st, 2008). It involves collaborating sites worldwide where data has been obtained from over 4000 healthy pregnant women at low risk of impaired fetal growth who are scanned with 2D and 3D ultrasound up to six times during pregnancy from $<14^{+0}$ weeks to term. All women are screened at study entry based on a set of criteria defining a “clinically healthy” woman, which particularly focus on excluding known risk factors for IUGR (e.g. smoking, chronic illness) and over-growth (e.g diabetes). Multiple pregnancy or major fetal abnormalities are also excluded from the study.

2D ultrasound images and manual measurements made on the images are being recorded including standard biometric measures of the head (BPD, OFD, HC), the abdomen (AC, TAD, APAD) and the femur (FL). Each manual measure is repeated 3 times. The 3D

volumes are obtained by scanning the head (at the level of the BPD), the abdomen (at the level of AC) and the femur. These volumetric scans are being used primarily for quality control, i.e. to assess the quality of 2D measures taken, and if necessary to retake or take additional measurements retrospectively. There are three novel components in the INTERGROWTH-21ST study:

i. It excludes mothers whose fetuses are unlikely to have optimal growth because of environmental constraints, i.e. malnutrition, smoking, illness etc.

ii. It is a first attempt to acquire clinical data on a large scale. Previous studies are typically conducted on a single site and much smaller than this.

iii. Inclusion of 3D ultrasound in the protocol for quality control purpose, i.e. to assess the quality of 2D fetal dimension measurement.

The 2D images and 3D volumes used in this thesis came from this study.

2.4 Challenges in Fetal Ultrasound Imaging

This section discusses two main challenges normally encountered in making reliable and accurate diagnostic measurement using ultrasound images.

2.4.1 Quality of Images

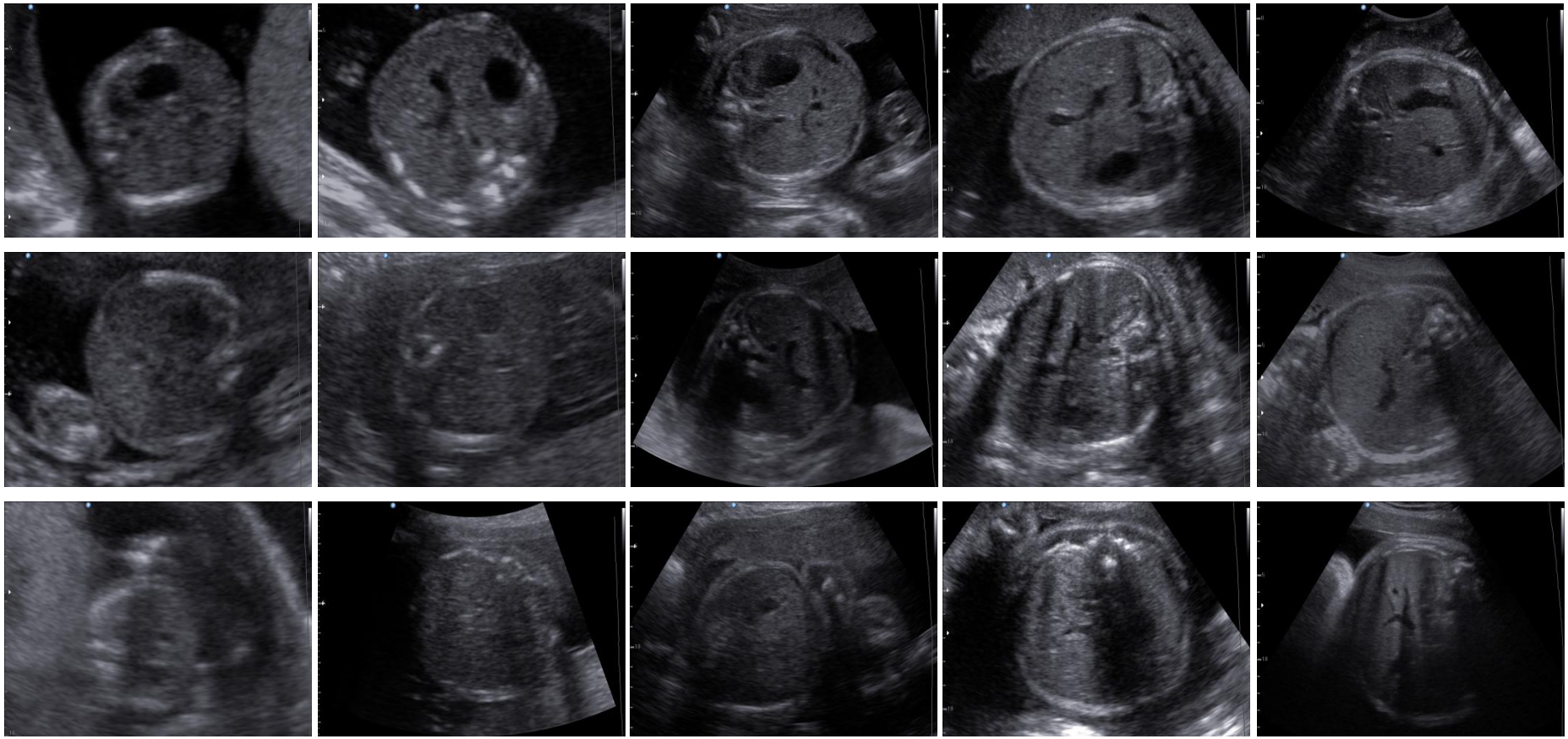
Ultrasound images are very challenging due to the presence of speckle, shadows and low contrast characteristic features. Ultrasound waves are poorly transmitted through air and adipose tissue, and are reflected by bone. The visualization of anatomical structures is dependent on factors including orientation of the transducer with respect to the object, signal attenuation and missing information due to signal dropouts and acoustic shadows. Therefore, there is a great emphasis for the clinician to carefully follow the designed acquisition protocols wherever possible.

In the case of fetal ultrasound images, fetus shape and anatomy varies during pregnancy. Inconsistent positioning of the probe on the patient body and the fetal position during the scan are among the factors for the variability. As the pregnancy advances and the size of the fetus increases and becoming more compressed within the womb, the quality of the images degrades with more appearance of artefacts and shadows (due to the reflection from the surrounding bones which has increased in density). The low quality could also be attributed to the increasing body mass index of the mother where the ultrasound waves have farther to travel and are attenuated along the way. Figure 2.5 shows examples of different qualities of fetal abdominal ultrasound scan for abdominal measurement at 5 stages of gestation.

2.4.2 Inter- and Intra-Operator Variability

The acquisition of ultrasound images and the fetal biometry measures are based on the operator's experience and subjective impression. Often significant training is required to acquire good data. The accuracy and the reproducibility of the measurements are subjected to inter- and intra-observer variability even when performed by trained sonographers, as well as specific conditions in each examination (Sarris et al., 2012). In addition, acquiring some of the more advanced measurements is tedious and time-consuming.

Standardization of procedures including the selection of the most appropriate plane for taking the measurement and the standard way of taking measurement using the calliper placement, is necessary to achieve maximum validity. The standardization of procedures for fetal biometry measures can in theory be achieved through a robust method of quality control.



(a) 17 weeks (b) 21 weeks (c) 28 weeks (d) 33 weeks (e) 38 weeks

Figure 2.5: Different quality of fetal abdomen ultrasound images at various gestational ages (17, 21, 28, 33 and 38 weeks). The high quality images (top row) have clear boundary of the abdominal wall and the standard anatomical landmarks (stomach and umbilical vein) are clearly visible. The medium quality images (middle row) have partial occlusion of the landmarks and/or less distinguishable borders. In low quality images (bottom row), there are problem of reduced contrast due to shadowing and poor visibility of the landmarks and borders.

2.5 Qualitative Measures in Fetal Biometry Images

Qualitative measures rely on assessing the image quality and appearance/absence of object features that should be included in an image if the correct measurement is to be taken. Subjective evaluation of images is common when reviewing an ultrasound examination performed by another operator. Although this is clinically acceptable, it does not provide good inter- or intra-reviewer reproducibility, even when a reviewer is trained and experienced in this type of evaluation (Salomon et al., 2006). More importantly, subjective evaluation is unlikely to improve training and practice on a large scale because it cannot provide either constructive comments or objective input for improvement (Salomon and Ville, 2005).

Objective quality assessment of ultrasound images and measurements has been proposed for various clinical examinations including those of the breast tumors (Van Limbergen et al., 1989), ovarian tumors (DePriest et al., 1993), cardiovascular evaluation (Crouse et al., 1986), and examination of the appendix (Tzanakis et al., 2005). Image scoring in prenatal ultrasound has been developed largely for the purpose of improving the standardisation of nuchal translucency (NT) measurement (Herman et al., 1998). It was adopted as the gold standard of quality control of nuchal translucency measurements at 11 to 14 weeks of gestation (Snijders et al., 2002, Fries et al., 2007).

Salomon et al. (2006) proposed a quality control method based on image scoring for fetal biometry in the second trimester akin to the clinically accepted image scoring practice adopted in the quality control of nuchal translucency ultrasound measurements. They performed experiments which involved both subjective and objective evaluation on a set of 300 images comprising of 100 images for each biometric measurement (abdomen, femur and head) by three different reviewers. The results showed that the subjective evaluation did not provide good inter- and intra-reviewer reproducibility, even when the reviewers are trained

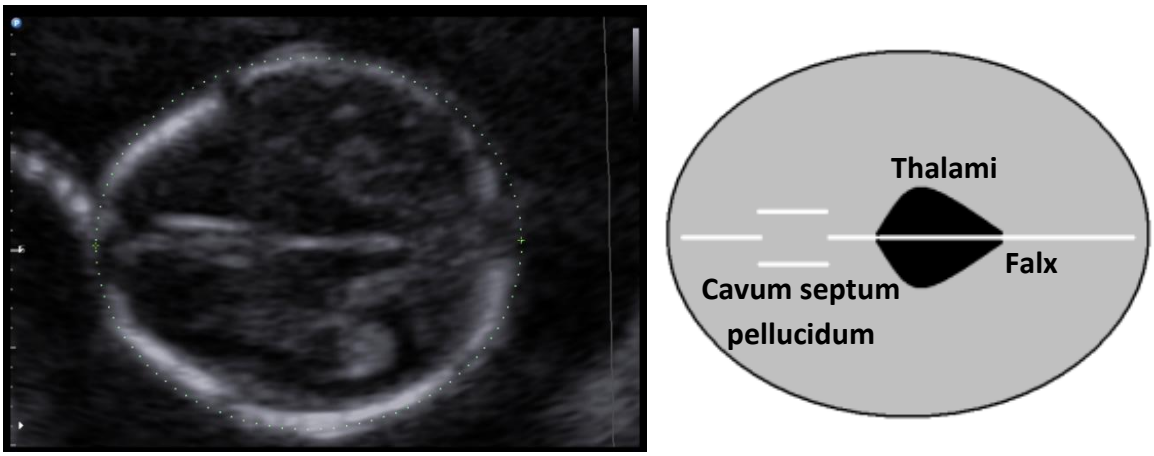
and experienced in that type of evaluation. In contrary, the objective score-based method allowed for good inter-reviewer reproducibility that showed good agreement in the scoring. However, they noted that extensive qualitative analysis would require significant human resources which is time-consuming and expensive.

Salomon's image scoring scheme has been adopted in the qualitative quality control component of INTERGROWTH-21st study. Dudley and Porter (1993) demonstrated the value of using a similar scoring system as a feedback mechanism for training and coaching the sonographers in their hospital. After receiving the feedback on the criteria that was not met, the sonographers achieved an improvement in the recognition of the relevant features and the technical skills required to meet the quality criteria. Quality assurance and control of the Ultrasound component of INTERGROWTH-21st study.

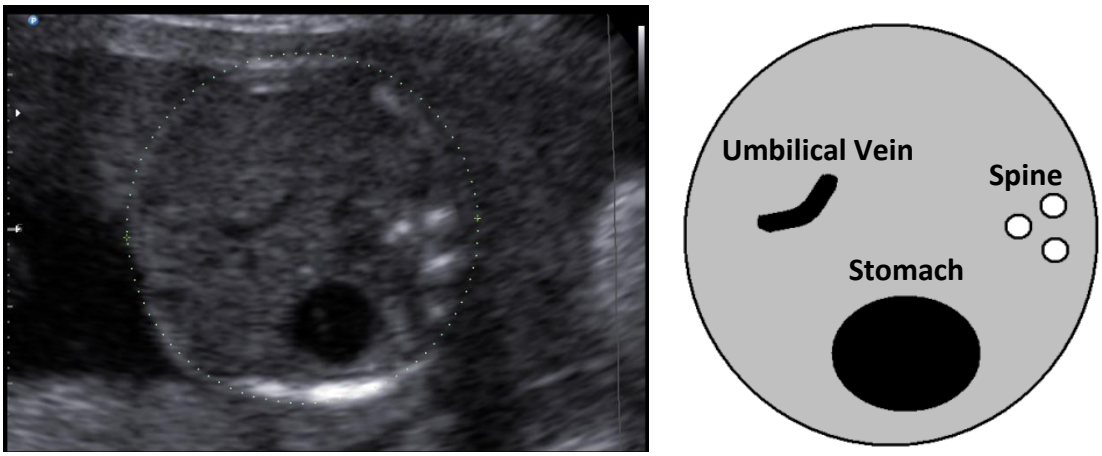
The objective scoring in Salomon's work was performed according to pre-defined criteria summarized in Table 2.1. Each criterion scores one point when correct and from the summed score, the image is classified as good in agreement (difference of $\leq$ 1 point), moderate in agreement (difference of 2 points) and poor in agreement (difference of $>$ 2 points) to the requirement of suitable frame for fetal biometry measurement. Examples of fetal head and fetal abdomen ultrasound images that satisfy the scoring are shown in Figure 2.6.

Table 2.1: Objective scoring system for still images (Salomon et al., 2006)

Fetal head measurements	Fetal abdominal measurements	Fetal femur measurement
Symmetrical plane	Circular plane	Both ends of the bone clearly visible
Plane showing thalami	Image shows the stomach bubble	<45◦ angle to the horizontal
Cavum septum pellucidum 1/3 along midline echo	Image shows umbilical vein along 1/3 of the abdomen	Femoral plane occupying at least 30% of the total image size
Cerebellum not visible	Kidneys not visible	Callipers placed correctly
Fetal head occupies at least 30% of the total image size	Abdomen occupies at least 30% of the total image size	
Callipers and dotted ellipse placed correctly	Callipers and dotted ellipse placed correctly	



(a)



(b)

Figure 2.6: Examples of fetal ultrasound scan that satisfy Salomon's scoring and diagram showing visible landmarks in standard fetal (a) head and (b) abdominal planes.

In summary, the image scoring system is an important element for reducing intra- and inter-operator variability in fetal biometric measurements. However, visual scoring of ultrasound images has proven to be a tedious and time-consuming process and also subjected to bias by the observer. Therefore, an automated image assessment with well-defined criteria could overcome the limitations of manual objective scoring.

2.6 Automated Image Analysis in Obstetric Ultrasound

In this section, we review prior relevant research on fetal ultrasound automated analysis. By using the SCOPUS database (www.scopus.com), 19 key papers[2] were identified. The published works have been grouped into two application categories: standard fetal biometric measurement and "other applications".

2.6.1 Automated Fetal Biometric Measurement

The majority of automatic methods developed for fetal ultrasound image analysis had focused on automation for standard fetal biometry segmentation and measurements. The variability introduced by the human operator was identified as motivation for producing standard measurements through automatic approach. A summary of prior relevant research is given in Table 2.2.

Initially, the approaches for automatic fetal anatomical segmentation in ultrasound images were mostly based on morphological operators (Thomas et al., 1991, Zador et al., 1991, Matsopoulos and Marshall, 1994, Hanna and Youssef, 1997, Lu et al., 2005). A series of steps such as edge detection, edge linking and the Hough transform were used in order to achieve head and femur segmentation. The segmentation results showed correlation coefficients larger than 0.97 when compared to the measurements provided by experts.

[2] 15 June 2012

However, such methods are heavily dependent on the threshold value selected in the pre-processing stage which can greatly differ from one image to another.

Chalana et al. (1996) and Pathak et al. (1997) employed active contour model for fetal head and abdomen segmentation. The disadvantage of the active contour method is that it can get stuck at local minima, which might require manual correction. Jardim and Figueiredo (2003, 2005) presented a method for the segmentation of femur and skull, based on the evolution of a parametric deformable shape. Pixels are grouped into two regions according to similarity in textures using the parametric model defined by the Rayleigh distribution. The authors noted that this algorithm does not guarantee an optimal solution and also requires an initial guess from the user, which makes the system semi-automatic.

Carneiro et al. (2008b) exploited the database-guided segmentation technique with discriminative constrained probabilistic boosting tree classifier for the detection and measurement of head, femur and abdominal structures. A large database of expert annotations of structures of interest was used for training the probabilistic boosting tree classifier, where the nodes of the binary tree are strong classifiers trained using AdaBoost. The method was extended to measure the humerus length (HL) and the crown-rump length (CRL).

Yu et al. (2007, 2008a, 2008b) experimented with the gradient vector field (GVF) method for fitting ellipses to the edges of head, femur and abdomen. The ellipse was initialized using the Hough transform. However, they found the GVF method to be erroneous due to its sensitivity to the edges of other structures, and it often generated poor results in the case of improper initialization.

2.6.2 Other Measurements

There have been a few related works in fetal 3D ultrasound. Nguyen et al. (2006) focused on extracting the frontal surface of a fetus automatically from a 3D fetal ultrasound

volume using a support vector machine (SVM) based texture classification. Experimental results showed that some amount of abdomen wall and fluids managed to be removed for better visualization of the fetus frontal surface. Anquez et al. (2008) presented a method for the segmentation of the 3D fetus by modelling the intensity of pixels belonging to fetal and amniotic fluid with a Rayleigh distribution model and an exponential distribution model, respectively. Carneiro et al. (2008a) extended their probabilistic boosting tree method that was used on 2D images to 3D ultrasound volume of fetal head, with the aim of finding anatomies in ultrasound volumes based on semantic keywords. The method presented in this work do not select the standardized plane using any pre-defined clinical protocol but it was left for the user to decide based on the selected anatomical landmarks in fetal brain (i.e. cerebellum, cistern magna).

Gooding et al. (2010) investigated the use of fusion methods to improve the quality of volumetric fetal cardiac imaging. Data were acquired from seven volunteers with gestation ages in the range 24 to 30 weeks. Multiple 4-D scans of the heart acquired at arbitrary orientations were aligned and then combined/fused into a single image. The results presented show reduction in the appearance of image artefacts and improvement in the contrast-to-noise ratio. They also investigated the effect of fusion on the reproducibility of left ventricle segmentation using VOCAL^{TM} software (GE Kretz, Zipf, Austria). The variability of volume estimates on fused image was found to be reduced by about 50% relative to measurement on a single scan.

Recent work by Yaqub et al. (2010b) proposed the use of Random Forest (RF) method to segment the femur in 3D fetal ultrasound volume. Their approach that used weighted class decision from each tree in RF outperformed the conventional RF method for the accuracy of the segmentation. Considering the low contrast of the 3D ultrasound volumes, they also experimented with a wavelet-based registration method to fuse multiple fetal femur

volumes to enhance the femur boundary definition (Yaqub et al., 2010a). They found that fused view volumes have better boundary definition of the femur compared to the single view images. This finding is found through the visual scoring result by an experienced sonographer who gave higher scores for the fused images compared to single view images.

Rueda et al. (2011) investigated the problem of segmenting the adipose tissue on the fetal arm which is proposed as a fetal nutrition indicator. The segmentation approach was based on the fuzzy connectedness framework with an affinity function using structural and edge information extracted from local phase features. Quantitative and qualitative results presented showed the superiority of the method compared to the framework that was based on intensity images only.

Table 2.2: Summary of research in fetal ultrasound image analysis in chronological order.

Author, Year	**ROI**	**Dataset**	**Techniques**	**Validation**
Standard Fetal Biometric Measurements				
(Zador et al., 1991)	Head	75 images (from frozen video)	Threshold - Edge detection (gradient operator) - Hough transform	BPD (r = 0.986, 1.87±1.94mm) OFD (r = 0.958, 2.82±4.13mm) HC (r = 0.972, -0.36±9.87mm) 8 sec on 10 MHz IBM
(Thomas et al., 1991)	Femur	24 scanned images	Morphological operators Pre-processing – enhancement – thresholding – extraction - opening/closing - skeletonisation	r = 0.9985, MSE 0.8133 10 minutes (PC spec not mentioned)
(Matsopoulos and Marshall, 1994)	Head	1 image (video)	Morphological operators Pre-processing - threholding – erosion and dilation – thinning – growing-filling – erosion - opening	Error = 1.050%
(Chalana et al., 1996, Pathak et al., 1997)	Head	35 images	Active contour model Specify initial points - Pre-processing – Cubic spline fitting - Active contour algorithm	BPD (r = 0.99, |mean| = 1.41) HC (r = 0.994, |mean| = 2.99) 32 s on Sun SparcStation 20/71 248 ms on MS5000 system

(Hanna and Youssef, 1997)	Head	Not specified	Morphological operators Pre-processing – de-noising - enhancement - threshold – ellipse fit - closing	BPD (r = 0.994) HC (r = 0.985) 4 minutes on 66 MHz PC
(Jardim and Figueiredo, 2003)	Head, Femur	50 pair images	Parametric deformable shape model	Qualitative. Difference of 1 day in estimate
(Lu et al., 2005)	Head	217 images	K-means classifier and Iterative Randomized Hough transform (IRHT)	BPD (r = 0.997, 0.12%) HC (r = 0.993, -0.52%) 1.6 s on 1987-MHz Athlon
(Carneiro et al., 2007, Carneiro et al., 2008b)	Head, Femur, Abdomen Body Humerus	Training (1426 head, 1168 femur, 1293 abdomen, 547 humerus, 325 body), Testing (30 for each structure)	Constrained probabilistic boosting tree (CPBT)	Average measurement from 15 experts as gold standard. 0.5 sec on dual-core PC Other quantitative measurement – see reference
(Yu et al., 2008a, Yu et al., 2008b)	Head Femur Abdomen	215 cases (with 103 set for fetuses delivered within 3 days)	De-noising (Anisotropic diffusion) – Fuzzy C-Means Clustering – Randomized Hough transform – Gradient vector field (GVF) snakes	BPD (r = 0.991, 2.89%) HC (r = 0.983, 1.85%) FL (r = 0.983, 3.34%) AC (r = 0.984, 7.53%
Other Automatic Fetal Ultrasound Measurements				
(Nguyen et al., 2006)	Whole fetus	Training (400 fetus) 40 volume datasets	Support vector machine (SVM) texture classification	Qualitative
(Anquez et al., 2008)	Whole fetus	4 volume datasets	Rayleigh and exponential distribution modelling	Quantitative result for only 1 dataset with 72% correctly classified pixels
(Carneiro et al., 2008a)	Fetal head	200 volumes from 13-35 weeks fetuses	Probabilistic boosting tree (PBT) algorithm	Error similar to inter-user variability. Performs under 10 sec on dual-core PC 1.7 GHz
(Gooding et al., 2010)	Fetal heart	6 volumes from 24-30 weeks fetuses	Fusion methods (mean, median, mean shift, maximum intensity, wavelet coefficient)	Δ(contrast) = 37.60% Δ(SNR) = 18.82% Δ(CNR) = 33.23% Δ(Std deviation) = 6.60
(Yaqub et al., 2010a, Yaqub et al., 2010b)	Femur volumes	20 volumes from 19 weeks±6 days fetuses (Cross validation - Training 18, Testing 2)	Random Forest (RF) with weighted voting tree – 3D Haar features – connected component	Recall 70±15% Precision 88±11% Bland-Altman plot

(Rueda et al., 2011)	Fetal arm	7 images from 21-40 weeks fetuses	Fuzzy connectedness with affinity function using local phase and feature asymmetry	Precision 93.51±1.91% Recall 82.77±5.74% Dice similarity 87.69±3.05%

2.6.3 Summary

Most of the fetal ultrasound image analysis literature is confined to the problems of (semi)-automatic segmentation of specific anatomical structure for accurate biometric measurement. All prior literature had assumed that the plane selected is correct, which may not be the case and will lead to a major source of error in practice.

2.7 Machine Learning in Medical Imaging

Machine learning can provide an effective way to automate the analysis for medical images. Early applications of machine learning in the biomedical field originated from the development of artificial neural networks (ANNs), where the concepts have been inspired by studies of human neurons (McCulloch and Pitts, 1943, Bishop, 1996). Examples of the applications of machine learning in medical imaging include *computer-aided detection and diagnosis systems* (e.g., CT lung structures (Ochs et al., 2007), mammography (El-Naqa et al., 2002), CT colonography (Wang et al., 2010)), *medical image segmentation* (e.g., lung (Prasad et al., 2008), femur volume (Yaqub et al., 2010b), endocardial borders (Wang et al., 2011)), *medical image registration* (e.g., brain MRI (Wu et al., 2006), multi-modality registration (Jiang et al., 2008)), and *content based image retrieval systems* (Reddy and Bhuyan, 2008, Huang et al., 2010, Simonyan et al., 2011).

The problem of accurate and robust detection of visual classes using learning-based techniques in medical imaging has not been extensively researched compared to its

application in the natural image domain. Applications of machine learning in natural images include various type of object detection (i.e. face (Viola and Jones, 2004), people (Ioffe and Forsyth, 2001), car (Schneiderman and Kanade, 2000)), face recognition (Turk and Pentland, 1991), handwritten digit recognition (Amit and Geman, 1999), and etc.

Machine learning algorithms used for object detection are mainly confined to supervised discriminative learning approaches. A d*iscriminative model* tries to find a mapping from input variables to an output variable, with the goal of discriminating between classes, rather than modelling a full representation of a class (*generative model*). Training samples are labelled, where each sample contains two parts: one is input observations or features and the other is output observations or labels (Hastie et al., 2009). A supervised discriminative learning algorithm analyses the labelled training data and produces an inferred discriminating function (a *classifier*), usually in the form of equations and numerical coefficients or weights. Some of the common examples of these algorithms that have found its application to object detection in medical imaging are AdaBoost (Freund and Schapire, 1997), Support Vector Machine (SVM) (Vapnik, 1995), etc.

The formulation of SVM learning is based on the principle of structural risk minimization. SVM attempts to minimize a bound on the generalization error i.e. the error made by the learning machine on test data not used during the training. El-Naqa et al. (2002) used the SVM approach for the detection of clustered microcalcifications (MCs) in mammograms. They demonstrated that such an approach could outperform several well-known methods in the literature, such as the image difference technique (IDT), the difference of Gaussian (DoG) method, the wavelet-decomposition based method and a multilayer neural network method. However, the computational complexity of the SVM classifier, both in training and testing can prove to be burdensome for real-time or near real-time applications (Burges, 1998). In an SVM classifier, the decision function is determined by a subset of

training samples (called support vectors), and the computational complexity of the decision function is linearly proportional to the number of support vectors. Too many support vectors can lead to a classifier that is computationally expensive. There is also the issue of optimal feature selection in an SVM implementation. Failure to discard irrelevant features affects the classification accuracy, computational efficiency and learning convergence of the algorithm. Various feature selection strategies had been proposed for SVM (Hermes and Buhmann, 2000, Cao et al., 2007, Moghaddam et al., 2007). However, separately performing these two steps might result in a loss of information relevant to classification tasks (Nguyen and de la Torre, 2010).

Boosting is a technique introduced by Freund and Schapire (1997) for combining multiple 'weak' classifiers to produce a strong classifier whose overall performance can be significantly better than that of any of the individual classifiers. A boosting technique produces highly accurate results even if the weak classifiers have a performance that is only marginally better than random, hence the term 'weak learners'. The weak classifiers are trained in sequence, and each classifier is trained using a weighted distribution of the data set. Adaptive Boosting (AdaBoost) is a widely used form of boosting. It is adaptive in the sense that subsequent classifiers are adapted in favour of samples misclassified by previous classifiers. This is done by assigning greater weights to the samples that were misclassified by one of the earlier classifiers, when training for the next classifier. Once all classifiers have been trained, their predictions are combined through a weighted majority vote. It is found that by iteratively combining weak classifiers in this way, the training error converges to zero quickly (Schapire et al., 1998).

Advantages of the AdaBoost algorithm are that it has no parameters to tune other than the number of iterations. Boosting provides an efficient algorithm for selecting a small number of highly relevant features from a very large number of potential features and the

feature selection process is done simultaneously during the training process, requiring no additional experiments. Tieu and Viola (2004) used boosting to select a small set of discriminatory sparse features out of a possible set of over 45,000 for tracking various different objects in natural scene. Inspired by the work of Tieu and Viola, Viola and Jones (2004) used boosting to build an extremely efficient face detector dependent on a small set of features.

Another important issue tackled through AdaBoost is the imbalanced class problem representation. In medical imaging, it is common to find some classes are represented by a larger number of samples than other classes. Direct classification may be biased towards the majority classes and result in poor performance on the minority classes. Boosting is therefore a feasible technique in tackling the class imbalance problem through the modification of the underlying data distribution and classifying in the re-weighted data space iteratively (Sun et al., 2006). This re-weighting exercise also supports a relevance feedback for the user where training examples with high weight can be used to flag potentially mislabelled examples.

While AdaBoost had proven to be an effective method for object detection in natural scenes (Viola and Jones, 2004, Isukapalli et al., 2006, Whitehill et al., 2009), it has also found application for object detection in medical images (Yuanzhong et al., 2006, Ochs et al., 2007, Zhou et al., 2007).

In a paper by Yuanzhong et al. (2006), AdaBoost classifiers were trained to classify true and false boundary positions of liver tumors from 30 Computed Tomography (CT) training images and tested on a separate 30 CT images. The authors claimed that the approach worked well despite the diverse intensity distributions of the tumor regions and tumor region shape variability. Only qualitative assessment was given in the paper.

Ochs et al. (2007) presented an automatic classification of lung bronchovascular anatomy in chest CT images using the AdaBoost learning algorithm. A set of ensemble classifiers (composed of 20 weak classifiers) were trained with AdaBoost to detect voxel parts of a specific structure: airway, fissure, nodule and vessel structure of airway. Their feature set consisted of voxel attenuation and a small number of features based on the eigenvalues of the Hessian matrix. The result on diverse dataset (29 chest CT scanned with different scanner models from 4 manufacturer) was shown to be promising with the AUC values for all the airways structures classifiers were between 0.931 to 0.984. However, there is no mention as to whether the training dataset was separate from the testing dataset.

Another implementation on clinical chest CT data for detecting the vessel bifurcation points also employed the AdaBoost algorithm in a detection framework (Zhou et al., 2007). Features were derived from first derivatives and second derivatives of 2D Gaussian filters to capture low frequency information, high frequency information (i.e., edges) and the local maxima (i.e., ridges). The method was trained with 100 positive and 100 negative examples and tested on 50 positive and 53 negative examples. The paper compared the performance of an AdaBoost classifier (with 20 iterations) to other classifiers; k-Nearest Neighbor classifier with a Euclidean distance measure between input images ($k = 3$), a naïve Bayes classifier, a neural network (1 hidden layer, learning rate of 0.3) and a support vector machine. The result showed that Adaboost is superior to the other methods with the lowest mean error rate of 3.4%.

In ultrasound imaging, Karavides et al. (2010) attempted to automatically detect the standard anatomical landmarks points (apex and mitral valve points) in the four-chamber and two-chamber views in echocardiogram using the AdaBoost algorithm. A rectangular box (sizes and orientations are not mentioned in the paper) were scanned over the 2D image and the image content of the box were represented using Haar features. Three multilevel

classifiers were used to detect the center of the left ventricle, the apex and the mitral valve center. The classifiers were trained using positive and negative image examples (no value given) generated from 60 patients 3D datasets while the testing were done using 25 patients datasets. The detection errors were calculated by measuring the Euclidean distance of detected and manually annotated landmarks. They found the detection approach using AdaBoost to be promising where the errors recorded by the method (7.5±3.3mm and 5.0±2.5mm for apex and mitral valve center, respectively) were comparable to the inter-observer and intra-observer variability of manual marker identification (7.1±2.9mm and 3.8±1.3mm for apex and mitral valve center, respectively).

2.8 Conclusions

This chapter provided a general overview of the current practice in fetal growth monitoring and pregnancy management using biometric measurements from ultrasound images and technical work on fetal ultrasound segmentation and detection. This thesis now sets out to consider the detection of structures in a fetal abdominal scan by using machine learning methods.

Chapter 3 Anatomical Object Detection in 2D Fetal Abdominal Ultrasound Image

In this chapter, we present the implementation of anatomical landmark detection in ultrasound images using a machine learning framework with a view of enabling a fast and more reproducible automated quality assessment of ultrasound image plane. We focus on the detection of the stomach and the umbilical vein in fetal abdominal scan. This detector is inspired by the approach of Viola and Jones (2004) who considered the problem of face detection. We have chosen to build a model using statistical learning given positive and negative training examples. A learning algorithm trains a classifier by selecting visual features. We chose to apply a learning technique known as Adaptive Boosting (AdaBoost) because of its dual ability of creating a strong classifier and feature selection. Our choice was also motivated by the Adaboost's ability for high accuracy classification during the detection task.

We introduce the problem of the detection of two anatomical landmarks in fetal ultrasound abdominal scans in Section 3.1. The relevant background on the machine learning framework used for object detection is described in Section 3.2. The experimental setup which includes the modules used in the detection, the datasets and the validation measures are given in Section 3.3. Quantitative and qualitative experimental results are presented and discussed in Section 3.4. Finally, we end the chapter with concluding remarks in Section 3.5.

3.1 Introduction

The detection of two anatomical landmarks in the abdominal image plane, specifically the stomach bubble and the umbilical vein is an important step in identifying the standard plane for taking the abdominal circumference (AC) measurement. Historically, the plane consisting of these two landmarks was described in the procedure proposed during the early introduction of using the abdominal circumference from ultrasound images as the biometric markers for fetal weight estimation (Campbell and Wilkin, 1975). It was also used in constructing the widely used chart for abdominal circumference size by Chitty et al. (1994).

In Figure 3.1, the position of the ultrasound probe used to acquire the standard abdominal image plane for measuring the abdominal circumference (AC) is illustrated. Two examples of the correct section selection for AC measurement are presented in Figure 3.2 showing the main anatomical landmarks: the stomach and the umbilical vein. In contrast, poor section selections are shown in Figure 3.3. In the 16 week fetus, the stomach is not visible and in the 38 week fetus, the entire elongated intra-abdominal umbilical vein is shown indicating that the plane is too oblique.

The aim of our work is to use a machine learning approach for the detection of the stomach and the umbilical vein in the ultrasound images. We chose to explore the established detection method based on a machine learning algorithm known as AdaBoost coupled with a set of Haar features extracted from the intensity image. Other features are considered in later chapters.

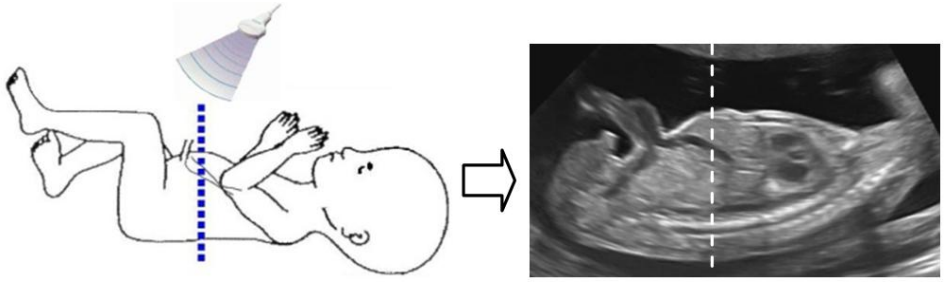


Figure 3.1: The position of the ultrasound probe for taking the standard abdominal image plane is represented by dotted lines.

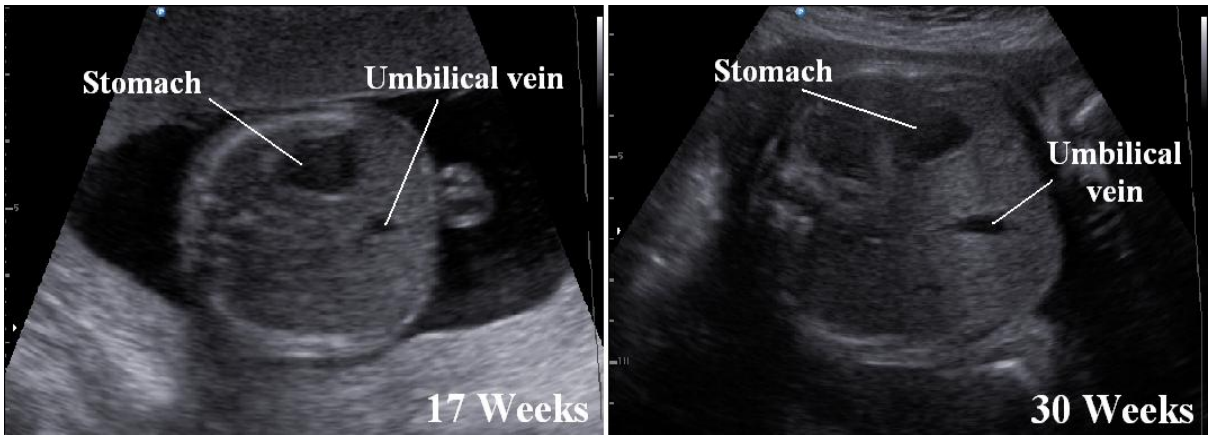


Figure 3.2: Examples of good section for fetal abdominal circumference in 17 weeks and 30 weeks fetuses.

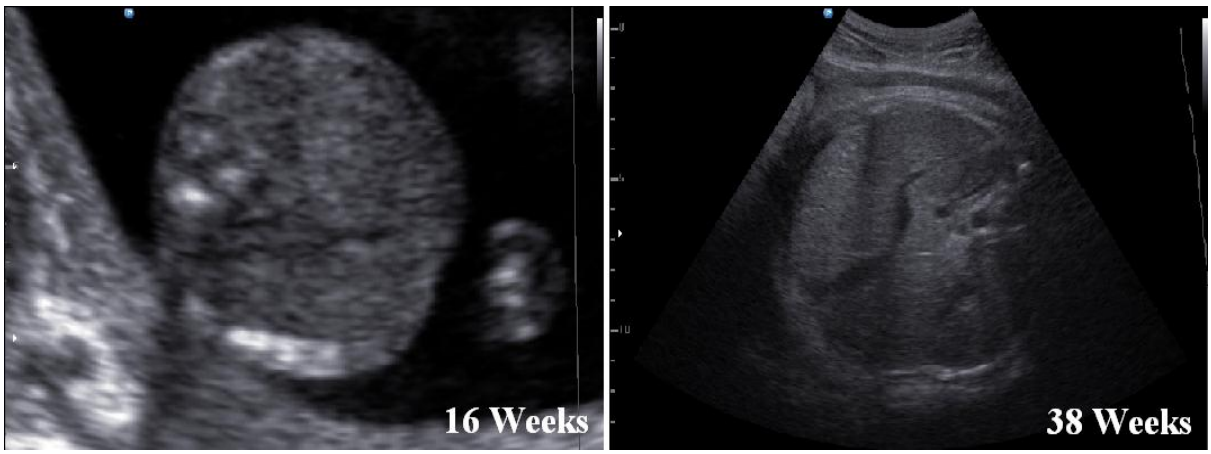


Figure 3.3: Two examples of the wrong section for fetal abdominal circumference in 16 weeks (stomach is invisible) and 38 weeks fetuses (umbilical vein is elongated).

3.2 Background on the Object Detection Framework

The implementation of object detection using a machine learning approach in our application is formulated as a supervised discriminative model where a classifier model is provided only for the target variable(s) conditional on the observed variables. A binary classifier is meant to discriminatively distinguish the object from the background. A classifier usually consists of a feature and its threshold. Several types of features have been used for image representation such as image intensities, intensity gradient magnitudes, intensity curvatures, filter-banks (Shen et al., 2003), steerable filters (Greenspan et al., 1994), etc.

Essentially, there are two motivations for using features instead of pixel intensities directly in our implementation. Firstly, features encode domain knowledge better than pixels. Significant variability in the patterns within the stomach and the umbilical vein boundaries and also the absence of constraint on the background make the direct analysis of pixel intensity inadequate. The second reason is that a feature-based system can be much faster than a pixel-based system (Viola and Jones, 2004).

3.2.1 Haar Features

Haar features and their variants are commonly found as an input to machine learning frameworks for object segmentation or detection as the Haar features provides rich image representation for effective learning and efficient implementation. They were first proposed for the detection of pedestrians in images (Oren et al., 1997) and later adapted by others for many different applications, such as face detection (Viola and Jones, 2004), hand gesture recognition (Qing et al., 2007), etc. In medical image analysis, applications include x-ray image retrieval (Reddy and Bhuyan, 2008), heart region detection (Pavani et al., 2010), ultrasound femur segmentation (Yaqub et al., 2011), etc. The features are based on Haar

wavelets which are single wavelength square waves (one high interval and one low interval). In two dimensions, a square wave is a pair of adjacent rectangles - one light and one dark.

Figure 3.4 shows the set of Haar features (two-, three- and four- rectangles) and the unary feature used in this experiment. The features are calculated by finding the difference of the pixels in the white region(s) from the pixels in the black region(s). The regions have the same size and shape and are horizontally or vertically adjacent. The task of the learning algorithm is to identify a set of features that consistently distinguishes the object of interest (stomach/umbilical vein) in the ultrasound image.

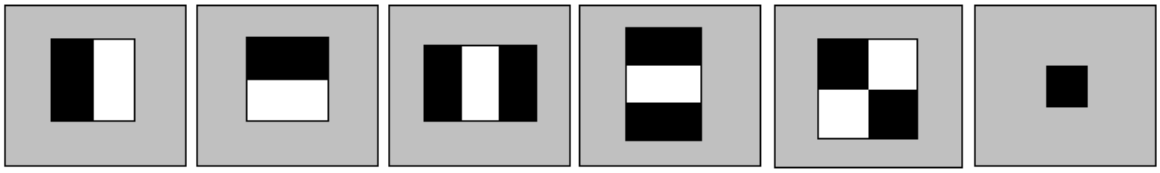


Figure 3.4: Prototypes of Haar features and unary feature used in our algorithm.

3.2.2 Integral Image

The integral image is an image representation that acts as a means for fast feature extraction and is proven to be an effective way to speed up the detection task (Viola and Jones, 2004). The integral image ii of an image I is defined as:

$$ii(x,y) = \sum_{x' \le x, y' \le y} I(x',y') \tag{3.1}$$

where $ii(x,y)$ is the integral image and $I(x,y)$ is the original image. The value of the integral image at the coordinate (x,y) is the sum of all the pixels above and to the left of (x,y), including itself.

The speed advantage of using an integral image lies in the fact that any rectangle sum in an image can be calculated from the corresponding integral image, by indexing the integral

image for only four times. In Figure 3.5, a rectangle specified by four points, with upper-left point as (x_1, y_1) and lower-right point as (x_2, y_2). The sum of pixel values which lies within rectangle $A(x_1, y_1, x_2, y_2)$ is evaluated by four integral image references (see Equation (3.2).

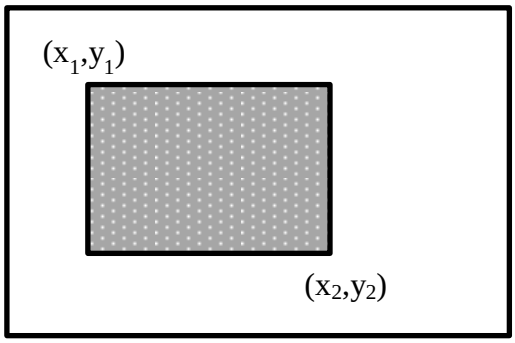


Figure 3.5: Example of integral image application.

$$A(x_1, y_1, x_2, y_2) = ii(x_1, y_1) + ii(x_2, y_2) - ii(x_1, y_2) - ii(x_2, y_1) \quad (3.2)$$

The two-rectangle features are computed with six references because the two rectangles are adjacent. The three-rectangle features require eight references and the four-rectangle features require only nine.

Even if the features are very simple to compute with the use of integral image, applying the complete set of 57,315 features (the derivation of the number is shown in Section 3.3.2), which is far larger than the number of pixels, during detection would be too computationally expensive and lead to over-fitting. Thus, the next stage in building the object detector is to use a learning function which selects a small set of rectangle features: the ones which best separate the positive and the negative samples. In this system, AdaBoost is used to both select the features and to train the classifier.

3.2.3 Adaptive Boosting (AdaBoost)

Boosting is the process of forming a strong hypothesis through linear combination of weak hypotheses. The motivation of boosting is derived from the observation that finding

several simple rules for classification can often be easier than finding a single highly accurate rule. Freund and Schapire (1997) formulated an iterative learning method known as Adaptive Boosting (AdaBoost) that allows the designer to keep adding "weak" classifiers until some desired (low) training error has been achieved. They presented the scenario of an expert gambler who would be unable to articulate a grand set of rules for selecting a winning horse in horse racing event but would have no trouble coming up with a "rule-of-thumb" when presented with the data for a specific set of races.

In the context of object detection which is a binary classification task, a weak hypothesis can be represented as the "weak classifier" that is derived from the extracted set of features. Such classifiers are called 'weak' because even the best classification function is not expected to classify the data well, they only need to classify correctly the examples in more than 50% of the cases.

A weak classifier h_j, where j is the jth classifier of a possible N, is a simple structure containing a feature f_j, threshold θ_j and parity p_j. The output of a weak classifier is binary and is defined below:

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j \theta_j \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

where θ_j is a threshold value that separates the value of feature f_j of positively labelled images (foreground) from the negatively-labelled images (background), and p_j is 1 if the positive examples are classified below the threshold or -1 if the positive examples are classified above the threshold.

The individual classifier thresholds and the parity of each classifier are determined across a large image dataset. We set the threshold value using the following equation:

$$\frac{1}{2}\left(\frac{1}{|C_o|}\sum_{x\in C_o} f_j(x) + \frac{1}{|C_1|}\sum_{x\in C_1} f_j(x)\right) \tag{3.4}$$

where C_o is the set of negative examples and C_1 is the set of positive examples. The parity p_j is the sign of the difference of the two means in the above equation.

AdaBoost is then used to concurrently select and combine relevant features from the feature set during the training of classifier, thus avoiding a separate feature selection process common with other classification methods. The implementation of AdaBoost is outlined in Table 3.1.

The main idea behind the use of AdaBoost is the application of a weight distribution to the sample set which is modified at every iteration. The weights of all the sample sets are first normalized and its distribution is flat at the beginning (Equation (3.6)). After each iteration, it is modified according to the hypothesis returned by each of the weak learners (Equation (3.7)). This modification indicates that the new weight is a product of its current weight and a factor β_t which in turn is a product of the minimum error classifier that is chosen. If classification by the weak classifier is correct for the sample, then the weight of the sample is reduced (seen as an *easier sample*), otherwise there is no change to its weight. This implies that samples from the data set that are incorrectly classified obtain a higher weight value (seen as *difficult sample*). It is this property of the AdaBoost algorithm which emphasises the bad classification (or rather hard to classify samples) that drives down the classifier error.

Once AdaBoost algorithm makes its choice, the weak classifier is assigned a measure of importance $\alpha_t = 0.5 * ln(1 - e_t)/e_t$ where e_t is the training error of the weak classifier at round t. After successive iterations of the algorithm the result is a weighted linear combination of T hypotheses where the ones with lowest classification error obtain higher

weighting (more importance). The final strong classifier is shown in Equation (3.8), where the binary classification is made based on the sum of the individual weights α (these represent the threshold) and the summed product of each weak classifiers weight and classification ($\alpha_t h_t(x)$) of a particular sample, x.

Table 3.1: AdaBoost Algorithm (modified from (Viola and Jones, 2004)).

- **Input:** Assume n sample images $(x_1 \ldots, x_n)$, and associated labels $(y_1 \ldots, y_n)$, where $y_i \in \{0,1\}$. y_i= 0 denotes a negative sample and y_i= 1 a positive one. m is the number of negatives samples and $l = n - m$ the number of positive samples.

- **Initialise:** Set the n weights to:

$$w_{1,i} = \begin{cases} (2m)^{-1} & if y_i = 0 \\ (2l)^{-1} & if\ y_i = 1 \end{cases} \quad (3.5)$$

- **For $\boldsymbol{t = 1, \ldots, T}$** (where T is the total number of weak classifiers):
  1. Normalize the weights,

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \quad (3.6)$$

  so that w_t is a probability distribution.

  2. For each feature *j*, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to the $w_{t,i}$'s as $e_j = w_{t,i}|h_j(x_i) - y_i|$.

  3. Choose the "weak" classifier h_t as the h_j with the lowest error e_j. Set e_t to that e_j.

  4. Update the weights:

$$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i} \quad (3.7)$$

  where $e_i = 0$ if example $x_i$is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{e_t}{1-e_t}$.

- **Output:** The final "strong" classifier *H(x)* is:

$$H(x) = \begin{cases} 1 \ if \ \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T} \alpha_t \\ 0 \qquad\qquad otherwise \end{cases}$$

  where $\alpha_t = log\frac{1}{\beta_t}$ (3.8)

The main objective of AdaBoost is to minimize the error over the training set. Freund and Schapire (1997) proved that the upper bound of the training error e_{tr} is $exp(-2\sum_t(0.5 - e_t)^2)$. The training error therefore drops exponentially with respect to the number of training rounds T. As long as the best base classifier is at least better than random, the upper bound on the error is $exp(-2T(0.5 - e_t)^2)$.

3.3 Experimental Setup

The anatomical object detection setup in our experiment was comprised of four modules which are the image, feature extraction, learning algorithm and detector modules, as described below:

i) **Image module:** Its primary role is to upload an image and present it in an appropriate form (cropped samples for learning module or extracted abdomen for detector module) and to provide the integral image for efficient feature extraction.

ii) **Feature extraction module:** This produces the array of weak classifiers (feature and its threshold) to be passed onto the learning algorithm or the detector module.

iii) **Learning algorithm module:** In our experiment we employed the AdaBoost algorithm.

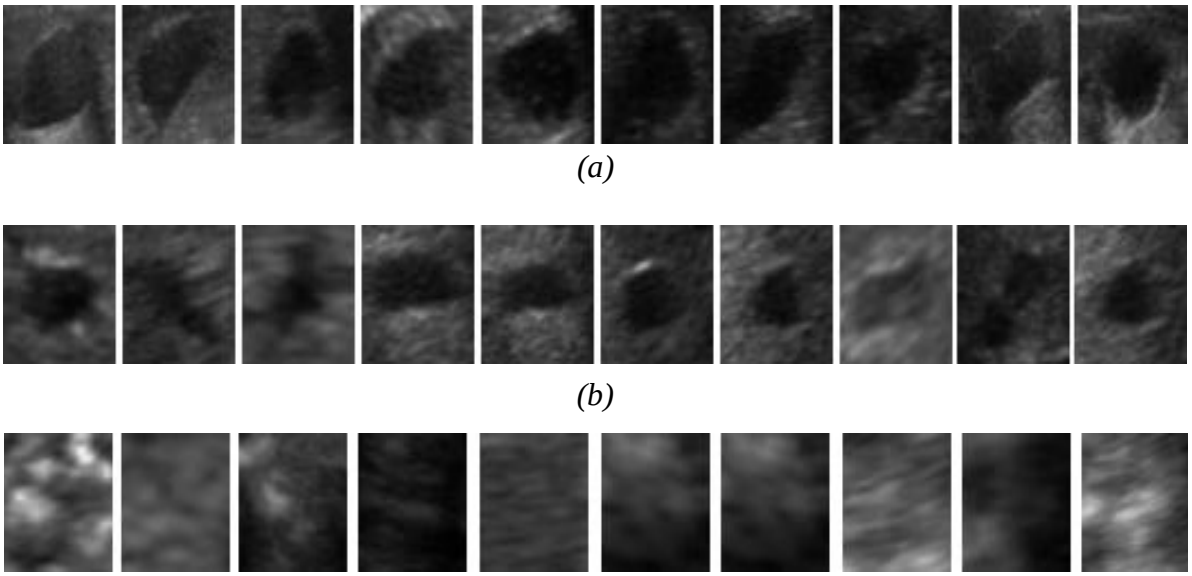
iv) **Detector module:** This uses the sliding window method (Viola and Jones, 2004) and uses the strong classifier produced by the learning module.

There are two stages in the development of an object detector. The first stage is the production of the strong classifier (training) where the image (training samples), feature extraction (as discussed in Section 3.2.1) and learning algorithm are used. The second stage is the detection process where only (test) image, feature (as indicated by the trained classifier) and detector components are needed.

3.3.1 Image Module

For a supervised learning algorithm, a training set of positive and negative samples are needed. The positive samples containing the objects of interest (stomach and umbilical vein) were cropped manually from the abdominal images. For the negative samples, we partitioned the ultrasound images into identically-sized sub-windows and then manually removed any sub-windowed images that appeared to contain the stomach or the umbilical vein.

The images were cropped loosely around the stomach and the umbilical vein to contain any additional visual information surrounding it. Our initial experiment with tightly cropped images for training resulted in poorer detection of the objects. Some typical stomach, umbilical vein and background examples used for training in our experiment are shown in Figure 3.6.



(a)

(b)

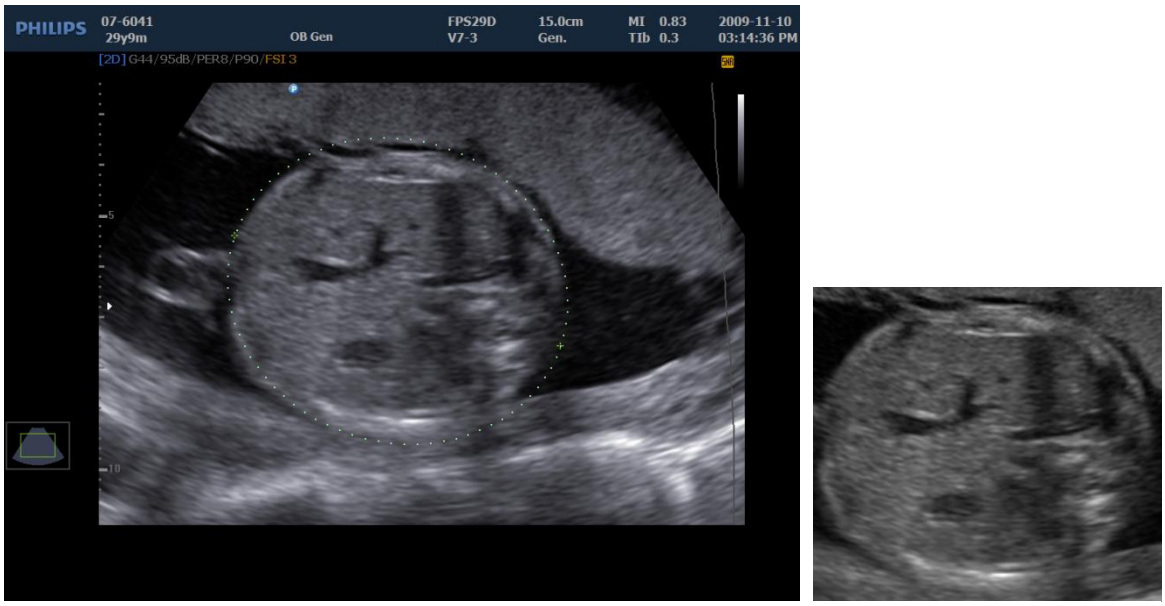
(c)

Figure 3.6: Example of (a) stomach images (b) umbilical vein images and (c) background images used for training.

We set the base resolution of the training samples as 100x100 pixels for the stomach and 75x75 pixels for the umbilical vein. The same values were used as the initial size of the

sliding window for detection. The values were selected because it provides an adequate size to capture the amount of information stored in the stomach (or umbilical vein) and surrounding, and also it corresponds to the value of the smallest stomach and the umbilical vein in our data collection.

For the image in the detector module, the original ultrasound image (Figure 3.7a) was pre-processed in order to extract the abdominal area. We applied an ellipse fitting algorithm by using the calliper points positioned around the fetus abdominal area in the image to extract the fetus abdominal area (Figure 3.7b).



(a) (b)

Figure 3.7: Fetal abdominal area extraction from original image using ellipse fitting.

We did not apply any normalization method on our training and testing images as proposed for the application in natural images. Normalization used in natural images is commonly based on the assumption of a Gaussian distribution model and this is not suitable for ultrasound images because of the speckle noise patterns and also significant variation in signal drop-out between images. Our experiment with variance normalization used by (Viola and Jones, 2004) on the ultrasound image produced an inferior result on detection accuracy.

3.3.2 Feature Module

The detailed implementation of feature extraction in our experiment is shown in Table 3.2. The initial size refers to the base width and length of the rectangles. The rectangles were shifted by 5 pixels in rows and columns. The increment refers to the increase of the rectangle size until the maximum fit in the image sample. The full set of 57,315 features derived from a 100x100 image sample was over-complete and was much larger than the number of pixels in the image, in this case 10,000 pixels.

Table 3.2: Number of features extracted from a 100x100 window.

Feature type	**Initial size (pixels)**	**Shift (pixels)**	**Increment (pixels)**	**Per prototype**	**Total number of features**
2-rectangles	10	5	+5	15390	30,780
3-rectangles	10	5	+5	8550	17,100
4-rectangles	5	5	+2	2874	2,874
unary	20	5	+10	6561	6,561
Total					**57,315**

3.3.3 Learning Module

Given a feature set and a training set of positive and negative images, we implemented the AdaBoost algorithm as described in Table 3.1. The result of the boosting process returns a strong classifier matrix. Abstractly, the final classifier produced by the algorithm is a set of weighted features that accurately classifies the two sets of labelled images such that the feature with the highest weights are relatively good at classifying the labelled examples. The first ten weak classifiers with higher weights selected by AdaBoost in our training implementation are shown in Figure 3.8.

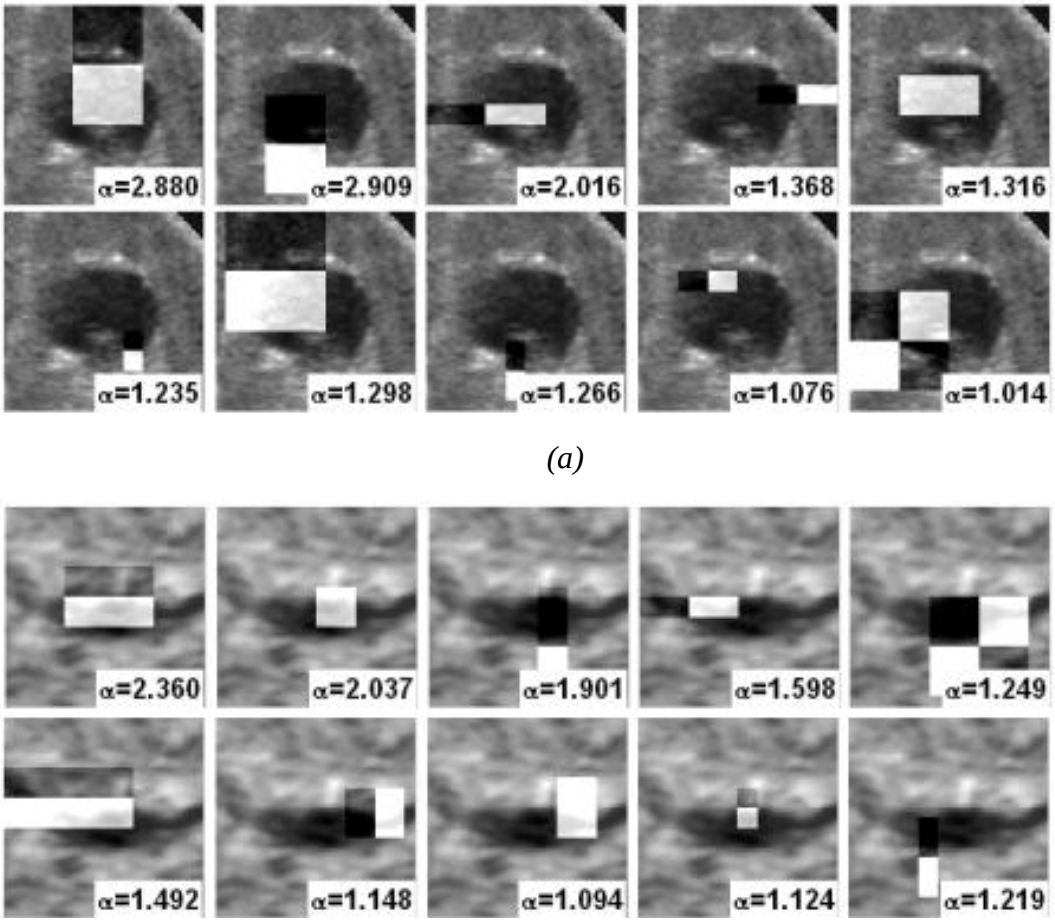
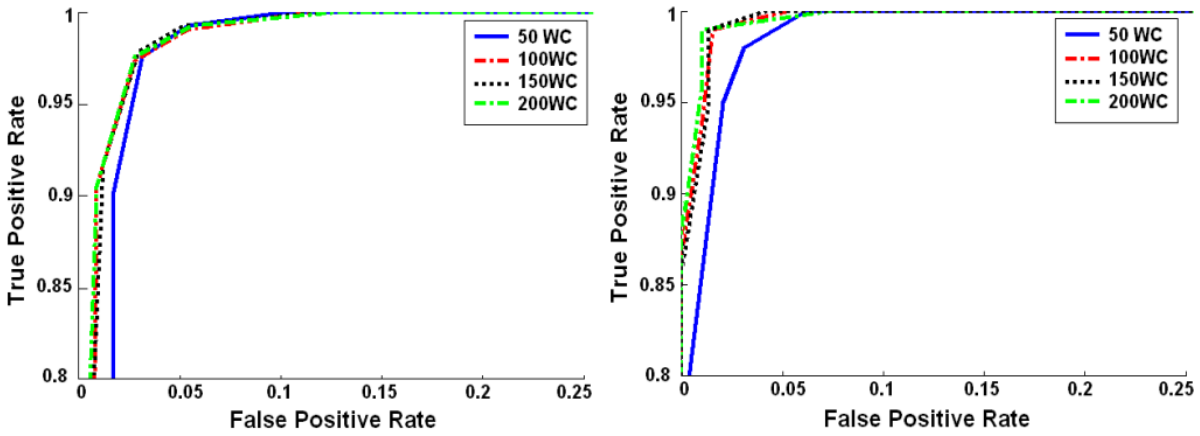


Figure 3.8: The first ten selected Haar features by AdaBoost are shown superimposed on some example images from the training set for (a) the stomach and (b) the umbilical vein.

The only parameter to tune in the algorithm was the number of training rounds that produce T weak classifiers. In our validation experiment, we used different numbers of weak classifiers (50, 100, 150, and 200) to classify 100 regions containing the stomach (or the umbilical vein) and 100 background region samples. The ROC analysis presented in Figure 3.9 indicates that the first 100 weak classifiers captured most of the distinguishing information for our model since there was no significant increase in classification accuracy by adding more weak classifiers after that. However we proposed the use of a larger value of 300 weak classifiers in the experiment for better accuracy in detection.



(a) (b)

Figure 3.9: ROC curve to analyse the effect different number of weak classifiers (WC) (50, 100, 150, and 200) in the classification of (a) the stomach and (b) the umbilical vein.

3.3.4 Detector Module

The detection module used the resultant strong classifier produced in the learning module which consists of an array of T elements, where each element has the following information (refer to Equation (3.3)): the type of feature, the height and width of the feature, the (x, y) location of the feature in the 100 x 100 sub-window, the threshold θ_t, the parity p_t, and the importance weight α_t. The detector only used a single strong classifier, hence is termed a monolithic detector (Viola and Jones, 2004, Bergboer et al., 2006).

The detector used a sliding window method which scanned the input at multiple locations and scales starting at the base size of 100x100 pixels for stomach and a smaller size of 75x75 pixels for the umbilical vein. The image was scanned at 4 scales each a factor of 1.25 larger than the last. The window size was increased in width and in height. The detector was also shifted across location. The shifting process was affected by the size of the detector and the detector shift size was equivalent to 10% of the detector size.

Only one sub-window (with the highest score) is considered as the candidate for the classification. The highest score was evaluated against a threshold value based on the sum of

individual weights of the weak classifiers ($\sum_T \alpha_t$) (refer to the strong hypothesis provided by the AdaBoost algorithm in Equation (3.8)) to give the binary classification result. The result will be positive (indicating the presence of object of interest) if the $\sum_T \alpha_t$ value is higher than the threshold value and negative (indicating absence of object of interest) if the value is lower than the threshold value.

3.3.5 Datasets

2D fetal abdominal ultrasound images for this study were retrieved from the image database of the Oxford Ultrasound Quality Control Unit of the INTERGROWTH-21st project (refer to Section 2.3). All ultrasound examinations were performed using a Philips HD9 ultrasound machine with a 2-5MHz probe by ultrasonographers trained to follow standardized data acquisition procedures.

Table 3.3: Details of the number of positive (+) and negative (-) images in the training, validation and testing datasets.

	Train+	**Train-**	**Valid+**	**Valid-**	**Test+**	**Test-**
Stomach	633	2073	100	100	2283	101
Umbilical Vein	224	851	100	100	2284	100

Images were divided into three sets (refer to Table 3.3): a large set of more than 2000 images was used for training, 100 images for parameter setting through validation and another large set of 2384 images was used for testing the algorithm performance. There is no overlap of images between the sets.

Our testing dataset of 2384 2D fetal abdominal US scans were divided into five clinically relevant gestational age groups (refer to Table 3.4). This was done for checking the performances of the algorithm in different age groups. The images were classified for the

presence and absence of the stomach and umbilical vein after consultation with trained sonographers.

Table 3.4: Distribution of images in the testing datasets for different gestational age groups.

	14^{+0} – 19^{+6} weeks	**20^{+0} – 25^{+6} weeks**	**26^{+0} – 31^{+6} weeks**	**32^{+0} – 37^{+6} weeks**	**38^{+0} – 42^{+6} weeks**
With SB	752	700	390	335	106
Without SB	30	18	11	28	14
With UV	721	710	396	341	116
Without UV	61	8	5	22	4

3.3.6 Validation Methodology

We used specificity, sensitivity, and balanced accuracy as the performance metrics for the detection method. The metrics are defined in the following equations:

$$Specificity = \frac{True\ Negative}{\sum All\ Negative\ Condition} \tag{3.9}$$

$$Sensitivity = \frac{True\ Positive}{\sum All\ Positive\ Condition} \tag{3.10}$$

$$Balanced\ Accuracy = \frac{Specificity + Sensitivity}{Total} \tag{3.11}$$

We empirically set that a stomach or an umbilical vein had been detected correctly if 75% of its area (manually labelled sub-window) is covered by the detector's output box.

False Positive in our experiment is considered for the following situations:

i) In the presence of the object of interest (stomach or umbilical vein) in the image, the algorithm detects area that does not contain the object or the overlap of the detected area is less than 75% of the ground truth area, or

ii) In the absence of the object of interest in the image, the algorithm gives positive detection result for the images.

False Negative is when the algorithm gives a negative detection result (detects nothing in the presence of the object of interest (stomach or umbilical vein) in the image.

We plotted the receiver operating characteristic (ROC) curves for showing the performance of the detector on the testing dataset by adjusting the threshold value (refer to Equation (3.8)). A lower threshold produced a higher detection rate and a higher false positive rate and vice versa. For the problem of confirming the image plane based on the presence of the objects, we needed to be strict on the False Positive at the cost of lower detection rates since we want to ensure that a wrong image plane does not get passed as acceptable. The threshold value was determined from the point on the receiver operating characteristic (ROC) curve nearest to the upper left corner of the "Overall" detection curve. The detections results recorded (using the threshold value) were used to compute the performance metrics in Equations (3.9) - (3.11).

3.4 Results and Discussion

ROC curves in Figure 3.10 demonstrate the performance of the stomach and the umbilical vein detection method in ultrasound images from different gestational groups. Table 3.5 shows the performance metrics recorded in our experiment.

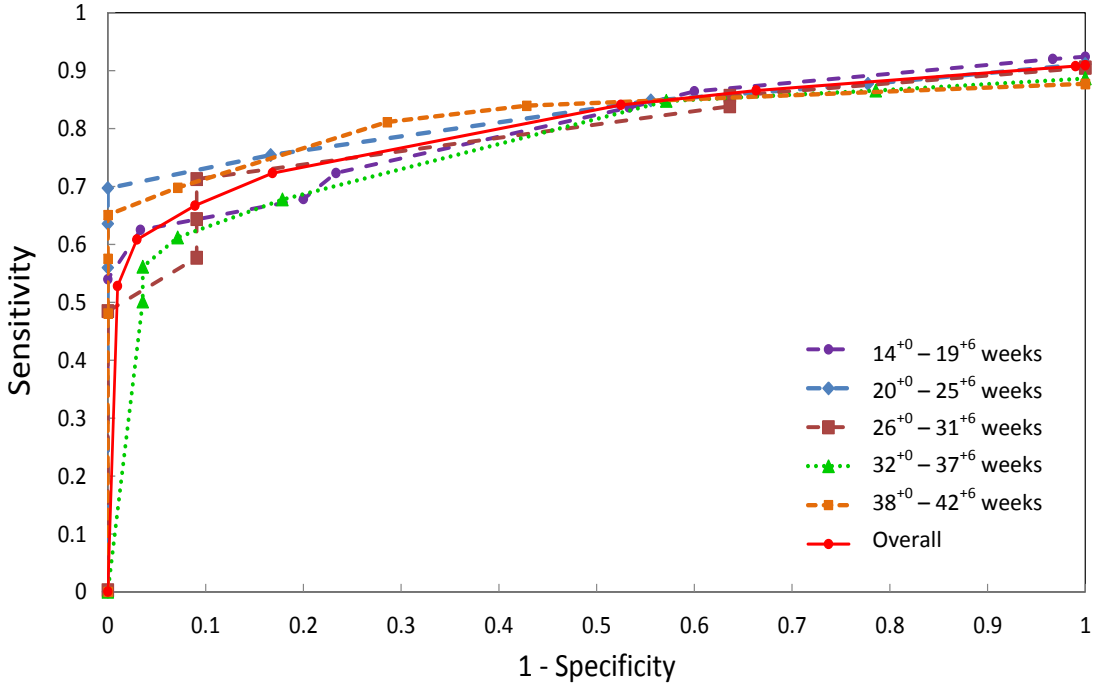
Table 3.5: Overall performance evaluation for the stomach and the umbilical vein detection

Stomach Detection

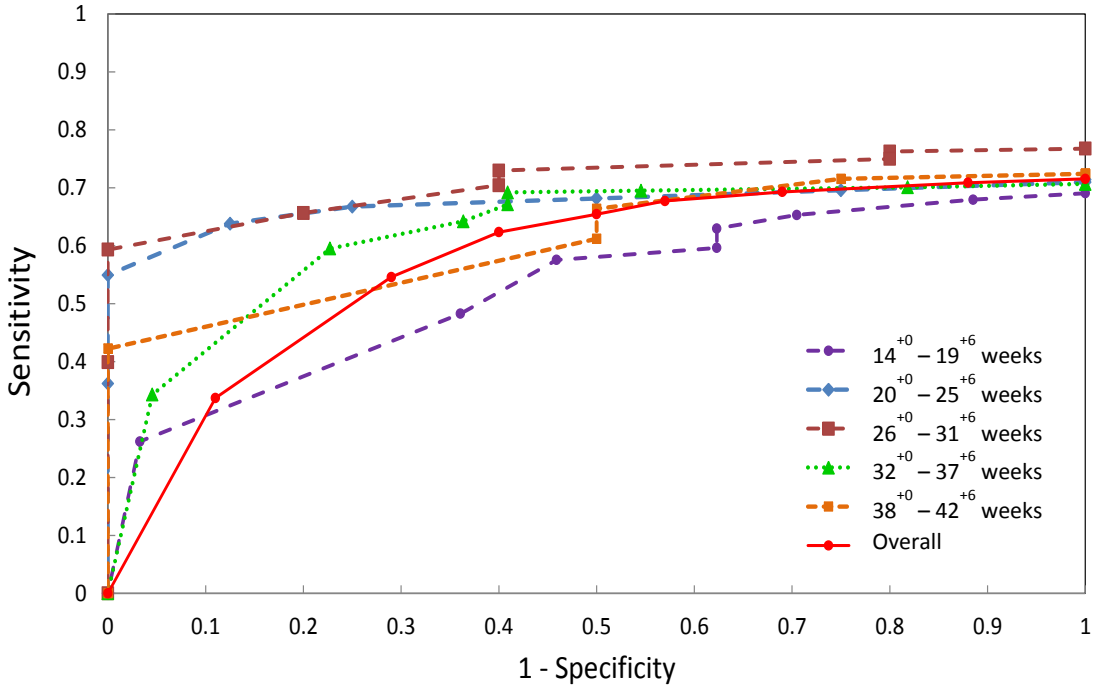
	$14^{+0} – 19^{+6}$ weeks	**$20^{+0} – 25^{+6}$ weeks**	**$26^{+0} – 31^{+6}$ weeks**	**$32^{+0} – 37^{+6}$ weeks**	**$38^{+0} – 42^{+6}$ weeks**	**Overall**
Balanced Accuracy	79.58%	81.79%	74.30%	74.49%	78.77%	78.94%
Sensitivity	0.63	0.64	0.58	0.56	0.58	0.61
Specificity	0.97	1.00	0.91	0.93	1.00	0.96
Area under curve (AUC)	0.80	0.82	0.79	0.77	0.82	0.80

Umbilical Vein Detection

	$14^{+0} – 19^{+6}$ weeks	**$20^{+0} – 25^{+6}$ weeks**	**$26^{+0} – 31^{+6}$ weeks**	**$32^{+0} – 37^{+6}$ weeks**	**$38^{+0} – 42^{+6}$ weeks**	**Overall**
Balanced Accuracy	56.10%	77.46%	79.67%	68.40%	55.60%	62.80%
Sensitivity	0.48	0.55	0.59	0.60	0.61	0.55
Specificity	0.64	1.00	1.00	0.77	0.50	0.71
Area under curve (AUC)	0.53	0.67	0.71	0.62	0.61	0.57



(a)



(b)

Figure 3.10: ROC curves for the detection of (a) the stomach and (b) the umbilical vein in different gestational age (GA) groups. GA is defined in weeks from conception.

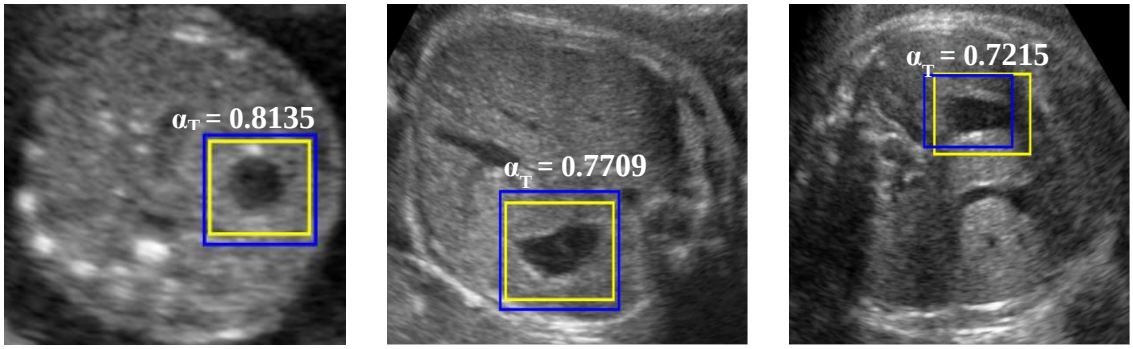
Based on the results, the detection of the stomach was more accurate and more consistent throughout all gestational ages compared to the umbilical vein and this agrees with general clinical understanding of difficulty of scanning. In some cases the algorithm detected other very similar looking objects in the abdominal scan that resembled the object of interest and this was especially prevalent for umbilical vein detection. There were other objects such as blood vessels in the image that have a similar shape and size to the umbilical vein.

The performance of the method in detecting the presence of both objects was best for images in mid-gestational ages (20^{+0} – 25^{+6} weeks and 26^{+0} – 31^{+6} weeks), which were the targeted age range since standard biometric measurement from the abdomen area are typically taken at 18^{+0} – 20^{+6} weeks of gestational age during fetal anomaly screening in the England (National Institute for Health and Clinical Excellence, 2008). Also any detection of abnormalities in terms of growth is preferably done at early gestational age for the development of intrauterine treatment or intervention.

The following factors can be attributed for poor detection of the umbilical vein detections at early and late gestational:

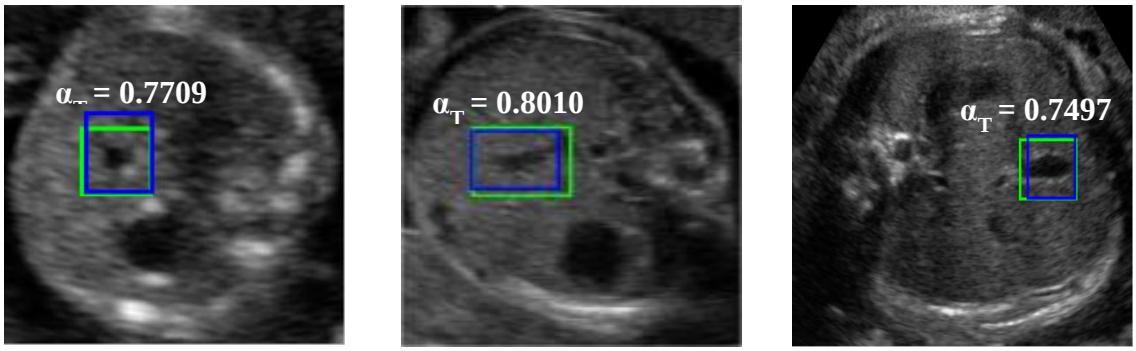
- In the earliest gestational weeks (14^{+0}-19^{+6}) the umbilical vein was difficult to locate as it is a very small structure and has poor contrast compared to background.
- The umbilical vein is a smaller structure compared to the stomach, and is often shadowed by the reflection from the spine and/or ribs at later gestational ages, making it prone to false positive detection or non-detection.

Figure 3.11 to Figure 3.18 present the visual results of the detection method for several sample images from the test datasets. The results were representative of the general trend of the detection in the different gestational age groups. Also presented are the maximum α values achieved using the detection method.



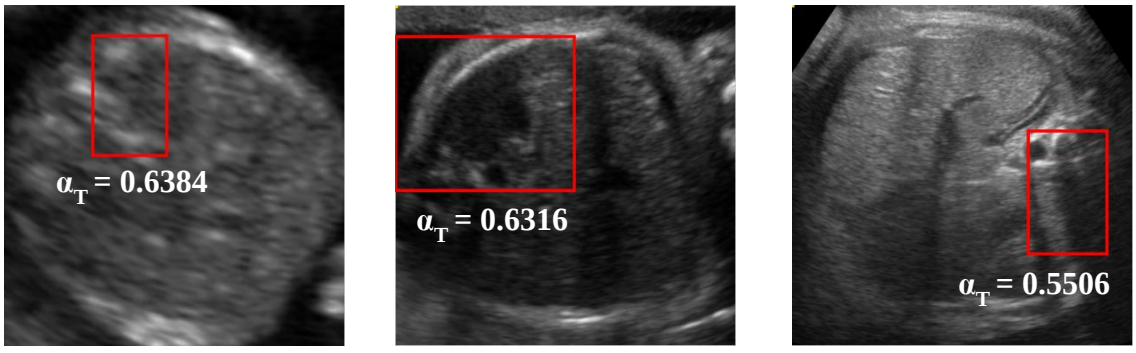
(a) (b) (c)

Figure 3.11: True positive results for stomach detection in different fetal scans at (a) 18 weeks (b) 28 weeks and (c) 38 weeks. Blue box represents the detection by the automated method and yellow box is the ground truth for the stomach.



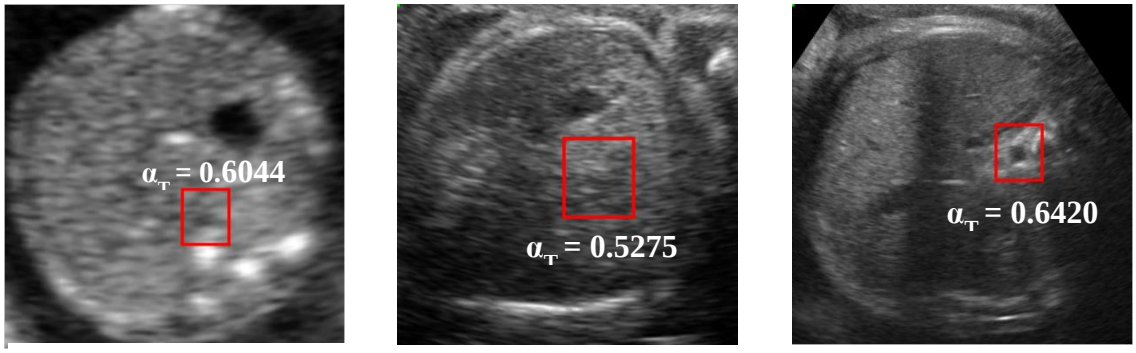
(a) (b) (c)

Figure 3.12: True positive results for umbilical vein detection in different fetal scans at (a) 16 weeks (b) 26 weeks and (c) 39 week. Blue box represents the detection by the automated method and green box is the ground truth for the umbilical vein.



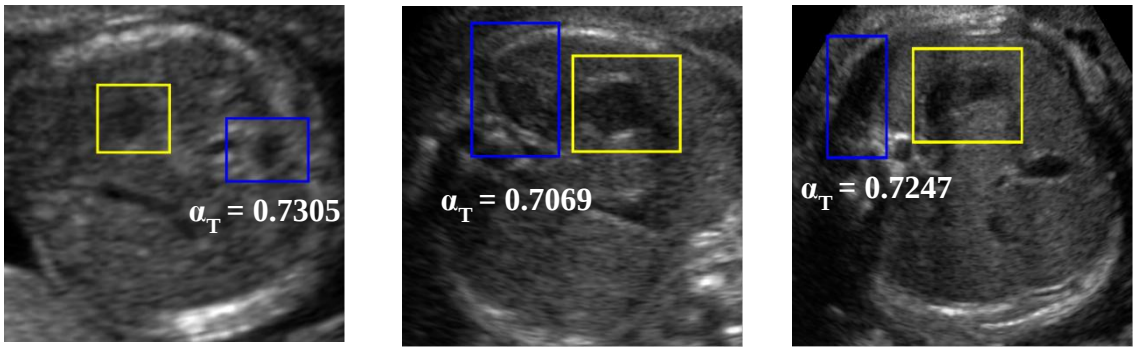
(a) (b) (c)

Figure 3.13: True negative results for stomach detection in different fetal scans at (a) 17 weeks (b) 30 weeks and (c) 38 weeks. Red box represents the sub-window that returned the maximum alpha value by the stomach detector and is less than the threshold value of 0.7.



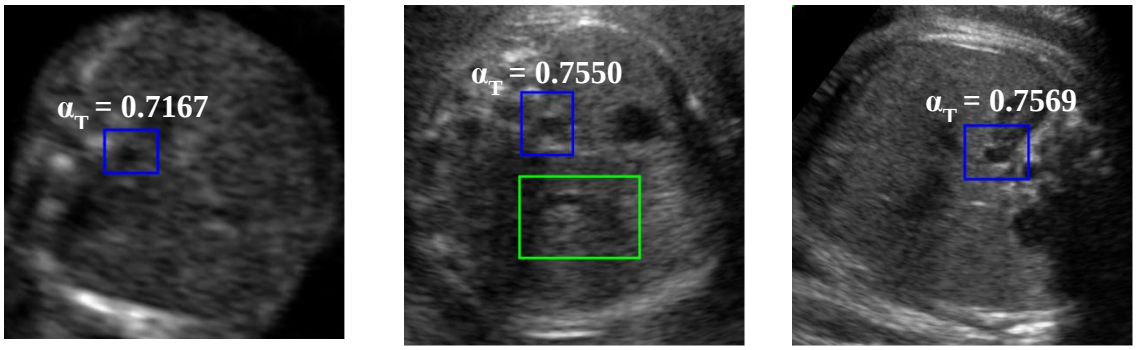
(a) (b) (c)

Figure 3.14: True negative results for umbilical vein detection in different fetal scans at (a) 17 weeks (b) 28 weeks and (c) 38 weeks. Red box represents the sub-window that returned the maximum alpha value by the umbilical vein detector.



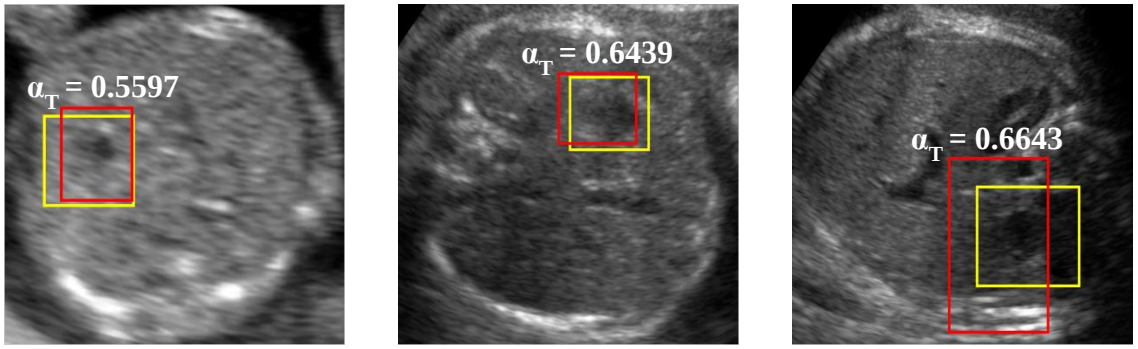
(a) (b) (c)

Figure 3.15: False positive results for stomach detection in different fetal scans at (a) 19 weeks (b) 29 weeks and (c) 39 weeks. Blue box represents the detection by the automated method and yellow box is the ground truth for the stomach.



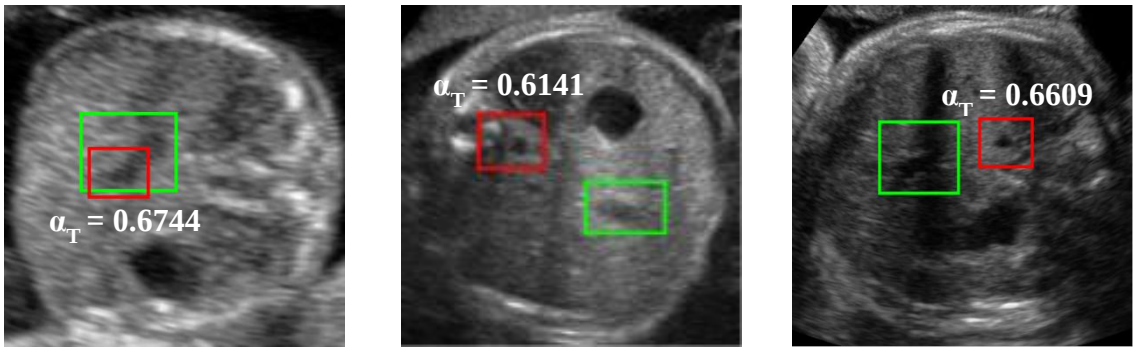
(a) (b) (c)

Figure 3.16: False positive results for umbilical vein detection in different fetal scans at (a) 16 weeks (b) 26 weeks and (c) 38 weeks. Blue box represents the detection by the automated method and green box is the ground truth for the umbilical vein.



(a) (b) (c)

Figure 3.17: False negative results for stomach detection in different fetal scans at (a) 19 weeks (b) 29 weeks and (c) 38 weeks. Red box represents the sub-window that returned the maximum alpha value by the stomach detector and yellow box is the ground truth.



(a) (b) (c)

Figure 3.18: False negative results for umbilical vein detection in different fetal scans at (a) 16 weeks (b) 30 weeks and (c) 40 weeks. Red box represents the sub-window that returned the maximum alpha value by the umbilical vein detector and green box is the ground truth for the umbilical vein.

The qualitative results in Figure 3.15 to Figure 3.18 confirm that shadowing, artefacts and poor contrast are common factors that triggered the false detection. This was especially prevalent for scans in later gestational age where the reflection from the spine often half-occludes the object of interest. The detector acted consistently accurate for images that were of good quality: objects clearly visible without any shadows from the surrounding and with good contrast compared to background (as shown in Figure 3.11 and Figure 3.12).

3.5 Conclusions

This chapter presented a machine learning approach for the detection of the stomach and the umbilical vein in fetal ultrasound images to assist in confirming the correct image plane for biometry measurement. The experimental results demonstrated that the proposed approach can assist in the detection of key anatomical landmarks in fetal ultrasound images. The average execution time was around 10 seconds and this could further be improved for on-line detection. The quality of the scan was an important factor which affects the accuracy of the detection and should be taken into account for any further improvement using the proposed framework. It was observed that the detection of the stomach was more accurate compared to the umbilical vein and this agrees with general clinical understanding of difficulty of scanning.

Chapter 4 Local Phase Feature from Monogenic Signal for Object Detection

In the previous chapter, we proposed a machine learning detection framework that used the AdaBoost algorithm as a learning tool which utilizes the Haar feature sets extracted from the intensity image of a 2D fetal abdominal ultrasound. In this chapter, we apply the same framework, this time supplied with a richer feature set consisting of the previous Haar features and features derived from local phase images. Our main objective is to evaluate the performance of the stomach and the umbilical vein detection algorithm when it is supplied with a richer feature sets that are derived from both intensity and local phase images.

Motivation for using local phase features in the ultrasound images is presented in Section 4.1. Section 4.2 describes the relevant background on local phase computation using the monogenic signal. Section 4.3 describes implementation details such as filter scale selection and types of features extracted from a local phase image along with their combination with intensity features for the classifier training. Experimental results are presented in Section 4.4 and are followed by a discussion in Section 4.5. The chapter ends with concluding remarks in Section 4.6.

4.1 Introduction

The AdaBoost algorithm, as presented in the previous chapter, relies on the combination of the selected feature sets that are combined to make a strong classifier. Its success relies heavily on the discrimination power of the feature sets that it is supplied during the training phase.

Different type of image features have been proposed for training machine learning algorithms. Among the features used in detection of objects in natural images are image intensities, image curvatures, intensity gradient magnitudes, steerable features (Freeman and Adelson, 1991), Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005), and Pyramid Histogram of Oriented Gradients (Bosch et al., 2007). In medical imaging, typical type of features that has been used with the AdaBoost algorithm are Haar features (Morra et al., 2008, Reddy and Bhuyan, 2008), first and second order derivatives of Gaussian filters (Zhou et al., 2007, Pescia et al., 2008), and statistical measures such as mean, standard deviation, median, maximum, minimum, etc (Madabhushi et al., 2005, Wei et al., 2005, Yuanzhong et al., 2006). All these features are extracted from the original intensity image.

To perform reliable detection, it is important that the features extracted from a training image be detectable even under changes in image scale, noise level and illumination. This imposes a great challenge for ultrasound images due to speckle, shadows and low contrast characteristic features. We approach the problem of introducing new feature sets for the stomach and the umbilical vein detection using an AdaBoost learning algorithm by extracting the features from local phase based analysis due to its proven reduced sensitivity to speckle (if scales are carefully chosen) (Rajpoot et al., 2009, Rueda et al., 2011) and intensity invariance.

We are encouraged by the works in the ultrasound domain where the feature information extracted from local phase-based processing of images has proven beneficial for a variety of image analysis tasks. Structural and edge information based on local phase is used to drive the segmentation of the fat adipose tissue in the ultrasound image of the fetal arm in Rueda et al. (2011). A local phase-based method has been employed for endocardial and epicardial boundary detection from 2D+time echocardiographic images (Mulet-Parada and Noble, 2000). The information derived from local phase images was integrated in the

speed term in the level set framework for the segmentation of the left ventricle in ultrasound (Belaid et al., 2011). Similarity measures derived from local phase-based method were used to guide the registration of apical and parasternal view of real-time 3D echocardiographic images (Grau et al., 2007) and MR-US registration (Weiwei et al., 2011). The features corresponding to tissue and bone derived from local phase image were utilized for bone segmentation in ultrasound images by Hacihaliloglu et al. (2009). The results from all these works suggest that local phase might be useful to improve the performance of the algorithm presented in Chapter 3.

In this chapter, we propose employing a feature set extracted from local phase based analysis within the AdaBoost machine learning framework for the stomach and the umbilical vein detection in 2D ultrasound images. Our aim is to evaluate the improvement achieved in the detection of the anatomical features by the introduction of these new features in the training datasets. We employed the monogenic signal which makes use of the isotropic Riesz filter to derive local phase and the feature symmetry (FS) measure (Felsberg and Sommer, 2001).

4.2 Background on Local Phase and Monogenic Signal

Phase information plays a crucial role in preserving important structural features in various types of signal (e.g., speech, images, volumetric data). Oppenheim and Lim (1981) demonstrated the importance of phase information by producing a hybrid image using the Fourier transform domain amplitude and phase spectra from two different images. We replicated their experiment and the results are illustrated in Figure 4.1.

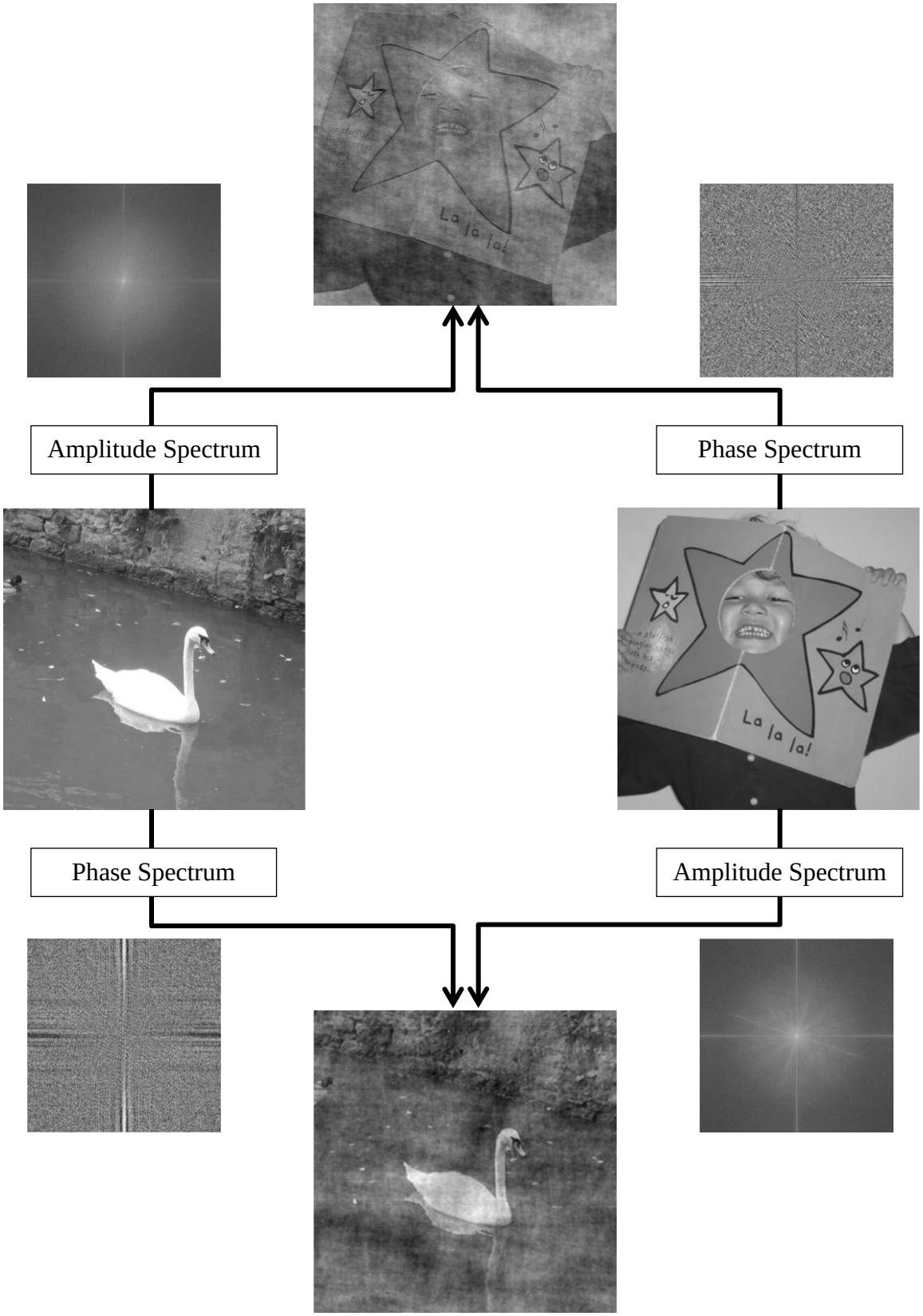


Figure 4.1: Illustration of the importance of phase where Fourier magnitude spectrum and Fourier phase spectrum were taken from separate images. Inverse Fourier transform was then performed to produce a new image.

Examples in Figure 4.1 demonstrate that the features seen in the hybrid image clearly correspond to those in the image from which the phase spectrum is taken. Almost no evidence from the image that supplied the amplitude information can be perceived in the hybrid image. This suggests that the information carried by the phase of the image appears to be more significant than the information carried by its amplitude.

Mathematically, the local phase and local amplitude (i.e. local energy) of a 1D signal are defined using the complex analytic signal $f_A(x)$ (Granlund and Knutsson, 1995). The analytic signal is formed by taking the original signal $f(x)$ as the real part and its Hilbert transform as the imaginary part $f_{\mathcal{H}}(x)$,

$$f_A(x) = f(x) + \imath f_{\mathcal{H}}(x), \tag{4.1}$$

where $\imath = \sqrt{-1}$. In the Fourier domain, the Hilbert transform $f_{\mathcal{H}}(u)$has the same magnitude as $f(u)$ but is rotated by $\pi/2$. The real and the imaginary parts of the analytical signal enable the definition of the local energy (E) and the local phase (φ) in the following way:

$$E(x) = \|f_A(x)\| = \sqrt{Re(f_A(x))^2 + Im(f_A(x))^2} \tag{4.2}$$

$$\varphi(x) = \arg(f_A(x)) = \arctan\left(\frac{Im(f_A(x))}{Re(f_A(x))}\right) \tag{4.3}$$

where $Re[\cdot]$ and $Im[\cdot]$ correspond to the real and complex components of the signal, respectively.

In practice, convolution with band-pass quadrature filters (i.e., filters which are $\pi/2$ phase shift version of each other) are used to transform a real-valued signal $f(x)$ to an analytical signal f_A(x).

$$f_A(x) = f_e(x) * f(x) - i\mathcal{H}\big(f_e(x) * f(x)\big)$$

$$= \big(f_e(x) - i\mathcal{H}(f_e(x)\big) * f(x)$$

$$= \big(f_e(x) - i(f_o(x)\big) * f(x) \tag{4.4}$$

where $*$ denotes the convolution operator, $f_e(x)$ is the band-pass filter and $f_o(x)$ is the Hilbert transform of $f_e(x)$, hence they are in quadrature (Granlund and Knutsson, 1995).

The convolution of the signal with a pair of quadrature filters produces the even-symmetric response, $even(x)$, residing in the real part of the result and the odd-symmetric response, $odd(x)$, residing in the imaginary part,

$$even(x) = f_e(x) * f(x) \tag{4.5}$$

$$odd(x) = f_o(x) * f(x), \tag{4.6}$$

Following the definitions in Equations (4.2) and (4.3), the local energy,$\hat{E}(x)$, and the local phase, $\hat{\varphi}(x)$, are then computed as follows,

$$\hat{E}(x) = \sqrt{[even(x)]^2 + [odd(x)]^2} \tag{4.7}$$

$$\hat{\varphi}(x) = \arctan\left(\frac{odd(x)}{even(x)}\right) \tag{4.8}$$

A common choice of quadrature filters is the log-Gabor filter (Field, 1987), which has a Gaussian transfer function when viewed on the logarithmic frequency scale (as shown in Figure 4.2). The log-Gabor filter has a transfer function of the form,

$$G(\omega) = exp\left(-\frac{log^2(\omega/k)}{2log^2(\sigma_\omega)}\right), \tag{4.9}$$

where k is the centre frequency of the filter (at which the power spectrum is at its maximum), and $0 < \sigma_\omega < 1$ is related to the spread of the frequency spectrum in a logarithmic function.

Log-Gabor filters allow arbitrarily large bandwidth filters to be constructed while still maintaining a zero DC component in the even-symmetric filters. Example of the quadrature pair of even- and odd-symmetric log-Gabor filters is shown in Figure 4.2.

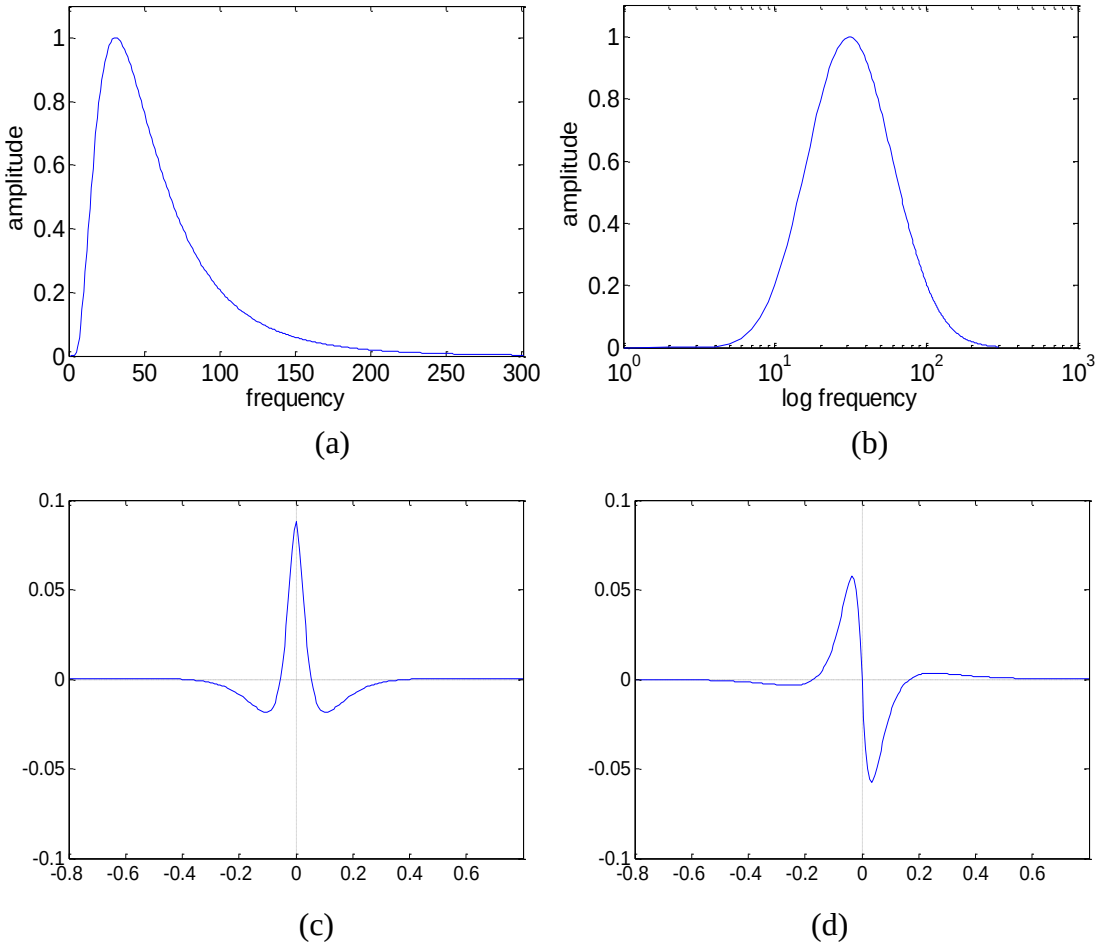


Figure 4.2: Examples of a log-Gabor filter. The transfer function of the filter viewed on both (a) linear and (b) logarithmic frequency scales. The quadrature pair of (c) even and (d) odd symmetric filters in spatial domain having bandwidth of 2 octaves and a center frequency of 1/50.

The even- and odd-symmetric filters are used to capture the even-symmetric features (i.e., ridges or valleys) and odd-symmetric features (i.e., positive or negative step edges) in the signal. The even symmetry of the signal (also known as local symmetry) exhibits a local

phase value of $\pm\pi/2$ and the local phase value of 0 or $\pm\pi$ denotes the odd symmetry of the signal (local asymmetry) (Rajpoot et al., 2009).

So far, the description of the analytical signal has been limited to a signal in one dimension. The extension of the analytical signal to a higher dimension is not straightforward since the Hilbert transform is only defined mathematically for a 1D signal. Typically, the extension of the local analysis techniques to 2D is performed using the construction of a quadrature pair of oriented band-pass filters by applying the 1D analysis over several orientations and combining the results in some way (Kovesi, 1996, Robbins and Owens, 1997, Mulet-Parada and Noble, 2000). This solution comes with the complexity of dealing with an oriented filter bank since it is then necessary to solve for the additional parameters relating to the choice of orientations.

Felsberg and Sommer (2001) have proposed a solution to this known as the monogenic signal. The monogenic signal is an isotropic extension of the analytic signal which preserves the core properties of the 1-D analytic signal that decomposes a signal into information about its structure (local phase) and energy (local amplitude). The analytic signal is generated using an isotropic vector valued odd filter known as the Riesz filter, which is a generalization of the Hilbert transform for higher dimensional signals. For 1D signal the Riesz transform is equivalent to the Hilbert transform. The spatial representations of these filters are as follows,

$$h_1(x,y) = \frac{-x}{2\pi(x^2+y^2)^{3/2}} \tag{4.10}$$

$$h_2(x,y) = \frac{-y}{2\pi(x^2+y^2)^{3/2}} \tag{4.11}$$

The convolution kernels of the Riesz filters in the frequency domain are,

$$H_1(u,v) = \frac{u}{\sqrt{u^2+v^2}} \tag{4.12}$$

$$H_2(u,v) = \frac{v}{\sqrt{u^2+v^2}} \tag{4.13}$$

In practice, the image $I(x,y)$ is first convolved with an even isotropic band-pass filter $b(x,y)$ that produces the even component of the monogenic signal,

$$even(x,y) = I_b(x,y) = b(x,y) * I(x,y). \tag{4.14}$$

The bandpassed image $I_b(x,y)$ is then filtered with the Riesz filter to produce the odd components,

$$\begin{aligned} odd_1(x,y) &= h_1(x,y) * I_b(x,y), \\ &= h_1(x,y) * b(x,y) * I(x,y). \end{aligned} \tag{4.15}$$

$$\begin{aligned} odd_2(x,y) &= h_2(x,y) * I_b(x,y), \\ &= h_2(x,y) * b(x,y) * I(x,y). \end{aligned} \tag{4.16}$$

$$odd(x,y) = \sqrt{\left(odd_1(x,y)\right)^2 + \left(odd_2(x,y)\right)^2}, \tag{4.17}$$

where $h_1(x,y)$ and $h_2(x,y)$ are the convolution kernels defined in Equations (4.10) and (4.11). The monogenic signal $I_M(x,y)$ of $I(x,y)$ is often expressed as,

$$\begin{aligned} I_M(x,y) &= [I_b(x,y), h_1(x,y) * I_b(x,y), h_2(x,y) * I_b(x,y)], \\ &= [even(x,y), odd_1(x,y), odd_2(x,y)]. \end{aligned} \tag{4.18}$$

From $I_M(x,y)$, the local energy $E(x,y)$, the local orientation $\theta(x,y)$, and the local phase $\varphi(x,y)$ of the image $I(x,y)$ are acquired through the following equations:

$$E(x,y) = \sqrt{even(x,y)^2 + odd(x,y)^2} \tag{4.19}$$

$$\theta(x,y) = arctan\left(\frac{odd_2(x,y)}{odd_1(x,y)}\right) \tag{4.20}$$

$$\varphi(x,y) = arctan\left(\frac{even(x,y)}{odd(x,y)}\right) \tag{4.21}$$

We are primarily interested in the extraction of features for detecting the stomach and the umbilical vein using local phase based analysis to augment intensity information in our machine learning framework. Local phase is invariant to changes in brightness and contrast within the images, which makes it particularly fit for ultrasound images, in which beam attenuation is present and echo intensity depends on the angle of incidence of the ultrasound beam.

Image local phase information had been previously used for processing ultrasound images as described in Section 4.1. However, to the best of our knowledge, this is the first attempt at evaluating the significance of using the local phase information in a machine learning framework for the detection of anatomical features in ultrasound images.

4.3 Experiments

Chapter 3 introduced the learning algorithm framework for the detection of the stomach and the umbilical vein in fetal abdominal ultrasound image. The features used in the previous chapter were Haar features extracted from the intensity image. In this experiment, we propose to introduce the features extracted from a local phase based processed image together with the Haar features into the machine learning framework and to evaluate the effectiveness in the detection of the stomach and the umbilical vein.

A key choice involved in computing the local phase based features from the monogenic signal was the selection of the quadrature band-pass filter. A comparison of common type of quadrature filters used in feature detection is presented in Boukerroui et al. (2004). We chose to experiment with log-Gabor filter which a well-known types of band-pass filters in computer vision applications. Log-Gabor filters allow arbitrarily large bandwidth zero DC filters to be constructed and also simplicity in implementation.

4.3.1 Features

In this experiment, we maintained the use of the same training set of positive and negative samples used in the previous chapter (described in Table 3.3).

We extracted unary features to best represent the local phase information from a local phase image. Unary features refer to the total sum of all the values in the specified rectangle. The details of the unary features extraction are shown in Table 4.1.

Table 4.1: Number of unary features extracted from a 100x100 window.

Feature type	**Initial size (pixels)**	**Shift (pixels)**	**Increment (pixels)**	**Count**
Unary	20	3	+5	**29,584**

Initial size refers to the base width and length of the rectangles. The rectangles were shifted by 3 pixels in rows and columns. The increment refers to the increase of the rectangle size until the maximum fit in the image sample.

4.3.2 Scale Selection

Filter scale selection is important for feature extraction. A very fine scale may detect non-structural response as potential features. By contrast, a very coarse scale may not be able

to capture some of the true feature points. We used two different ways of producing a local phase image:

1) Single-scale local phase image (SSLP) where the local phase image was derived using one scale of the filter.

2) Multi-scale local phase image (MSLP) where the local phase image was derived from the integration of the local phase images produced at different scales.

Local phase image was produced from original ultrasound image (832 x 569 image size as in Figure 3.7a but with the surrounding black frame containing device headers and patient information cropped out).

The MSLP image $\varphi_{MS}(x, y)$ was formed by averaging the local phase computed at N different scales:

$$\varphi_{MS}(x, y) = \frac{1}{N} \sum_{s} \varphi_s(x, y)$$

For the scale determination in producing SSLP image, we experimented with unary features extracted from several different scales as listed in Table 4.2. The value of σ_ω of the filter used in all the local phase calculations was set to 0.5. Then, all the features were combined and boosting was performed using AdaBoost to select the best scale for implementation.

Table 4.2: Filter scales for the stomach and the umbilical vein.

Stomach (pixels)	**Umbilical Vein (pixels)**
50, 100, 150, 200, 250, 300	15, 30, 50, 100, 150, 200, 250, 300

The first ten features selected through the boosting process are shown in Table 4.3 and the examples of the images produced at different scales are shown in Figure 4.3. Based on the list of features in Table 4.3 we decided to use the filter scale of 100 pixels and 30 pixels for

the single-scale local phase (SSLP) detection of the stomach and the umbilical vein, respectively. This is because these two scales (100 pixels for stomach and 30 pixels for umbilical vein) were selected as the first feature and also populate half of the list.

Table 4.3: The scale of filter and the weight of the first ten features selected by AdaBoost for the parameter determination of single-scale local phase implementation

Stomach		**Umbilical Vein**	
Scale of filter (pixels)	**Weight (α)**	**Scale of filter (pixels)**	**Weight (α)**
100	2.489	30	2.521
50	1.021	250	1.505
100	0.786	100	1.173
150	0.769	100	1.280
200	0.668	30	1.355
50	0.517	150	1.171
100	0.460	30	0.955
100	0.449	250	0.848
300	0.502	30	0.818
100	0.470	30	0.994

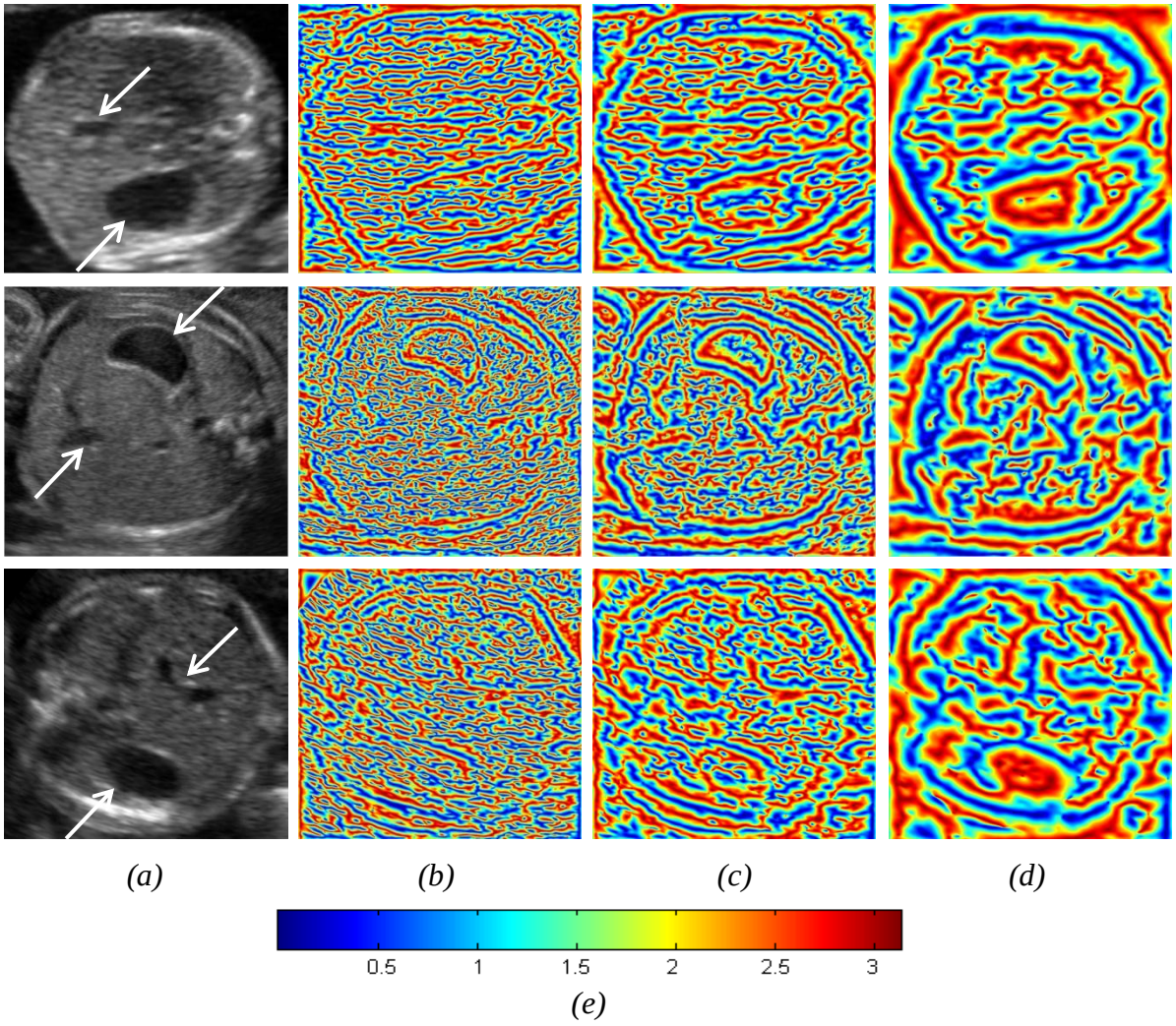


Figure 4.3: (a) Original abdominal ultrasound images and its corresponding local phase images for filter scale of (b) 30 (c) 50 and (d) 100 pixels. The arrows indicate the stomach and the umbilical vein in the images. (e) Colour bar for the images.

For the multi-scale local phase image, three filter scales were chosen from a fine, medium and coarse scale range. These scales were empirically found to give good exclusion of the stomach and the umbilical vein from the surrounding in the majority of images (refer to Figure 4.3). We found that acquiring a local phase image through a multi-scale method was less sensitive to noise while enabling the detection of the key feature points.

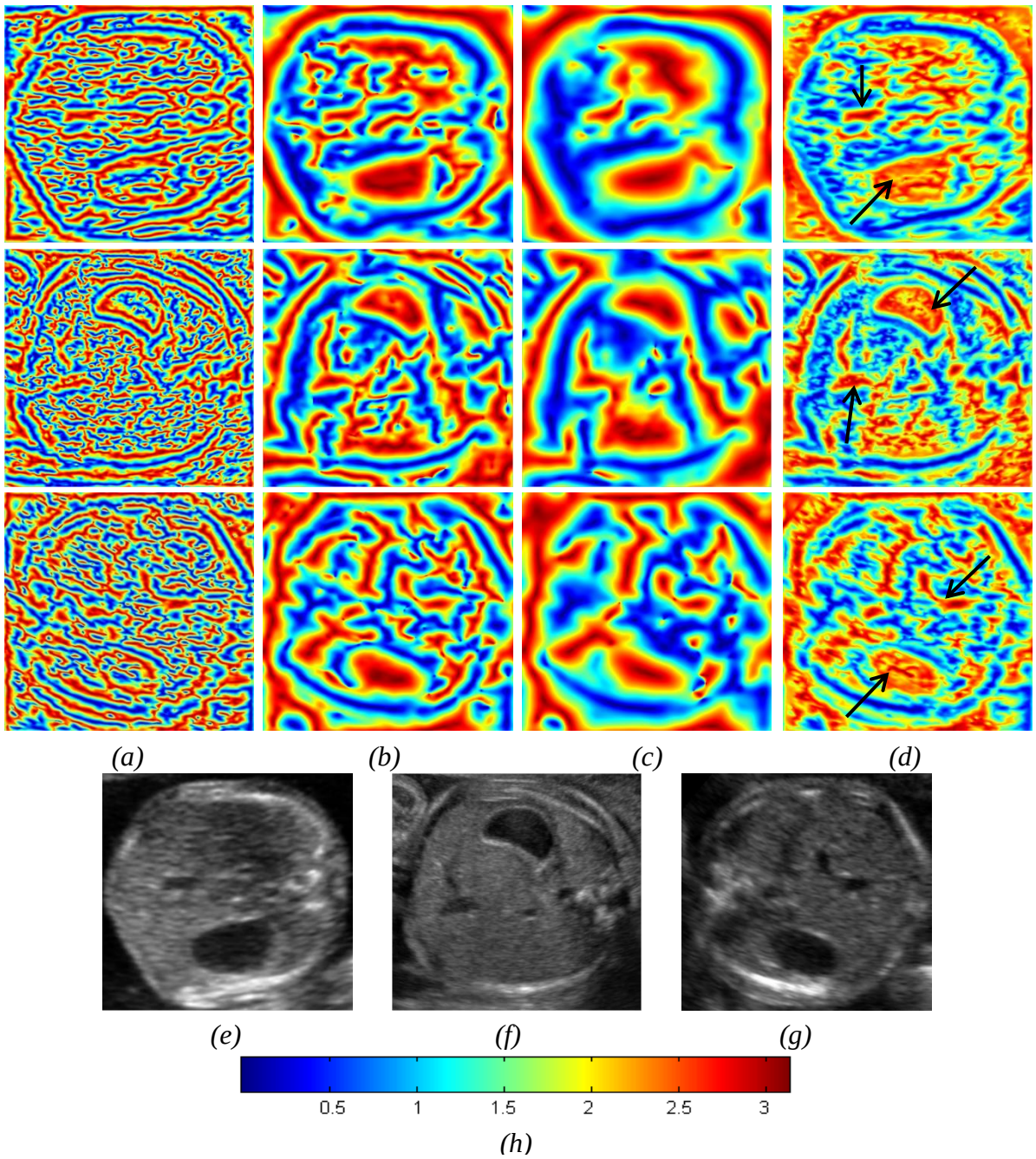
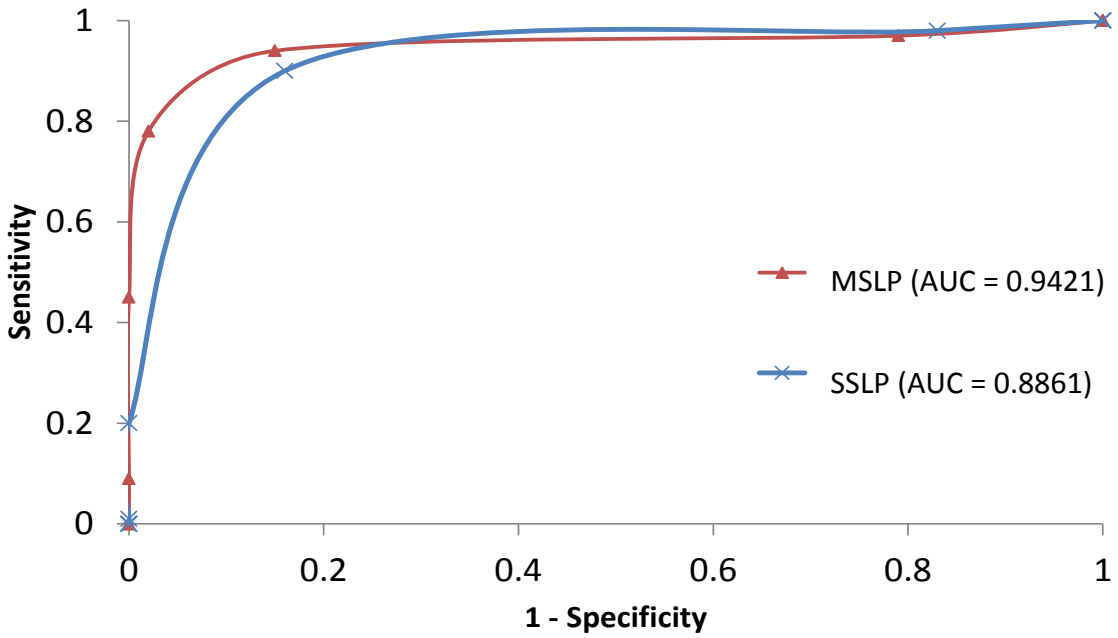


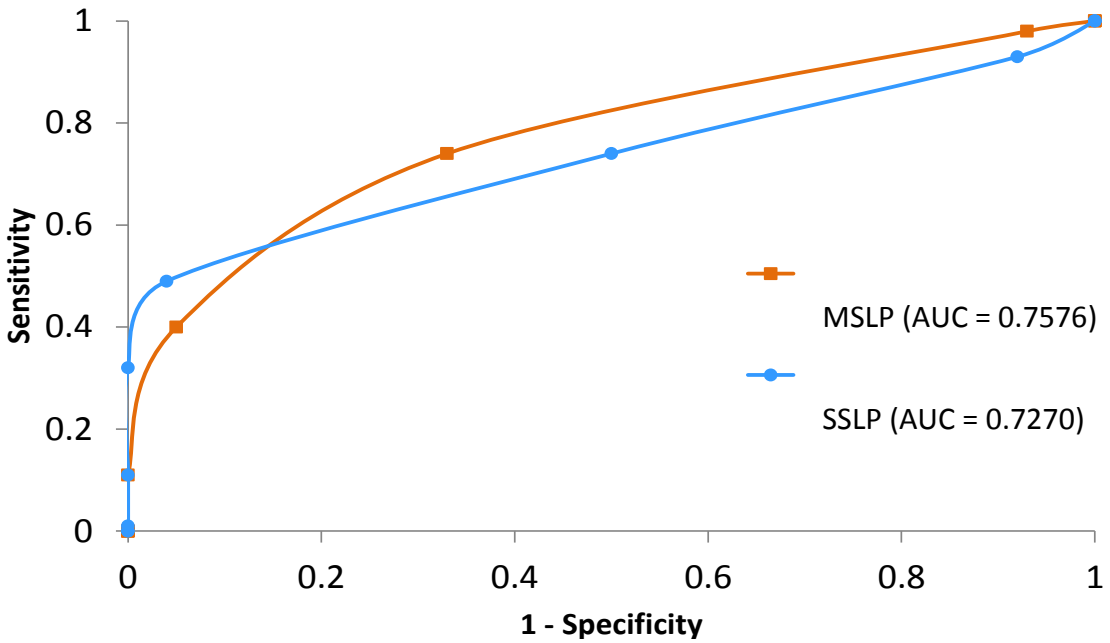
Figure 4.4: Example of local phase images produced with the filter scale of (a) 50 (b) 150 and (c) 250 pixels which are integrated to produce (d) the multi scale images, (e) - (g) are the original intensity images for each row of local phase images(h) Colour bar for the images. The arrows indicate the stomach and the umbilical vein in the images. The intensity images shown here are the same as in Figure 4.3.

We used the validation set (identical to Section 0) to check the performance of the classifiers trained with local phase images derived using the single-scale and multi-scale filters. As demonstrated in Figure 4.5, the classification of both the stomach and the umbilical

vein in the validation set were superior when the features derived from the multi-scale local phase images were used. Therefore, we proceed with the use of the multi-scale local phase features for the detection experiment.



(a)



(b)

Figure 4.5: ROC analysis for the classification result of (a) the stomach and (b) the umbilical vein in the validation set using local phase features from multi scale (MSLP) and single scale (SSLP) filters. Area under the curve(AUC) for each plot are also given.

4.3.3 Classifier Training

We trained the classifier using the AdaBoost learning algorithm supplied with the feature sets extracted from the positive and the negative images of the local phase images, and also on the combination with the features extracted from the intensity (as used in the previous chapter Section 3.2.1). All the feature sets (and their combinations) with their notations used for the training in this experiment are listed below:

1) Multi-scale local phase features (MSLP)

2) Intensity Haar features and multi-scale local phase features (Intensity + MSLP)

Similar to Chapter 3, we performed 300 rounds of boosting on these feature sets to select the weak classifiers and assign its weights of importance in the classification (α).

The resulting first ten weak classifiers chosen through the boosting process for the stomach and the umbilical vein are illustrated in Figure 4.6 and Figure 4.7, respectively. Note that in both the stomach and the umbilical vein trained models, the local phase features were selected as the first weak classifier by the learning algorithm and assigned with a large classifier weight (α) which indicates relatively high discriminating power of the feature. The selected local phase features were mainly focused on the homogenous dark region contained inside the objects while the selected Haar features were more focused on the boundaries of the objects as shown in Figure 4.6 and Figure 4.7.

Table 4.4 enables a more detailed analysis of the weak classifiers selected from the pool of local phase features in order to understand their influence on the classification. It can be seen that the umbilical vein model has a higher number of LP features (60 weak classifiers from the overall total of 300) with the weight ratio equivalent to 20% of its overall weight. LP features in stomach detection model had a lesser impact with 44 features selected and its weight accounted for 15.16% of the total classifier weight.

Table 4.4: Details on the selected local phase (LP) features in the stomach and the umbilical vein trained classifier.

	Stomach	Umbilical Vein
Total LP features	44	60
Ratio of LP weights	0.1516	0.2009

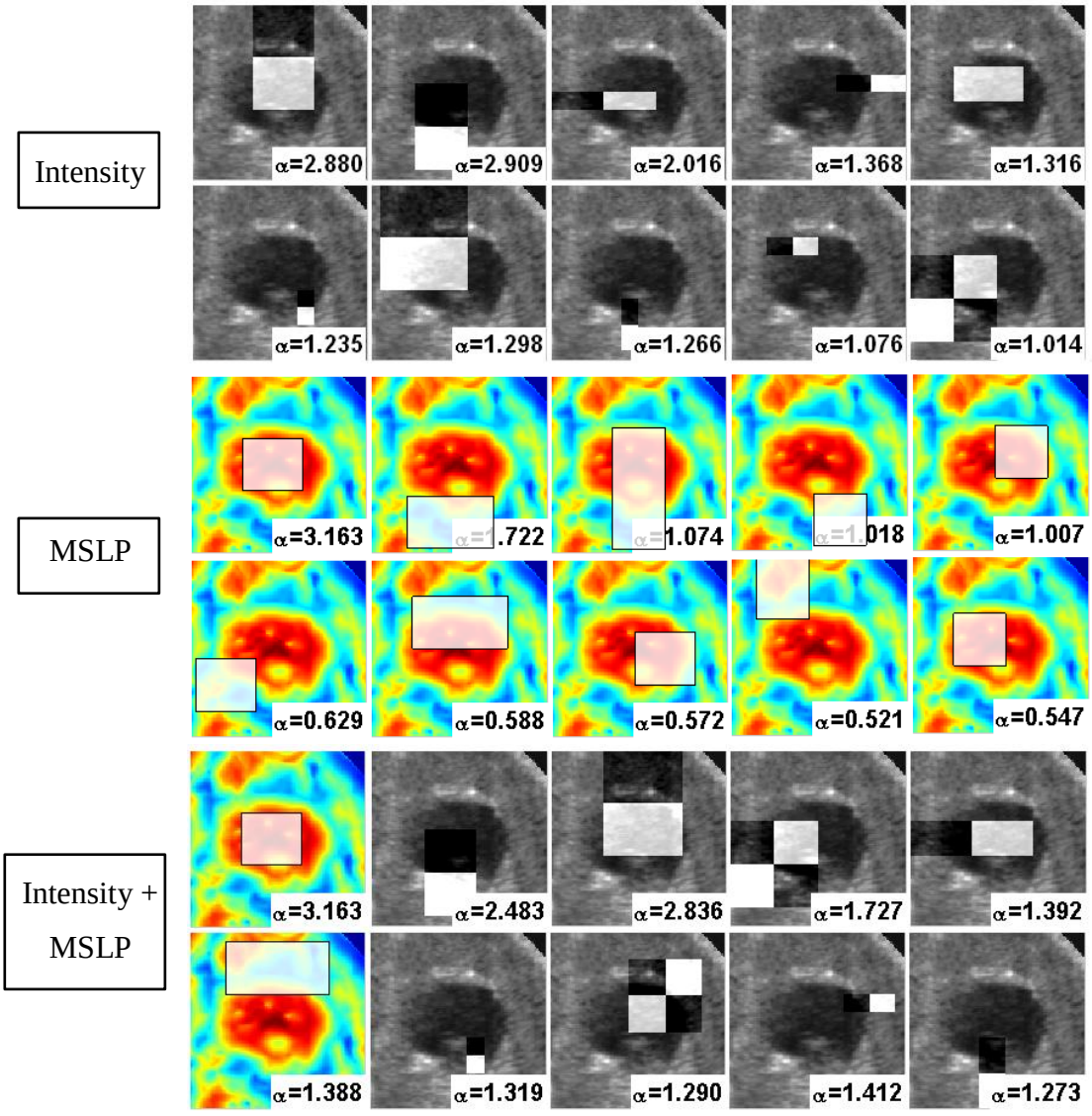


Figure 4.6: The first ten features together with its weight (α) designated by AdaBoost algorithm for the stomach detection superimposed on the sample image. The greyscale and the coloured images indicate that the features selected were from the intensity feature set and local phase feature set, respectively. Different rows represent the different feature set used in the training process.

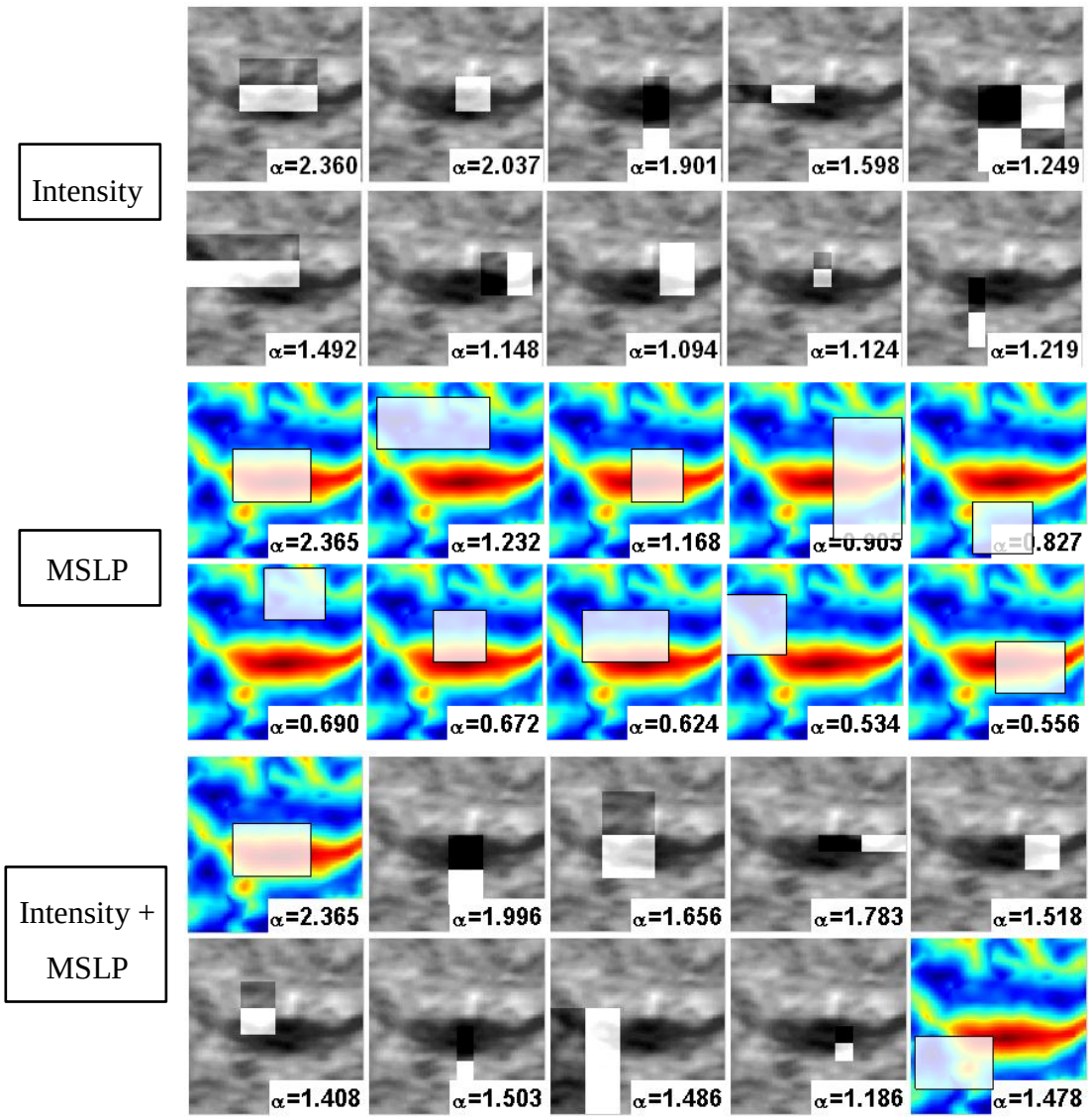


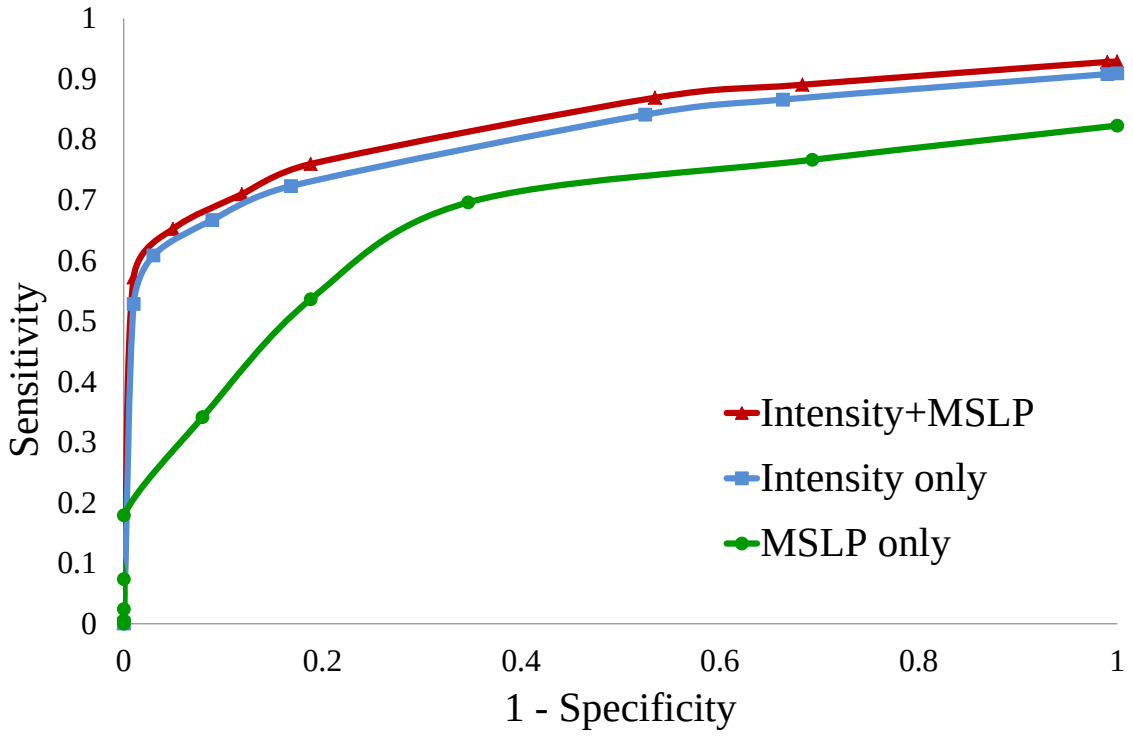
Figure 4.7: The first ten features together with its weight (α) designated by AdaBoost algorithm for the umbilical vein detection superimposed on the sample image. The greyscale and the coloured images indicate that the features selected were from the intensity feature set and local phase feature set, respectively. Different rows represent the different feature set used in the training process.

4.3.4 Validation Measures

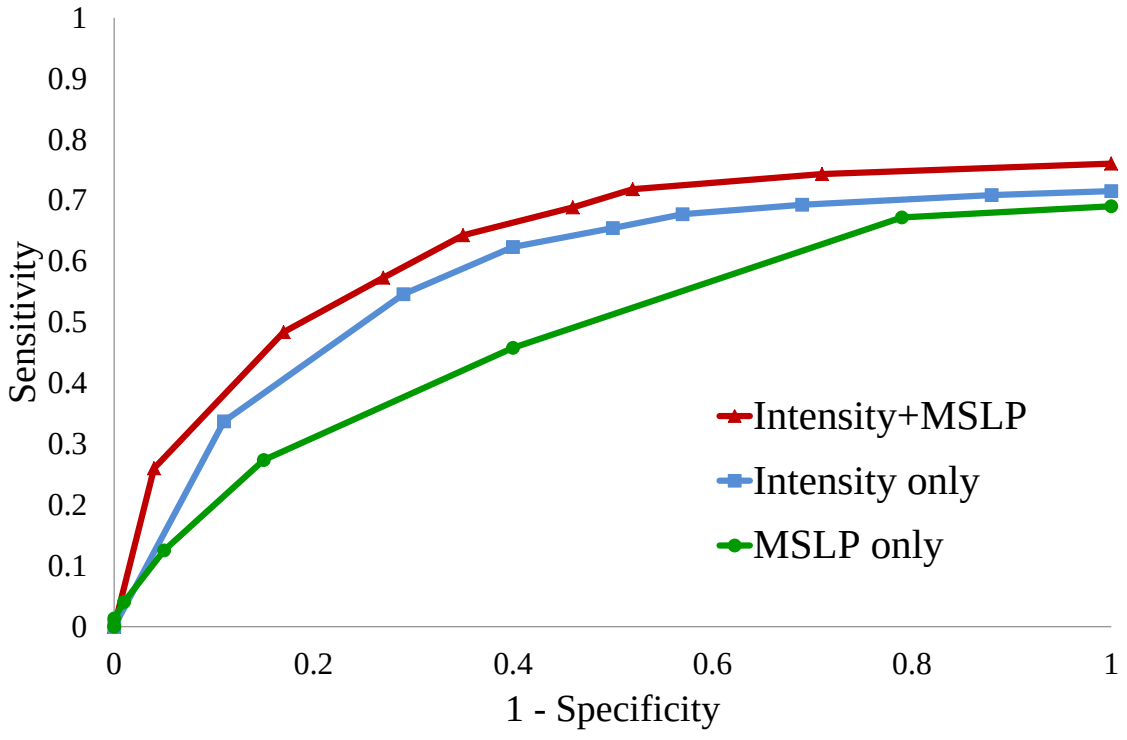
We used the same validation measures as in the previous chapter (see Section 3.3.6) for the performance evaluation of the detection task. We also included Area under Curve (AUC) as an additional measure for the evaluation of the usage of different feature sets in detection. This measure is commonly employed for comparing different machine learning techniques.

4.4 Results

The methods were tested on the same datasets as used in the previous chapter (described in Section 0). The ROC curves produced for the detection of the stomach and the umbilical vein are shown in Figure 4.8. The ROC curves comparing both the "*Intensity+MSLP*" and "*Intensity only*" methods in different gestational age groups are shown in Figure 4.9. Next, using the threshold point determined from the ROC plot, we computed the performance metrics for all the methods which were summarized in Table 4.5. The qualitative results that highlight the comparison between each method are shown and discussed in Section 4.5.

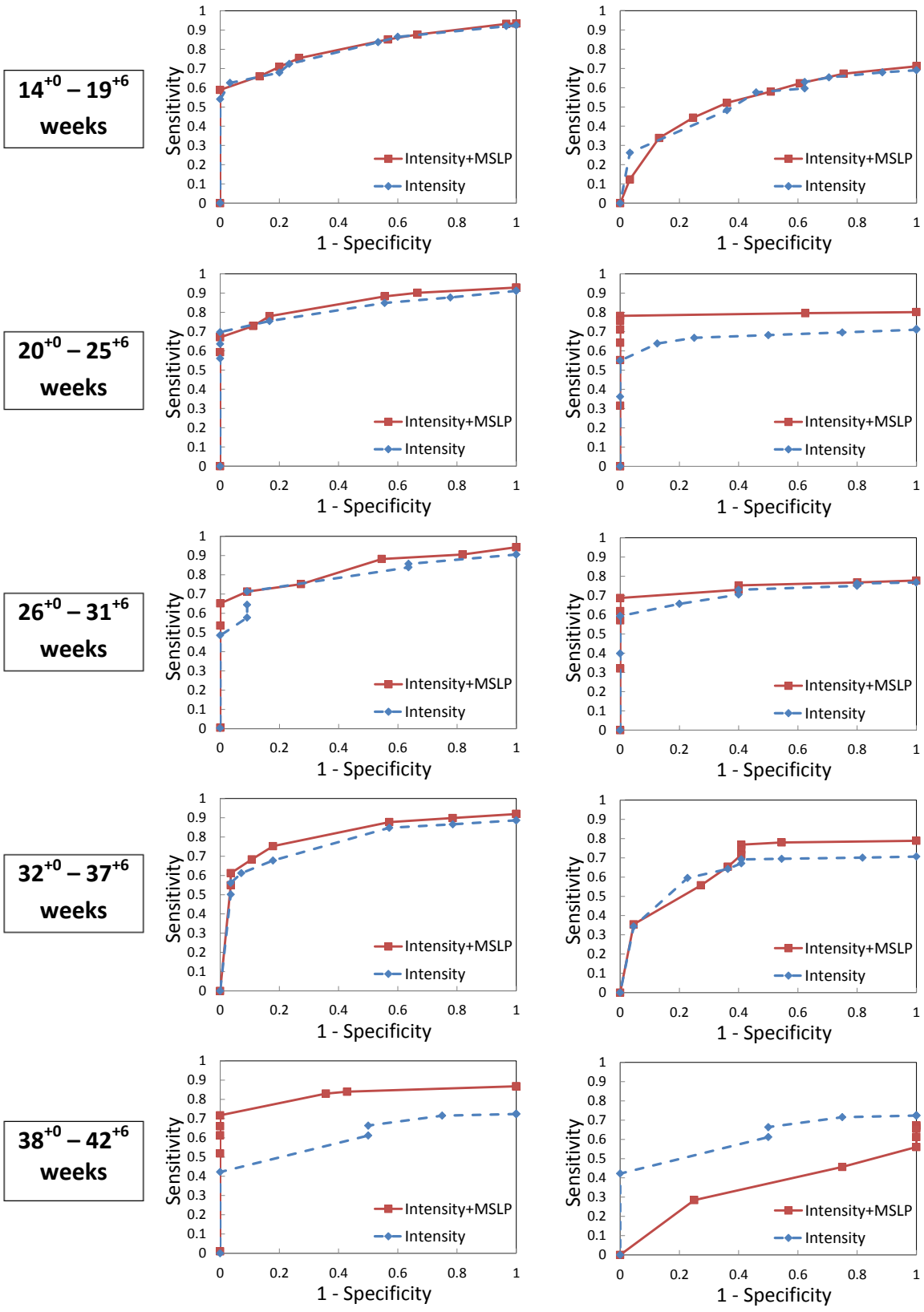


(a)



(b)

Figure 4.8: ROC curves for the detection of (a) the stomach and (b) the umbilical vein using three different types of feature sets.



(a) Stomach detection (b) Umbilical vein detection

Figure 4.9: Comparison of ROC curves between "Intensity+MSLP" and "Intensity only" methods in different gestational age groups.

Table 4.5: Performance evaluation for the detection of the stomach and the umbilical vein between "Intensity+MSLP" and "Intensity only" methods in different gestational age groups.

Stomach Detection

	14^{+0} – 19^{+6} weeks		**20^{+0} – 25^{+6} weeks**		**26^{+0} – 31^{+6} weeks**		**32^{+0} – 37^{+6} weeks**		**38^{+0} – 42^{+6} weeks**		**Overall**	
	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP
Balanced Accuracy (%)	79.58	76.31	81.79	83.50	74.30	82.56	74.49	78.81	78.77	80.66	78.94	80.14
Sensitivity	0.63	0.66	0.64	0.67	0.58	0.65	0.56	0.61	0.58	0.61	0.61	0.65
Specificity	0.97	0.87	1.00	1.00	0.91	1.00	0.93	0.96	1.00	1.00	0.96	0.95
Area under curve (AUC)	0.80	0.81	0.82	0.85	0.79	0.83	0.77	0.81	0.82	0.82	0.80	0.83

Umbilical Vein Detection

	14^{+0} – 19^{+6} weeks		**20^{+0} – 25^{+6} weeks**		**26^{+0} – 31^{+6} weeks**		**32^{+0} – 37^{+6} weeks**		**38^{+0} – 42^{+6} weeks**		**Overall**	
	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP	Int	Int+MSLP
Balanced Accuracy (%)	56.10	59.90	77.46	82.11	79.67	80.93	68.40	64.52	55.60	28.02	62.80	65.16
Sensitivity	0.48	0.44	0.55	0.64	0.59	0.62	0.60	0.65	0.61	0.56	0.55	0.57
Specificity	0.64	0.75	1.00	1.00	1.00	1.00	0.77	0.64	0.50	0.00	0.71	0.73
Area under curve (AUC)	0.53	0.53	0.67	0.79	0.71	0.74	0.62	0.66	0.61	0.34	0.57	0.63

4.5 Discussion

In comparison to the intensity-based features method, the approach with the combined features (*Intensity* + *MSLP* feature sets) achieved an increase of 1.20% and 2.36% in the accuracy of the stomach and the umbilical vein detection, respectively. There was an increase of 100 true positive (TP) detections of stomach but with a decrease of 2 true negative (TN) cases. This reflected in the increase of the sensitivity from 0.61 to 0.65 and the decrease of the specificity from 0.96 to 0.95. For umbilical vein detection, there was an increase of 62 TP and 2 TN detections, giving rise to an increase in both the sensitivity and the specificity by 0.02 to 0.57 and 0.73, respectively.

The improvements in the stomach and the umbilical vein detection that were achieved using the "*Intensity* + *MSLP"* method over the *"Intensity"* method are shown in Figure 4.10 and Figure 4.11, respectively. It is observed from the experimental results that improvements in the detection using the "*Intensity* + *MSLP"* method were generally achieved in images which had one or more of the following characteristics:

a) Non-uniform intensity inside the stomach or the umbilical vein region (i.e. first image in 38^{+0} – 42^{+6} weeks for stomach and first image in 20^{+0} – 25^{+6} weeks for umbilical vein - images are denoted with purple borders).

b) The stomach or the umbilical vein did not have a complete separation from the background for instance edges were blurred and gradually blended with the surrounding. (i.e. first image in 14^{+0} – 19^{+6} weeks for stomach and third image in 26^{+0} – 31^{+6} weeks for umbilical vein - images are denoted with green borders)

c) Poor overall contrast. (i.e. second image in 32^{+0} – 37^{+6} weeks for stomach and third image in 14^{+0} – 19^{+6} weeks for umbilical vein - images are denoted with orange borders)

Generally, the classifier scores by the method with the combined features were found to be higher than the *"Intensity"* method for umbilical vein detection (refer to Figure 4.11). Our results suggest that this is due to the higher number of local phase features chosen as weak classifiers in the "*Intensity + MSLP"* method.

Examples of incorrect detection were presented in Figure 4.12.

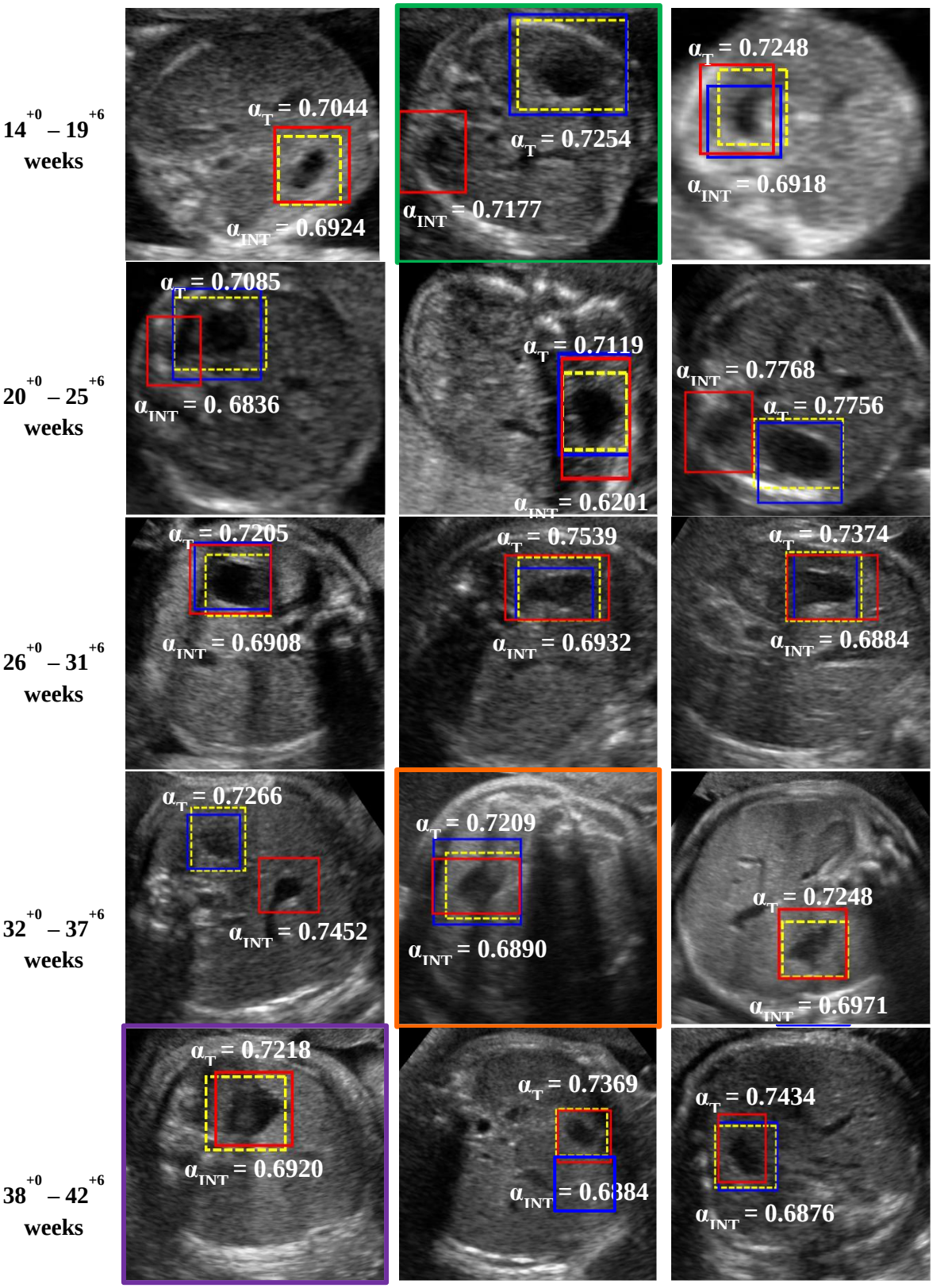


Figure 4.10: True positive results for stomach detection by the "Intensity+MSLP" method (blue box with α_T) where it corrected the false detection result achieved by using "Intensity" features (red box with α_{INT}). Yellow box represents the ground truth. (In some images, α_T appears without the blue box which indicates identical overlaps with red box)

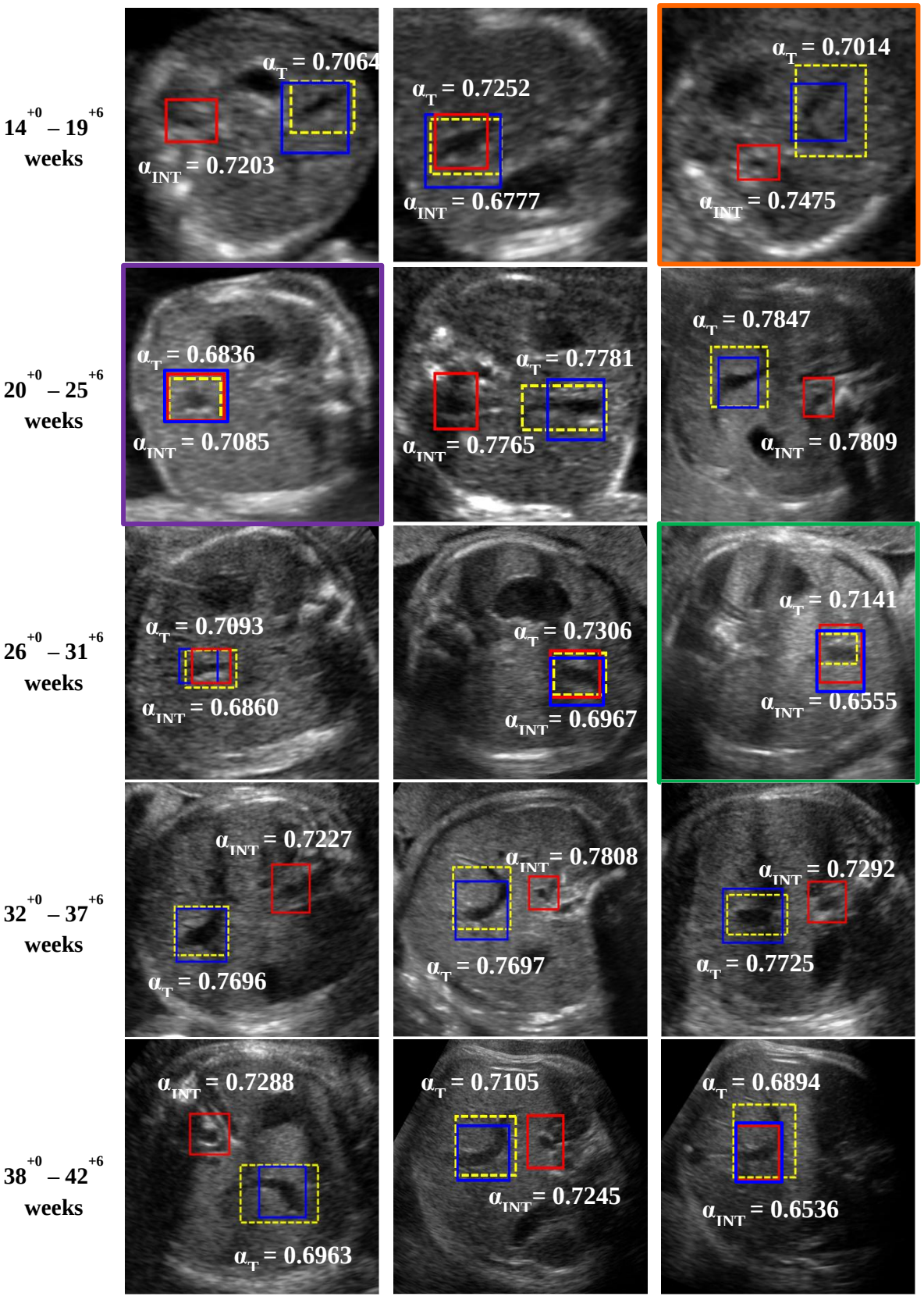
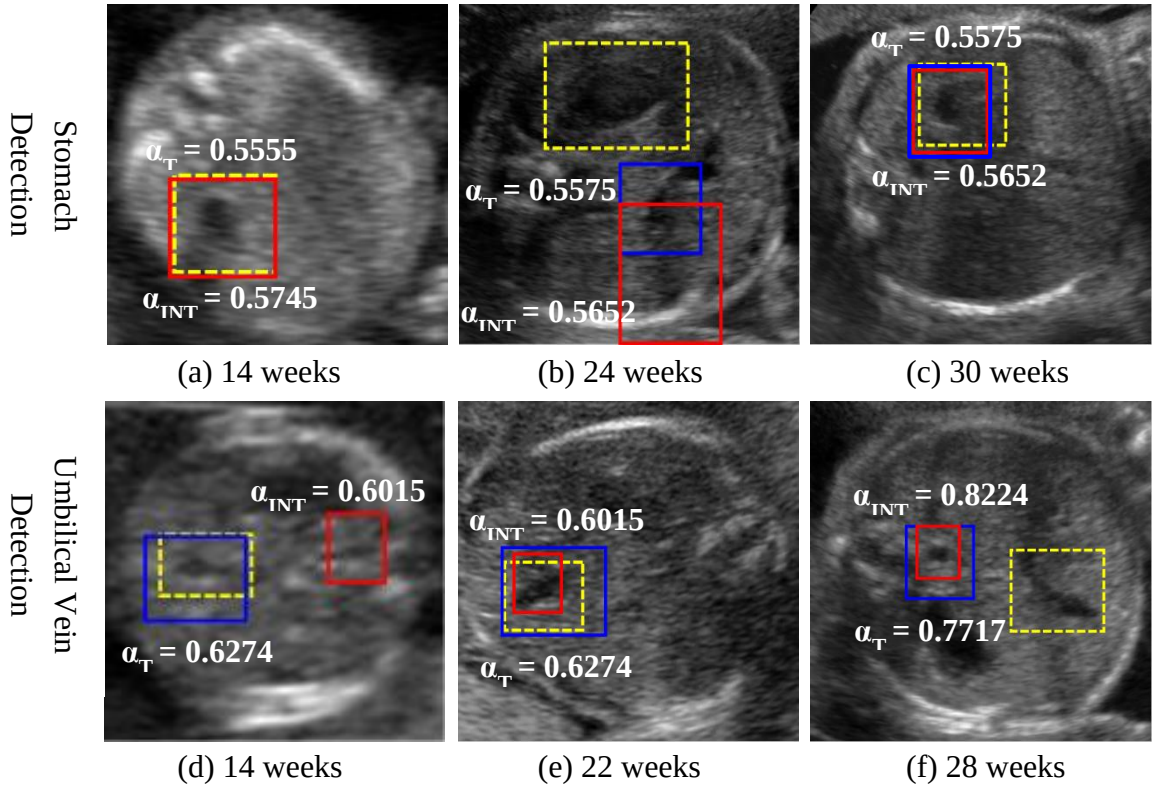


Figure 4.11: True positive results for umbilical vein detection by the "Intensity+MSLP" method (blue box with α_T) where it corrected the false detection result achieved by using "Intensity" features (red box with α_{INT}). Yellow box represents the ground truth.

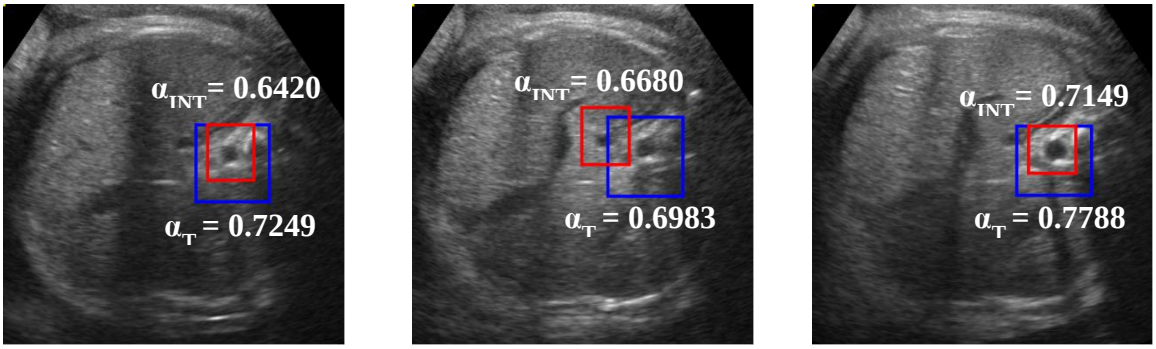


(a) 14 weeks (b) 24 weeks (c) 30 weeks

(d) 14 weeks (e) 22 weeks (f) 28 weeks

Figure 4.12: Example of images where objects were failed to be detected correctly. The blue box and α_T represent the detection by the "Intensity+MSLP" method, and the "Intensity" method detection is represented by the red box and α_{INT}. The yellow box represents the ground truth. Top row is for the stomach detection and the bottom row is for the umbilical vein detection. In (a) – (e), the scores were lower than the threshold value and resulted in false negative detections. In (f), a false positive result produced when other blood vessel was detected with higher classifier score.

Based on the ROC plots in Figure 4.9 the method that used the "*Intensity + MSLP*" feature sets achieved better performance in the detection of the stomach and the umbilical vein for all gestational age ranges except for the umbilical detection in the most advanced age range (38^{+0} – 42^{+6} weeks). However, there were only 4 negative images in this age range so strong conclusions cannot be drawn from this. In all these images, the "*Intensity + MSLP*" method detected the blood vessels near the spine which were assigned with high classifier scores. This caused a low specificity values in the detection, and lowered the ROC plot for the age range in Figure 4.9 (b). Three of these images are shown in Figure 4.13.



(a) (b) (c)

Figure 4.13: The false positive results for umbilical vein detection in images at 38 weeks. The umbilical veins were too elongated, hence not acceptable under the scoring criteria. The detection methods detected other vessels near the spine (blue and red boxes). The classifier scores achieved by "Intensity+MSLP" method (α_T) were higher than the threshold value, which lead to false positive detections in all images. The "Intensity" method achieved lower scores (α_{INT}) in (a) and (b) enabling a true negative detection but a higher score in (c) resulted in a false positive detection.

4.6 Conclusions

In this chapter, local phase images derived using the monogenic signal with multi-scale log-Gabor filters were introduced into the machine learning framework for object detection in fetal abdominal images. The proposed features were combined with the intensity features to produce a strong classifier. The high discriminative power of local phase features enabled them to be selected as the first weak classifiers with large classifier weight (α) in both (the stomach and the umbilical vein) detector models. The results demonstrated that the local phase features in combination with intensity features, managed to improve the detection of the stomach and the umbilical vein. Another alternative way of utilizing the information from the local phase images is the subject of the next chapter.

Chapter 5 Feature Symmetry for Efficient Object Detection

The object detection frameworks of Chapter 3 and Chapter 4 utilise the sliding window approach where an exhaustive scan, using the sliding window technique, is performed for all possible translations and for a sparse set of scales to find the anatomical object in the query image. In this type of approach, a strong monolithic classifier is applied to all sub-windows within an image and one takes the maximum of the classification score as an indication of the presence or absence of an object. One inherent disadvantage of this approach is the significant redundancy in computation because of the large number of candidate sub-windows considered which are "weak candidates" i.e. have very low classifications score. Moreover, the number of sub-windows grows in proportion to the image size, which is computationally too expensive when dealing with large datasets. In this chapter, an alternative approach of using a feature measure derived from the local phase image known as feature symmetry is proposed for fast object detection. As we will see, this leads to an improved result over an exhaustive evaluation of the strong classifier over all sub-window regions in an image.

The motivation for the approach is presented in Section 5.1. The proposed method is described in Section 5.2 followed by the experimental setup in Section 5.3. The results are presented in Section 5.4 along with the discussion. The chapter ends with the concluding remarks in Section 5.5.

5.1 Introduction

Sliding window approach has been the method of choice for generic object localization in natural images where the target is to find a bounding box around the object. In this approach, a quality function f, e.g. a classifier score, is evaluated over many sub-window regions of the image and taking its maximum as the object's location. The attempts to overcome the computational cost of such approach and speeding up the search can be grouped into two categories (Lampert et al., 2008). The first category consists of limiting the the coarse grid of possible rectangle locations and allowing only rectangles of certain fixed sizes as candidates (Dalal and Triggs, 2005, Ferrari et al., 2008). The second category involves local optimization method where promising regions in the images is first identified and then f is maximized by a discrete gradient ascent procedure (Bosch et al., 2007, Chum and Zisserman, 2007). In this work, the latter approach is adopted where the possible promising regions are identified through the application of the global detector on the image before the quality function (strong local classifier) is evaluated over these regions.

In this work, we propose the feature symmetry (FS) measure (Kovesi, 1997) derived from the local phase-based method as a global (coarse) feature for the localization of the region containing the stomach and the umbilical vein in fetal abdominal ultrasound image. A local (fine) detector which was trained in the previous experiment is then applied only to the locations deemed probable by the global features. This avoids the exhaustive sliding window method and as we will see achieves a faster computational speed.

5.1.1 Feature Symmetry

A dimensionless contrast invariant measure of symmetry (and asymmetry) constructed from local phase has been proposed by Kovesi (1997). It allows one to detect features (steps, ridges, valley) at points where there are consistency of local phase response (phase

congruency) over a range of scales. The stomach and the umbilical vein in fetal ultrasound images, typically appear as black blobs, with non-uniform intensity and sizes in different scans, and substantial shadowing and artefacts surrounding the objects. The local phase values within the stomach and the umbilical vein are found to be almost equivalent to π. Therefore, the phase congruency of the stomach and the umbilical vein could be computed using the multi scale feature symmetry measure derived from the monogenic signal components (Rajpoot et al., 2009). The feature symmetry (FS) measure is defined as:

$$FS(x,y) = \sum_{s} \frac{\lfloor [|even_s(x,y)| - |odd_s(x,y)|] - T_s \rfloor}{\sqrt{even_s(x,y)^2 + odd_s(x,y)^2} + \varepsilon} \tag{5.1}$$

where s represents the scale of the band-pass filter, ε is a small positive constant that avoids division by zero, $\lfloor . \rfloor$ operator denotes zeroing of any negative values, and T_s is scale specific noise compression term defined similarly in (Kovesi, 1996) as:

$$T_s = exp\left(mean\left(log\left(\sqrt{even_s(x,y)^2 + odd_s(x,y)^2}\right)\right)\right) \tag{5.2}$$

5.1.2 Scale Selection

To determine the appropriate scales of the band-pass filter for producing a feature symmetry image, we performed a feature selection experiment using the images from the training dataset of Chapter 3. Unary features were extracted from feature symmetry images produced using different types of scales combinations (refer to Table 5.1 for the different combinations used). The bandwidth of the filter (σ_ω) used in all the feature symmetry calculation was empirically set to 0.5. The AdaBoost algorithm was then used to find the features with the highest discriminative power. The first 10 selected features are shown in Table 5.2.

Feature symmetry values vary from a maximum of 1, indicating a very significant feature, down to 0 indicating no significance. Examples of feature symmetry images produced using three different combinations of scales with different threshold levels of feature significance are illustrated in Figure 5.1.

Table 5.1: Different combinations of band-pass filter scale used to produce the feature symmetry measure.

Notation	**Different Scales Combination (in pixels)**
ABC	[50, 100, 150]
DEF	[200, 250, 300]
BCD	[100, 150, 200]
CDE	[150, 200, 250]
ACE	[50, 150, 250]
BDF	[100, 200, 300]

Table 5.2: The filter scales combination and the weight of the first ten features selected by AdaBoost from the pool of unary features extracted from feature symmetry images.

Stomach		**Umbilical Vein**	
Scales combination of filter (pixels)	**Weight (α)**	**Scales combination of filter (pixels)**	**Weight (α)**
ACE	3.301	ABC	2.623
ABC	1.329	ACE	1.385
BDF	1.048	ACE	0.929
DEF	0.693	CDE	1.103
DEF	0.685	DEF	0.869
ACE	0.623	BDF	0.919
CDE	0.681	ACE	0.765
ACE	0.745	BCD	0.832
ABC	0.764	ABC	0.792
CDE	0.634	DEF	0.647

Based on the list of features in Table 5.2 and the visual observation of images (some examples shown in Figure 5.1), the scales combination of ACE ([50, 150, 250] pixels) was selected for the computation of the feature symmetry measure for the stomach and the umbilical vein. This is because majority of the first ten weak classifiers came from this scale combination. The features were assigned with high weights for its discriminative power in distinguishing the stomach and the umbilical vein from the non-objects and the background. Although the first feature selected for umbilical vein classification came from the scale combination of ABC [50, 100, 150], the feature symmetry produced using this scale also picks up a lot of noise, as shown in Figure 5.1.

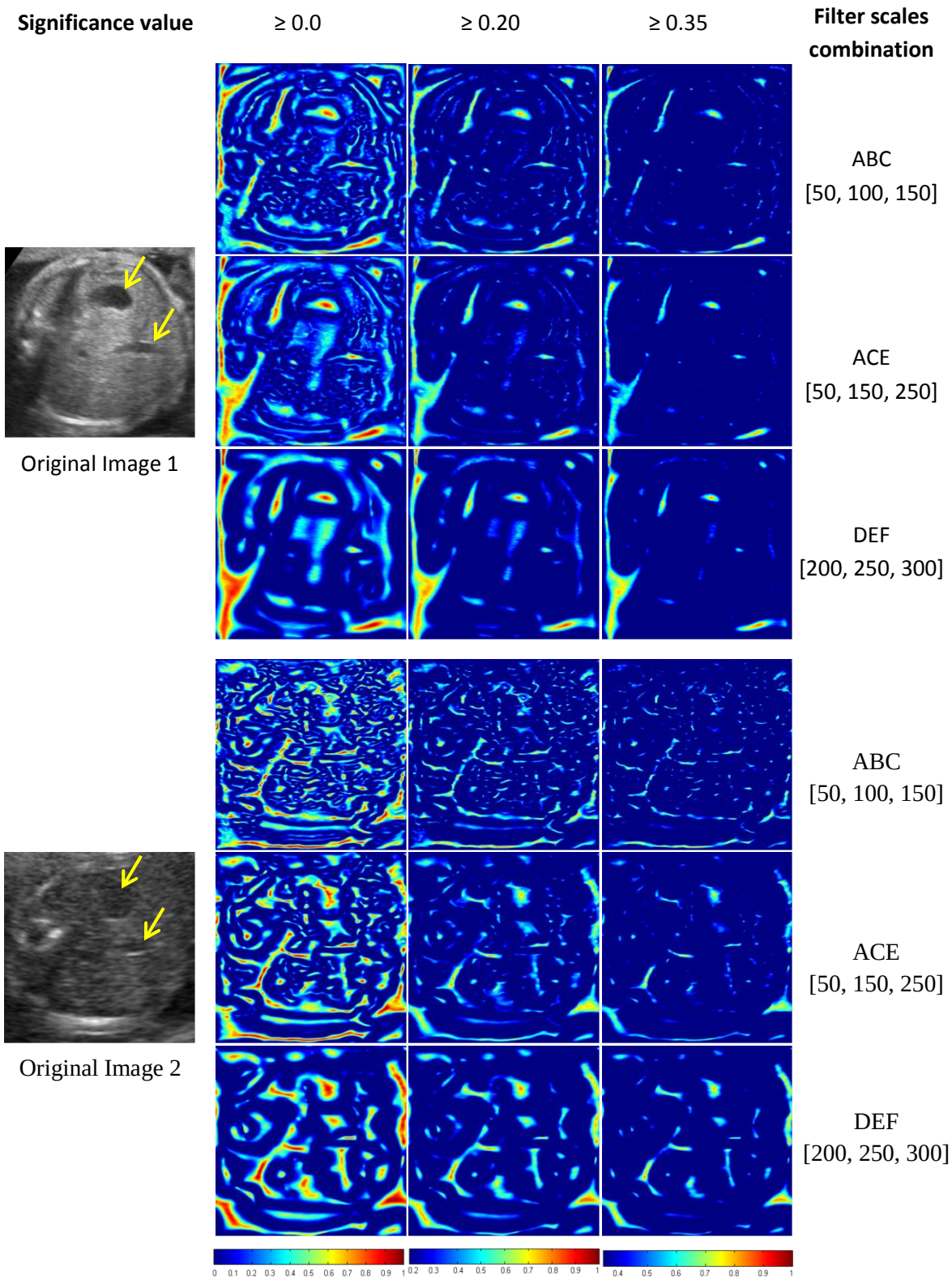


Figure 5.1: Examples of feature symmetry images produced using different scales combinations and threshold with three feature significance values. The yellow arrows on the original images indicate the location of the stomach and the umbilical vein.

We found that by setting the significance value for the feature symmetry measure to be 0.35, the stomach and the umbilical vein were identified from the feature symmetry image without having too many false positive candidates. This can be seen in Figure 5.1 where the images in the last column (significance value $\geq$ 0.35) are shown to have fewer objects compared to images produced with lower threshold of significance values (second column) and all significance values (first column).

Having selected candidate area, connected component labelling was then applied to the filtered feature symmetry image and the centre of the elements (centroid of the labelled components) estimated as candidate search centres. In summary the process to find the candidate location using the feature symmetry measure is as follows:

i) Produce the feature symmetry image,

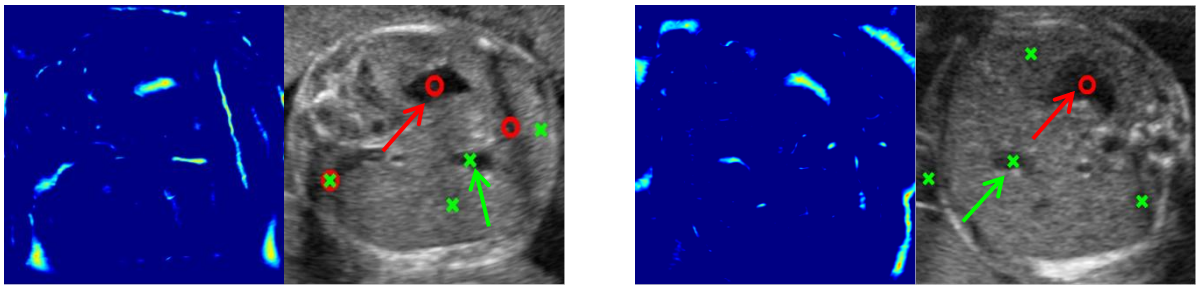
ii) Apply a threshold (significance value $\geq$ 0.35),

iii) Retain connected components, and

iv) Estimate the centroid of the labelled components.

Figure 5.2 shows the prediction of the candidate location of objects based on the global feature symmetry measure.

The classifier that was trained with the intensity features in Chapter 3 was used in this experiment as the local discriminative classifier. The local discriminative classifier was applied around search centres. In our validation set, the size of the stomach component in the labelled image usually varies from 800 to 2500 pixels and the umbilical vein from 300 to 1000 pixels. There might be overlap of potential components for the stomach and the umbilical vein but the task to distinguish them is passed to the local detector trained with the local features.



(a) (b) (c) (d)

Figure 5.2. Two examples of the global detector application. (a) and (c) show the feature symmetry map with significant features (>0.35) which were used to find the candidate locations for the stomach (red circles) and umbilical vein (green crosses) shown superimposed on the original intensity image in (b) and (d). The red and green arrows denote the correct positions for the stomach and the umbilical vein, respectively.

5.2 Experiments

Three different detection methods were compared:

1) The **"Local"** method where images were exhaustively scanned at multiple scales using the sliding window method. The detector was trained with Haar features extracted from the intensity images (described in Chapter 3)

2) The **"Global"** method where images were exhaustively scanned at multiple scales using the sliding window method. The detector was trained with unary features extracted from the feature symmetry images.

3) The **"Hybrid"** method where the detector trained with the local features (extracted from the intensity images) was applied at the candidate locations predicted by the global features (feature symmetry images).

In this experiment, the same testing dataset of positive and negative samples was used as described in Table 3.3.

5.3 Results and Discussions

The ROC curves produced for detection of the stomach and the umbilical vein of all the test images using the three different methods are shown in Figure 5.3. The ROC plots comparing the performance of the "*Hybrid*" and the "*Local*" methods in different gestational age groups are shown in Figure 5.4. The area under the curve (AUC) and the accuracy were recorded in Table 5.3 along with the average execution time. Performance metrics of the "*Hybrid*" and "*Local*" methods in different gestational age groups are shown in Table 5.4. The qualitative results that highlight the comparison between the methods were illustrated in Figure 5.5 to Figure 5.11.

Table 5.3: The overall performance of the three different methods in the detection of the stomach and the umbilical vein in fetal abdominal images.

	Area Under Curve (AUC)		**Accuracy (%)**		**Mean Execution Time (secs)**
	Stomach	**Umbilical Vein**	**Stomach**	**Umbilical Vein**	
Local	0.80	0.57	78.94	62.80	10.27
Global	0.71	0.53	69.28	57.99	10.65
Hybrid	0.88	0.75	82.75	72.55	0.94

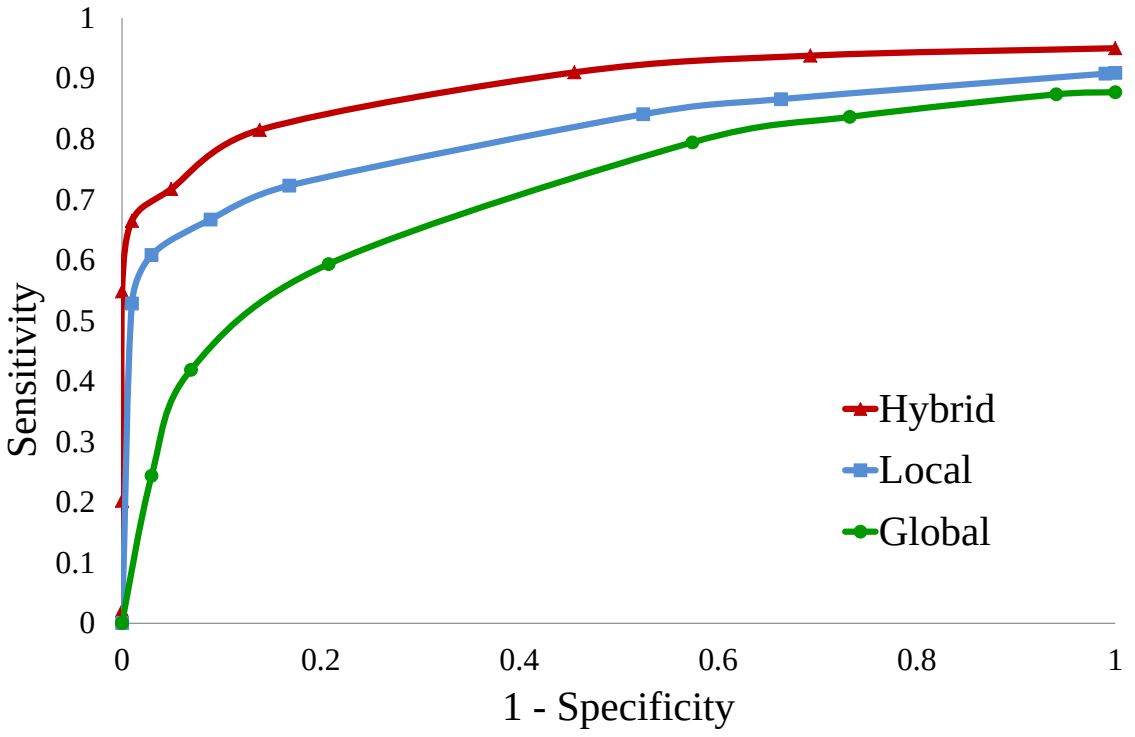
Table 5.4: Performance evaluation for the detection of the stomach and the umbilical vein between "Local" and "Hybrid" methods in different gestational age groups. Results for "Global" method were much lower than the other two methods and had been omitted.

Stomach Detection

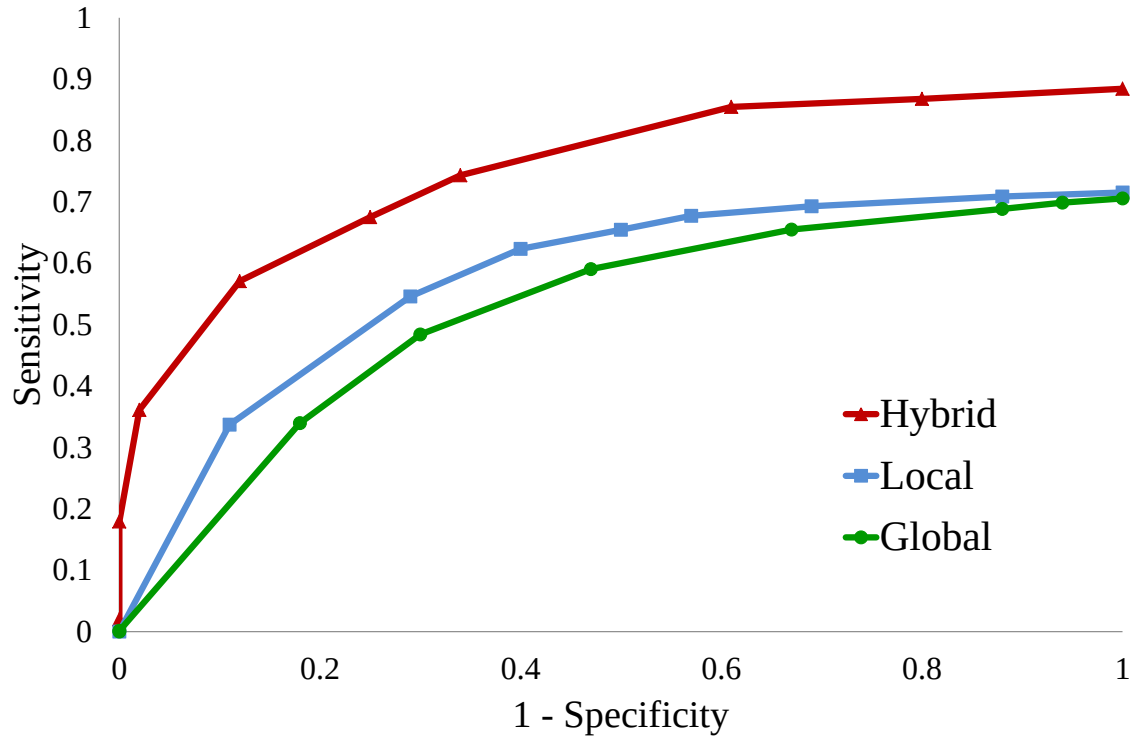
	14^{+0} – 19^{+6} weeks		**20^{+0} – 25^{+6} weeks**		**26^{+0} – 31^{+6} weeks**		**32^{+0} – 37^{+6} weeks**		**38^{+0} – 42^{+6} weeks**		**Overall**	
	Local	Hybrid	Local	Hybrid	Local	Hybrid	Local	Hybrid	Local	Hybrid	Local	Hybrid
Balanced Accuracy (%)	79.58	82.58	81.79	86.00	74.30	81.41	74.49	80.00	78.77	75.00	78.94	82.75
Sensitivity	0.63	0.68	0.64	0.72	0.58	0.63	0.56	0.60	0.58	0.50	0.61	0.66
Specificity	0.97	0.97	1.00	1.00	0.91	1.00	0.93	1.00	1.00	1.00	0.96	0.99
Area under curve (AUC)	0.80	0.85	0.82	0.90	0.79	0.91	0.77	0.89	0.82	0.86	0.80	0.88

Umbilical Vein Detection

	14^{+0} – 19^{+6} weeks		**20^{+0} – 25^{+6} weeks**		**26^{+0} – 31^{+6} weeks**		**32^{+0} – 37^{+6} weeks**		**38^{+0} – 42^{+6} weeks**		**Overall**	
	Local	Hybrid	Local	Hybrid	Local	Hybrid	Local	Hybrid	Local	Hybrid	Local	Hybrid
Balanced Accuracy (%)	56.10	67.93	77.46	83.24	79.67	86.99	68.40	65.69	55.60	31.90	62.80	72.55
Sensitivity	0.48	0.41	0.55	0.66	0.59	0.74	0.60	0.59	0.61	0.39	0.55	0.57
Specificity	0.64	0.95	1.00	1.00	1.00	1.00	0.77	0.73	0.50	0.25	0.71	0.88
Area under curve (AUC)	0.53	0.67	0.67	0.91	0.71	0.92	0.62	0.72	0.61	0.41	0.57	0.75



(a)



(b)

Figure 5.3: ROC curves for the detection of (a) the stomach and (b) the umbilical vein, using "Local", "Global" and "Hybrid" methods.

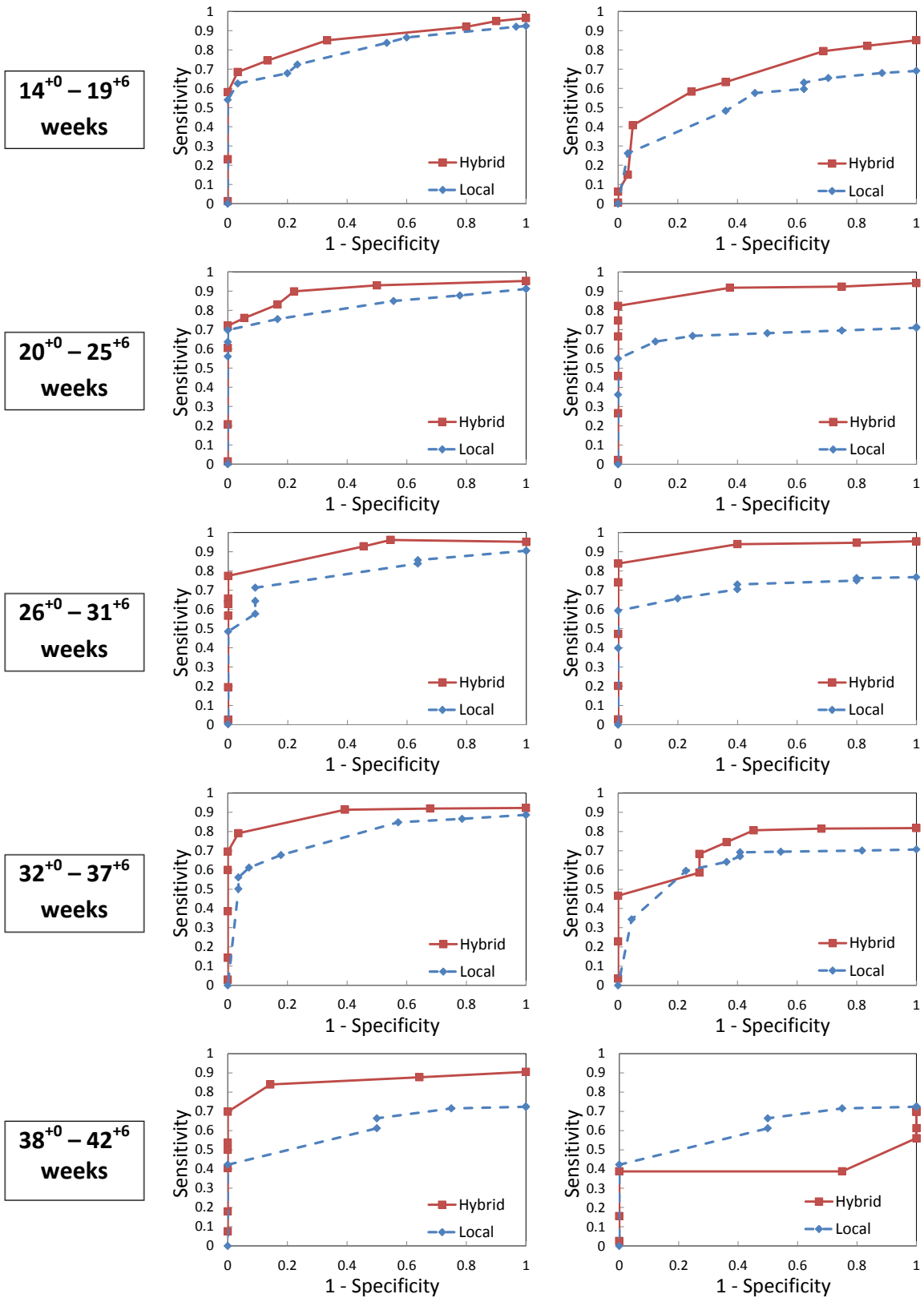


Figure 5.4: Comparison of ROC curves between "Hybrid" and "Local" methods in different gestational age groups.

The improvements achieved in the detection using the "*Hybrid*" method compared to the "*Local*" method are reflected in the increased percentage of the overall balanced accuracy from 78.94% to 83.92% for the stomach and 62.80% to 72.55% for the umbilical vein. Examples of the improved true positive detection for the stomach and the umbilical vein are shown in Figure 5.5 and Figure 5.6, respectively. Comparison of the performance of both methods for different gestational age groups reveals the same trend, except in the advanced age group ($38^{+0} – 42^{+6}$ weeks) where the "*Hybrid*" method performed poorly compared to the "*Local*" method.

The accuracy of the stomach detection in the $38^{+0} – 42^{+6}$ weeks group decreased by 2.77% and the umbilical vein detection decreased from 55.60% to 31.90%. This decrease in performance in the advanced age group was mainly attributed to the shadowing which obscured the anatomical object of interest. Examples of these misdetections are given in Figure 5.7 and Figure 5.8. The feature symmetry map failed to pick up the correct location of the objects for it to be passed as an input to the local detector. The shadowing mainly affected the umbilical vein detection in this age group, hence the decrease in accuracy. Umbilical vein, being a much smaller object than the stomach, was half-obscured in most of the images in this age group.

However, the "*Hybrid*" method achieved a notable improvement for the detection of both anatomical features in the crucial age groups for the clinical application ($20^{+0} – 25^{+6}$ weeks and $26^{+0} – 31^{+6}$ weeks where the abdominal circumference measurements are normally taken from the fetus abdominal scan). The specificity values of 1.00 were reached for the stomach and the umbilical vein detection in both the age groups. The highest sensitivity value of 0.72 for stomach detection was attained in the $20^{+0} – 25^{+6}$ weeks and 0.74 for the umbilical vein detection in the $26^{+0} – 31^{+6}$ weeks.

In general, we found that using global features from the feature symmetry images eliminates a lot of false positives caused by using local detector alone. In the case of detecting the absence of the stomach, "*Hybrid*" method achieved a total increase of 3 true negative results compared to the "*Loca*l" method. The images were from the gestational age groups of 26^{+0} – 31^{+6} weeks and 32^{+0} – 37^{+6} weeks, and the detections were illustrated in Figure 5.9. The only negative image that was missed (resulted in False Positive detection) was shown in Figure 5.10. This was also missed when the *"Local"* method was used. As for the umbilical vein detection, a large increase in the specificity was achieved in the earliest age groups (14^{+0} – 19^{+6} weeks) where 19 more true negative results were recorded. Some of the examples of the results are shown in Figure 5.11.

Another major contribution of the "*Hybrid*" method is the fast computation time, with an average of 0.94 seconds, a nine-fold decrease compared to the average computation time of the sliding-window method.

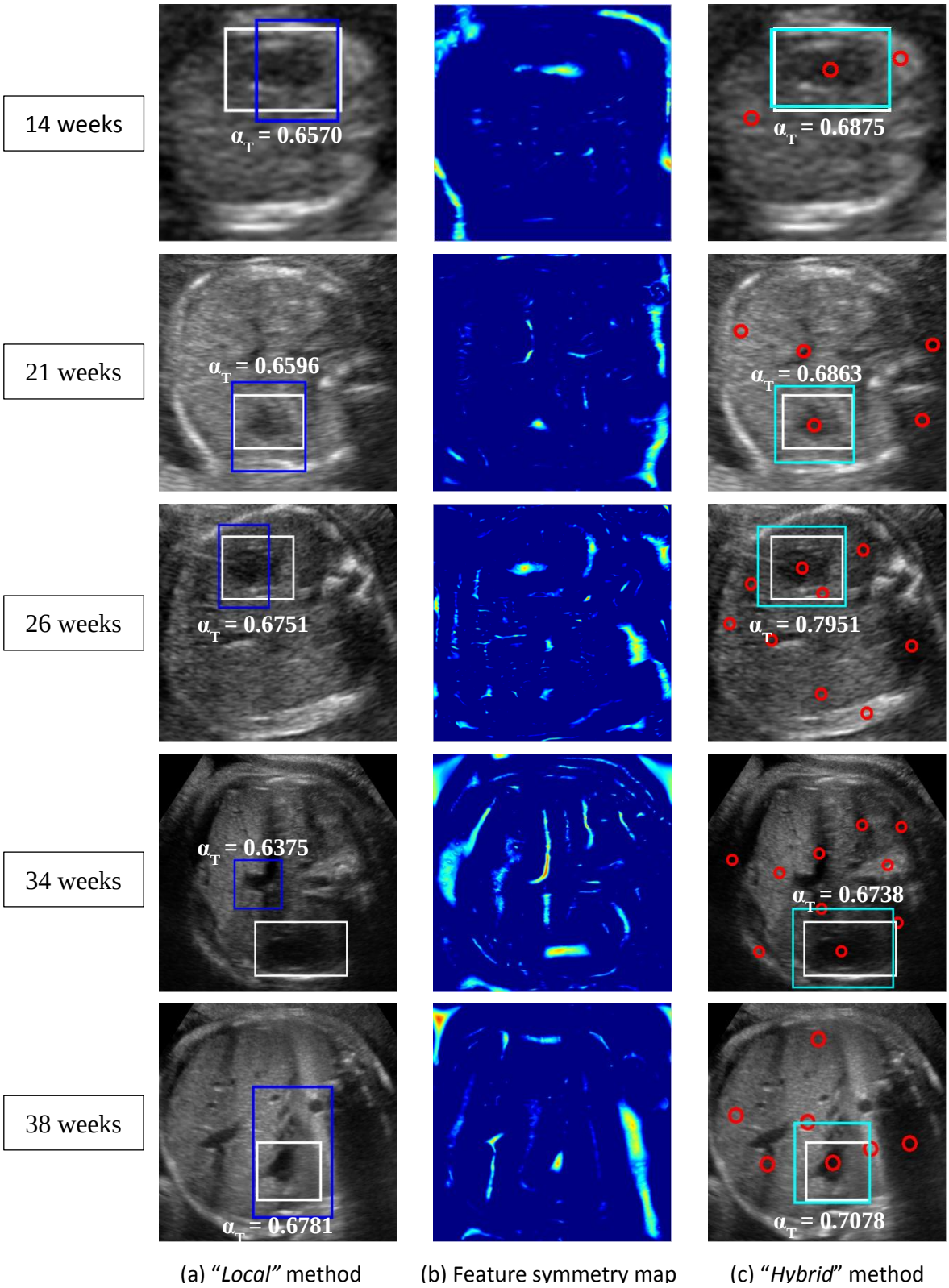


Figure 5.5: Examples of stomach detection where the false negative detection by the "Local" method were corrected by the "Hybrid" method. The global feature symmetry map inputs the candidate locations (red circles) for the local detector. White box represents the ground truth and the coloured boxes represent the detections by each method with classification scores α_T.

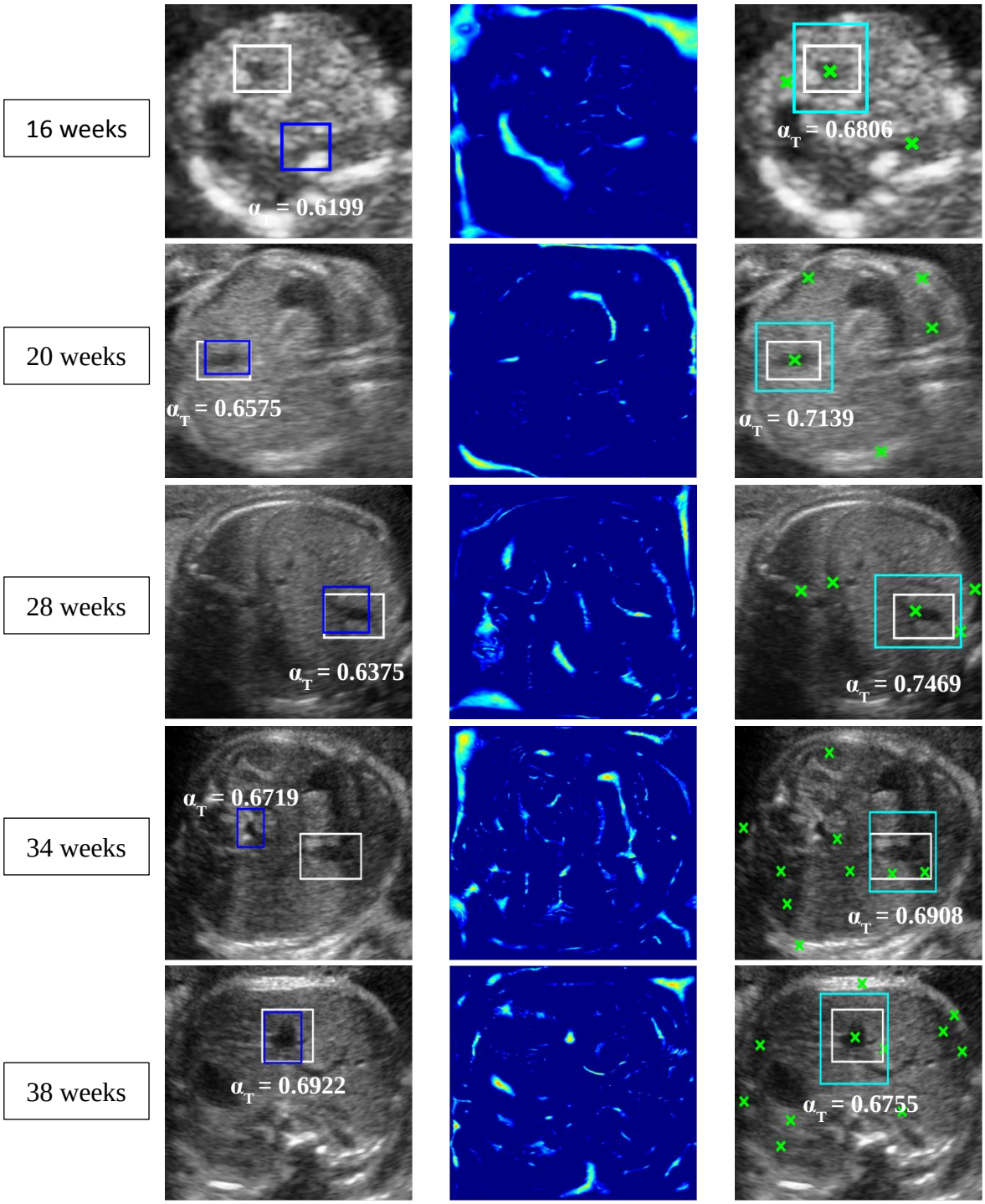
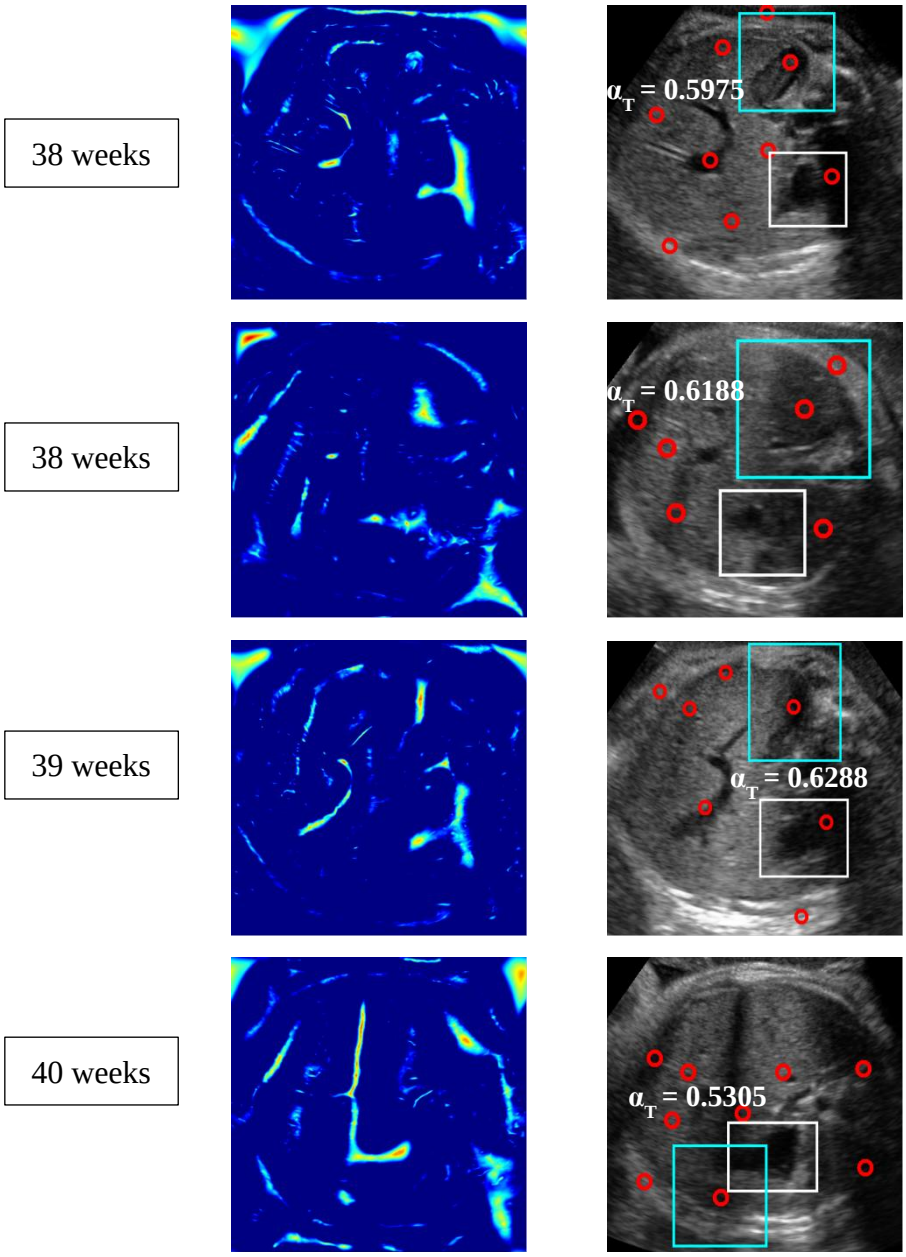
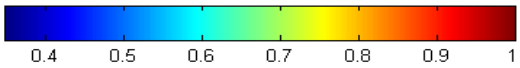


Figure 5.6: Examples of umbilical vein detection where the false negative detection by the "Local" method were corrected by the "Hybrid" method. The global feature symmetry map inputs the candidate locations (green crosses) for the local detector to be applied on. White box represents the ground truth and the coloured boxes represent the detections by each method with classification scores α_T. (d) Colour bar for the feature symmetry ma

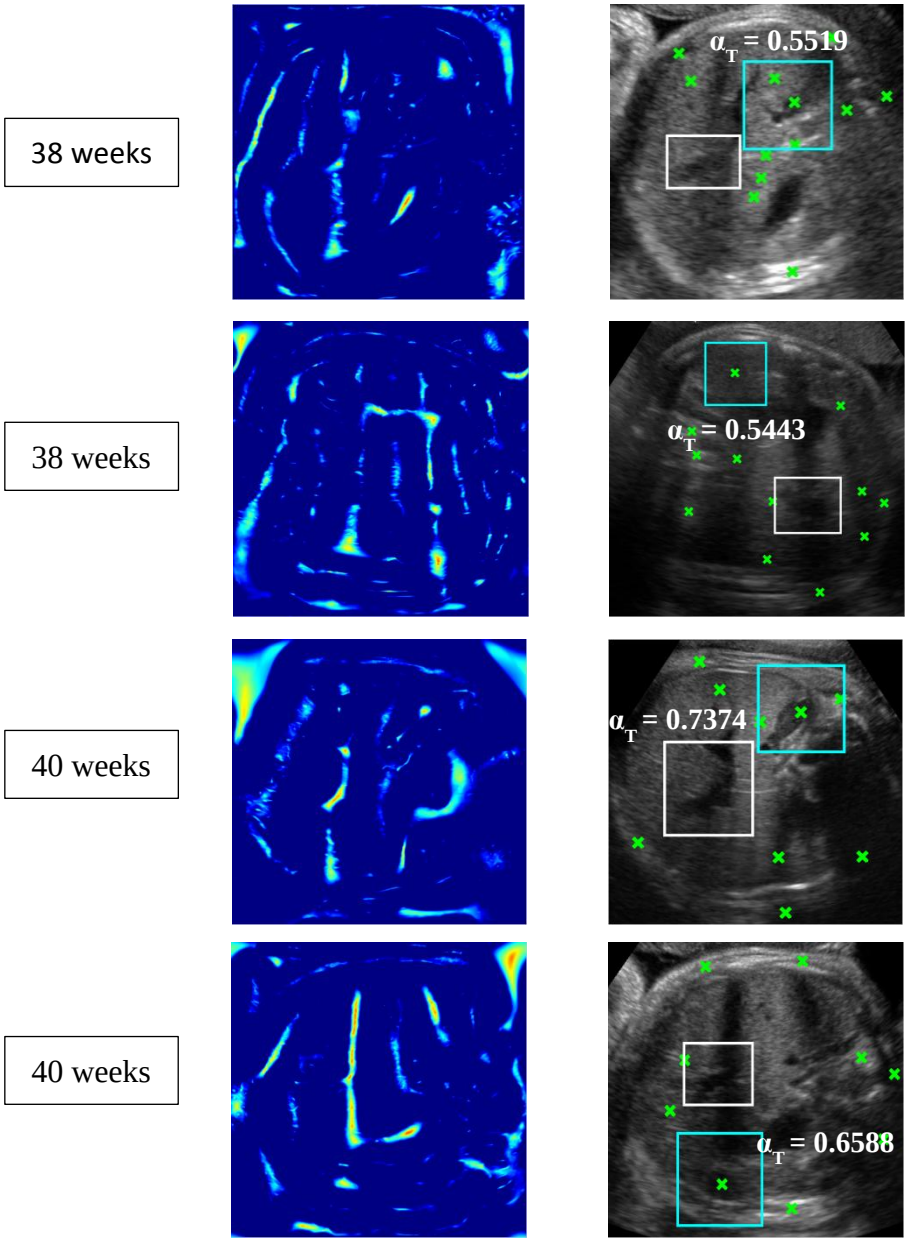


(a) Feature symmetry map (b) *"Hybrid"* method detection



(c) Colour bar for the feature symmetry map

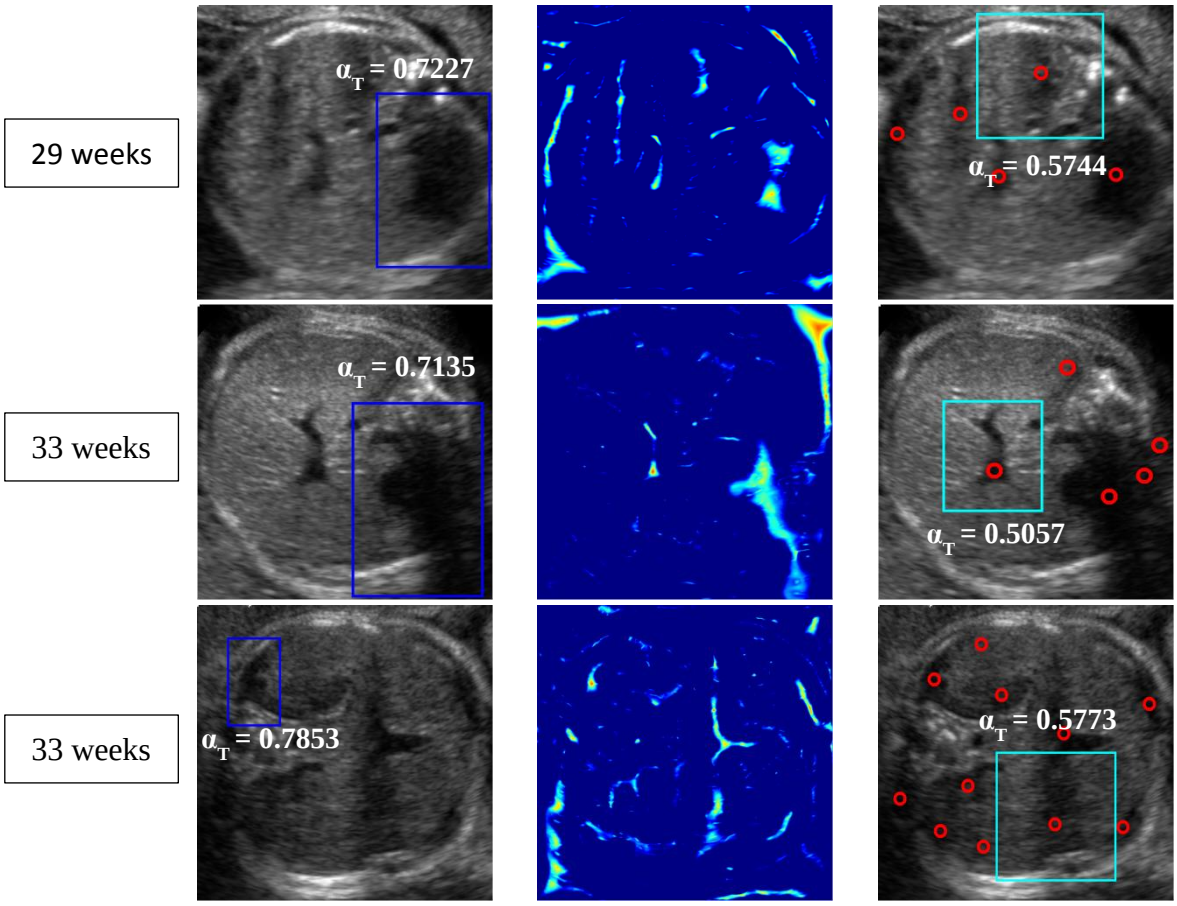
Figure 5.7: Examples of the misdetection of the stomach in the 38^{+0} – 42^{+6} weeks images. The location of the stomach could not be identified correctly through the feature symmetry map due to the shadowing effect over the stomach area. White box represents the ground truth and the coloured boxes represent the detections by the "Hybrid" method with classification scores α_T. The three top rows resulted in False Positive detections since the classification scores were higher than the threshold value and the bottom row resulted in False Negative detection.



(a) Feature symmetry map (b) *"Hybrid"* method detection

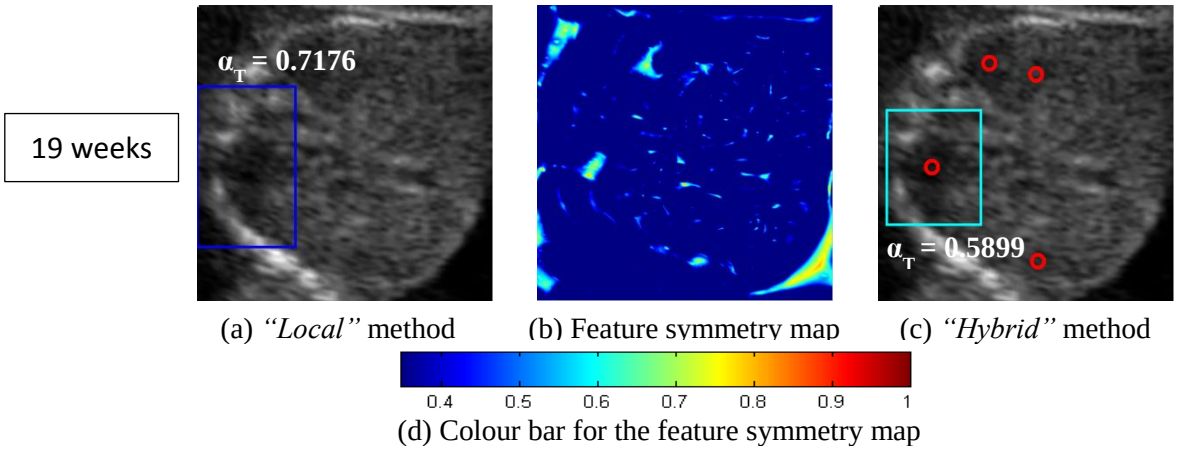
(c) Colour bar for the feature symmetry map

Figure 5.8: Examples of the umbilical vein misdetections in the 38^{+0} – 42^{+6} weeks images. The location of the umbilical vein could not be identified correctly through the feature symmetry map due to the shadowing effect over the umbilical vein area. White box represents the ground truth and the coloured boxes represent the detections by the "Hybrid" method with classification scores α_T. The two top rows produced False Negative results (low α_T) and the two bottom rows produced False Positive results (high α_T).



(a) *"Local"* method (b) Feature symmetry map (c) *"Hybrid"* method

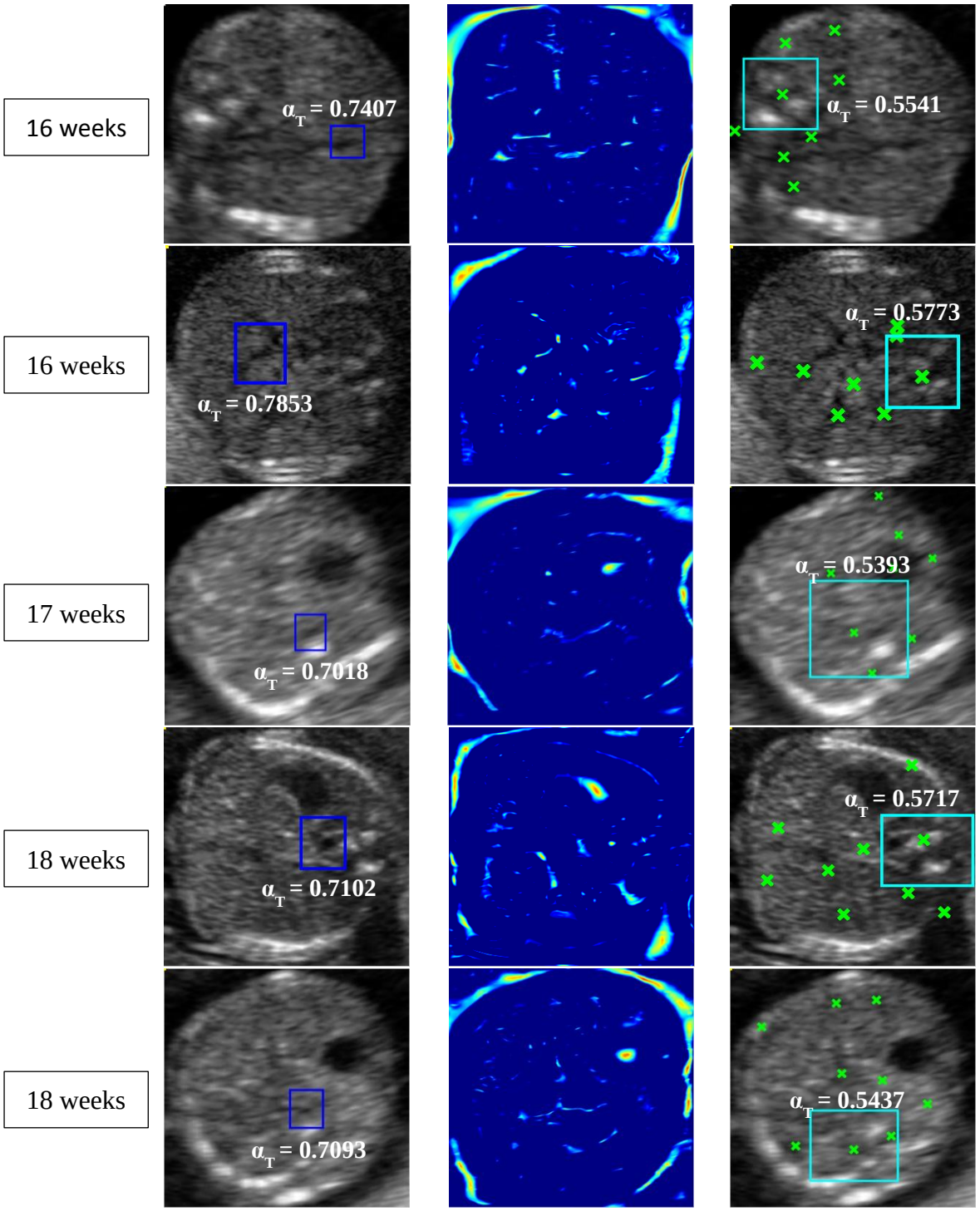
Figure 5.9: The three negative stomach images where false positive detection (high α_T score) by the "Local" method were corrected by true negative result (low α_T score) using the "Hybrid" method. The global feature symmetry map inputs the candidate locations (red circles) for the local detector to be applied on. The coloured boxes represent the detections by each method.



(a) *"Local"* method (b) Feature symmetry map (c) *"Hybrid"* method

(d) Colour bar for the feature symmetry map

Figure 5.10: The negative image that was missed by both "Local" and "Hybrid" methods. The scores achieved by both methods were higher than the threshold value and resulted in a false positive detection. The coloured boxes represent the detections by each method.



(a) *"Local"* method (b) Feature symmetry map (c) *"Hybrid"* method

(d) Colour bar for the feature symmetry map

Figure 5.11: Examples of negative umbilical vein images where false positive detection (high α_T score) by the "Local" method were corrected by true negative result (low α_T score) using the "Hybrid" method. The global feature symmetry map inputs the candidate locations (green crosses) for the local detector to be applied on. The coloured boxes represent the detections by each method.

5.4 Conclusions

This chapter presented an alternative approach to the sliding window method of Chapter 3 for the detection of anatomical objects in fetal ultrasound image. The global feature symmetry map derived from the local phase computation of the images was integrated within a machine learning framework that trains a local classifier using local Haar features from intensity images. This provides a computationally cheap step before invoking a local object detector to be applied in plausible locations. The proposed method exhibits better generalization capability when tested on 2384 images with an accuracy of 82.75% and 72.55% for the detection of the stomach and the umbilical vein, respectively. It also has faster computation time than the typical local object detector with the sliding window approach. It was observed that the method achieved high accuracy detection by focusing only on the high probability region and discarding many false positives candidates as in the sliding-window method. However, it is not suitable for applications in images where the objects are obscured partially since the correct locations cannot be predicted through the use of the feature symmetry map. For the application presented in this work, the object occlusion mostly happened in the advanced age group and less in the crucial age groups of the application.

Chapter 6 Two Pilot Studies to Illustrate Potential Clinical Utility

This chapter presents the application of the proposed machine learning detection method of Chapter 5 in two clinical pilot studies. In the first application, the detection method was used to record the presence and absence of the stomach and the umbilical vein in a standard fetal abdominal ultrasound scan. The method is compared with 4 expert observers. The second application applies the method to the problem of selecting the diagnostic 2D plane from a 3D fetal ultrasound volume for fetal biometry measurement.

6.1 Pilot Study 1: Comparison with Inter-Experts Agreement

In this section, we assess the agreement between the proposed automated method in Chapter 5 with expert observers in assessing the presence and absence of the stomach (SB) and the umbilical vein (UV) in the fetal abdominal biometry scan.

6.1.1 Experiments

Images used in this experiment were retrieved from the image database of the Oxford Ultrasound Quality Control Unit of the INTERGROWTH-21st project, an international multicentre research effort on fetal, neonatal and infant growth. All ultrasound scans were performed using a Philips HD9 ultrasound machine with a 2-5MHz 2D probe by ultrasonographers trained to follow standardized procedures for the study.

Two experiments were conducted in two separate settings with two different datasets. In the first setting, 100 2D ultrasound images of the fetal abdomen from singletons between

18^{+0} – 37^{+6} weeks of gestation were selected manually from the database. Four expert ultrasonographers (Dr. Aris Papageorghiou, Dr. Ippokratis Sarris, Dr. Christos Ioannau and Dr. Caroline Knight) blindly and independently assessed each image for the presence or absence of the stomach and the umbilical vein. Three of the experts (Dr. Aris Papageorghiou, Dr. Ippokratis Sarris, Dr. Christos Ioannau) had been directly involved on the quality control aspect of the INTERGROWTH-21st study and were trained on the scanning protocol adopted by the unit.

The second setting involved a new separate datasets where 300 2-dimensional ultrasound images of the fetal abdomen from singletons between 18^{+0} – 37^{+6} weeks of gestation were randomly selected by the computer. The images were assessed blindly and independently by three expert ultrasonographers (Dr. Ippokratis Sarris, Dr. Christos Ioannau and Dr. Caroline Knight).

Note that the difference in the two datasets above was in the method used for image selection, where the first dataset were done manually to ensure that it contained a good amount of negative samples and hence potentially selected with human bias. The second datasets were introduced to guarantee random images and to eliminate any possible bias in the image selections.

6.1.2 Results

Table 6.1 shows confusion matrices between the automated method and the experts and Table 6.2 illustrates the confusion matrices for the classification of the stomach and the umbilical vein in Dataset 1 between the experts. Confusion matrices for Dataset 2 classification between the automated method and experts are illustrated in Table 6.3 and classification between the experts in Table 6.4.

The inter-observer agreements were reported by computing the agreement between the experts and also between the automated method and the experts. The description of observer agreement statistics generally used in clinical imaging and benchmark scales to interpret the results are given in Appendix A. Because the rates of absence of objects in the images were much lower than its presence, we decided to report the prevalence-adjusted bias-adjusted kappa (PABAK) values (see Section A.3) to avoid the paradoxical conclusion with Cohen's kappa (see Section A.2). The percentage of agreement and the adjusted kappa values for Dataset 1 and Dataset 2 are summarized Table 6.5 and Table 6.6, respectively.

Table 6.1: Confusion matrices for the classification of the stomach and the umbilical vein in 100 ultrasound images (Dataset 1) between the automated method and the experts.

Stomach

	Automated Method		
Expert 1	**Present**	**Absent**	Total
Present	68	10	78
Absent	9	13	22
Total	77	23	100

	Automated Method		
Expert 2	**Present**	**Absent**	Total
Present	65	8	73
Absent	12	15	27
Total	77	23	100

	Automated Method		
Expert 3	**Present**	**Absent**	Total
Present	58	5	63
Absent	19	18	37
Total	77	23	100

	Automated Method		
Expert 4	**Present**	**Absent**	Total
Present	44	7	51
Absent	33	16	49
Total	77	23	100

Umbilical Vein

	Automated Method		
Expert 1	**Present**	**Absent**	Total
Present	83	7	90
Absent	8	2	10
Total	91	9	100

	Automated Method		
Expert 2	**Present**	**Absent**	Total
Present	79	6	85
Absent	12	3	15
Total	91	9	100

	Automated Method		
Expert 3	**Present**	**Absent**	Total
Present	81	4	85
Absent	10	5	15
Total	91	9	100

	Automated Method		
Expert 4	**Present**	**Absent**	Total
Present	72	8	80
Absent	19	1	20
Total	91	9	100

Table 6.2: Confusion matrices for the classification of the stomach and the umbilical vein in 100 ultrasound images (Dataset 1) between the experts.

Stomach

	Expert 1		
Expert 2	**Present**	**Absent**	Total
Present	66	7	73
Absent	12	15	27
Total	78	22	100

	Expert 1		
Expert 3	**Present**	**Absent**	Total
Present	61	2	63
Absent	17	20	37
Total	78	22	100

	Expert 1		
Expert 4	**Present**	**Absent**	Total
Present	42	9	51
Absent	36	13	49
Total	78	22	100

	Expert 2		
Expert 3	**Present**	**Absent**	Total
Present	57	6	63
Absent	16	21	37
Total	73	27	100

	Expert 2		
Expert 4	**Present**	**Absent**	Total
Present	42	9	51
Absent	31	18	49
Total	73	27	100

	Expert 3		
Expert 4	**Present**	**Absent**	Total
Present	38	13	51
Absent	25	24	49
Total	63	37	100

Umbilical Vein

	Expert 1		
Expert 2	**Present**	**Absent**	Total
Present	82	3	87
Absent	8	7	13
Total	90	10	100

	Expert 1		
Expert 3	**Present**	**Absent**	Total
Present	83	2	85
Absent	7	8	15
Total	90	10	100

	Expert 1		
Expert 4	**Present**	**Absent**	Total
Present	77	3	80
Absent	13	7	20
Total	90	10	100

	Expert 2		
Expert 3	**Present**	**Absent**	Total
Present	77	8	85
Absent	8	7	15
Total	85	15	100

	Expert 2		
Expert 4	**Present**	**Absent**	Total
Present	73	7	80
Absent	12	8	20
Total	85	15	100

	Expert 3		
Expert 4	**Present**	**Absent**	Total
Present	73	7	80
Absent	12	8	20
Total	85	15	100

Table 6.3: Confusion matrices for the classification of the stomach and the umbilical vein in 300 ultrasound images (Dataset 2) between the automated method and the experts.

Stomach

	Automated Method		
Expert 1	**Present**	**Absent**	Total
Present	274	15	289
Absent	8	3	11
Total	282	18	300

	Automated Method		
Expert 2	**Present**	**Absent**	Total
Present	266	12	278
Absent	16	6	22
Total	282	18	300

	Automated Method		
Expert 3	**Present**	**Absent**	Total
Present	276	14	290
Absent	6	4	10
Total	282	18	300

Umbilical Vein

	Automated Method		
Expert 1	**Present**	**Absent**	Total
Present	273	12	285
Absent	12	3	15
Total	285	15	300

	Automated Method		
Expert 2	**Present**	**Absent**	Total
Present	262	8	270
Absent	23	7	30
Total	285	15	300

	Automated Method		
Expert 3	**Present**	**Absent**	Total
Present	274	11	285
Absent	11	4	15
Total	285	15	300

Table 6.4: Confusion matrices for the classification of the stomach and the umbilical vein in 300 ultrasound images (Dataset 2) between the experts.

Stomach

	Expert 1		
Expert 2	**Present**	**Absent**	Total
Present	277	1	278
Absent	12	10	22
Total	289	11	300

	Expert 1		
Expert 3	**Present**	**Absent**	Total
Present	282	8	290
Absent	7	3	10
Total	289	11	300

	Expert 2		
Expert 3	**Present**	**Absent**	Total
Present	274	16	290
Absent	4	6	10
Total	278	22	300

Umbilical Vein

	Expert 1		
Expert 2	**Present**	**Absent**	Total
Present	267	3	270
Absent	18	12	30
Total	285	15	300

	Expert 1		
Expert 3	**Present**	**Absent**	Total
Present	275	10	285
Absent	10	5	15
Total	285	15	300

	Expert 2		
Expert 3	**Present**	**Absent**	Total
Present	261	24	285
Absent	9	6	15
Total	270	30	300

Table 6.5: Percentage of agreement and adjusted kappa value between the automated method (AM) and the experts (E1, E2, E3, E4) for Dataset 1.

	Stomach		**Umbilical Vein**	
Observers Compared	**Percentage of agreement**	**PABAK (95% CI)**	**Percentage of agreement**	**PABAK (95% CI)**
AM – E1	81%	0.62 (0.46 – 0.77)	85%	0.70 (0.56 – 0.84)
AM – E2	81%	0.60 (0.44 – 0.76)	82%	0.64 (0.49 – 0.79)
AM – E3	76%	0.52 (0.35 – 0.69)	86%	0.72 (0.58 – 0.86)
AM – E4	60%	0.20 (0.01 – 0.39)	73%	0.46 (0.29 – 0.63)
E1 – E2	81%	0.62 (0.47 – 0.77)	89%	0.78 (0.65 – 0.90)
E1 – E3	81%	0.62 (0.47 – 0.77)	91%	0.82 (0.71 – 0.93)
E1 – E4	55%	0.10 (-0.09 – 0.30)	84%	0.68 (0.54 – 0.82)
E2 – E3	78%	0.56 (0.40 – 0.72)	84%	0.68 (0.54 – 0.82)
E2 – E4	60%	0.20 (0.01 – 0.39)	81%	0.62 (0.47 – 0.77)
E3 – E4	62%	0.24 (0.05 – 0.43)	81%	0.62 (0.47 – 0.77)

*CI – Confidence Interval

Table 6.6: Percentage of agreement and adjusted kappa value between the automated method (AM) and the experts (E1, E2, E3, E4) for Dataset 2.

	Stomach		**Umbilical Vein**	
Observers Compared	**Percentage of agreement**	**PABAK (95% CI)**	**Percentage of agreement**	**PABAK (95% CI)**
AM – E1	92.33%	0.85 (0.79 – 0.91)	92.00%	0.84 (0.78 – 0.90)
AM – E2	90.67%	0.81 (0.75 – 0.88)	89.67%	0.79 (0.72 – 0.86)
AM – E3	93.33%	0.87 (0.81 – 0.92)	92.67%	0.85 (0.79 – 0.91)
E1 – E2	95.67%	0.91 (0.87 – 0.96)	93.00%	0.86 (0.80 – 0.92)
E1 – E3	95.00%	0.90 (0.85 - 0.95)	93.33%	0.87 (0.81 – 0.92)
E2 – E3	93.33%	0.87 (0.81 – 0.92)	89.00%	0.78 (0.71 – 0.85)

*CI – Confidence Interval

6.1.3 Discussion

Generally, the agreement recorded for images in Dataset 1 were lower than the agreement for Dataset 2. This could be explained by the effect of the bias imposed during the manual selection procedure. There were more cases in Dataset 1 compared to Dataset 2, of the case of the object's presence is "borderline" and being difficult to classify. This is especially prevalent for the stomach, where the percentage of agreement between experts in Dataset 1 ranged from 66% - 81% compared to 81% - 91% for umbilical vein. It is also evident from the adjusted kappa values that ranged from 0.10 – 0.62 and 0.62 – 0.82 for the stomach and the umbilical vein, respectively. The process of manual selection was intended to include all type of images: negative images, positive images and borderline cases. The selection of borderline cases was biased towards the presence/absence of the stomach as its main reference point.

The experience of the expert was also another factor that affected the agreement. One of the experts[3] recently joined the INTERGROWTH team. The three other experts were involved in the quality control aspect of the INTERGROWTH project since 2008 and had been active in many discussions to overcome differences in image quality assessment including the assessment of presence and absence of the anatomical landmarks. This variability was reflected in the "moderate" to "good" agreement (according to Altman's benchmark's scale – see Table A.3 in Appendix A) achieved between these three experts for stomach classification in Dataset 1 (adjusted kappa value ranged from 0.56 – 0.62) in contrast to the "poor" to "fair" agreement for inter-expert agreement that involved E4 (adjusted kappa value ranged from 0.10 - 0.24).

[3] Expert E4 who analysed Dataset 1 is the same person as E3 in Dataset 2.

The stomach classification in Dataset 2 recorded the percentage agreements of 93.33% - 95.67% (inter-expert) and 90.67% - 93.33% (method-expert) with the adjusted kappa values of 0.87 - 0.91 (inter-expert) and 0.81 - 0.87 (method-expert) all reflecting “very good” agreement.

In Dataset 2, the agreement for umbilical vein classification between the method and the experts were in consistent to the agreement between experts. The percentage agreements are 89.00% - 93.33% and 89.67% - 92.67% for inter-expert and method-expert, respectively. The corresponding adjusted kappa values of 0.78 - 0.86 (inter-expert) and 0.79 - 0.85 (method-experts) indicate “good” to “very good” agreements.

6.1.4 Conclusion

This section presented the agreement between the automated method and the experts in detecting the presence and absence of the stomach and the umbilical vein. The results indicate that the agreement between the automated method and the experts were very good for computer-random-selected images and the agreement was comparable to inter-experts agreement. The variation in datasets and experts suggests potential agreement in daily clinical practice.

6.2 Pilot Study 2: Selection of Optimal Plane from Ultrasound Volume

The search for the standardized planes in a 3D ultrasound volume is a hard and time consuming process even for expert physicians. A scheme for finding the standardized planes would be beneficial in advancing the use of volumetric ultrasound for clinical diagnosis. Previous research conducted on finding diagnostic planes in 3D ultrasound images has mostly confined to the domain of 3D echocardiography. In one of the approach (Leung et al., 2008), slice to volume registration method was used to find standard anatomical views for rest and stress images. That approach however is not suitable for fetal ultrasound because of the variability introduced by the positioning of the probe on the patient body and the arbitrary fetal position during scanning.

Usually, expert users need to find several landmarks in order to identify the diagnostic plane. For example, the user must search for the stomach and the umbilical vein in order to find the standard plane for measuring the abdominal circumference. Figure 6.1 illustrates the image planes acquired at different locations in a fetal abdominal ultrasound volume showing different visible anatomical landmarks.

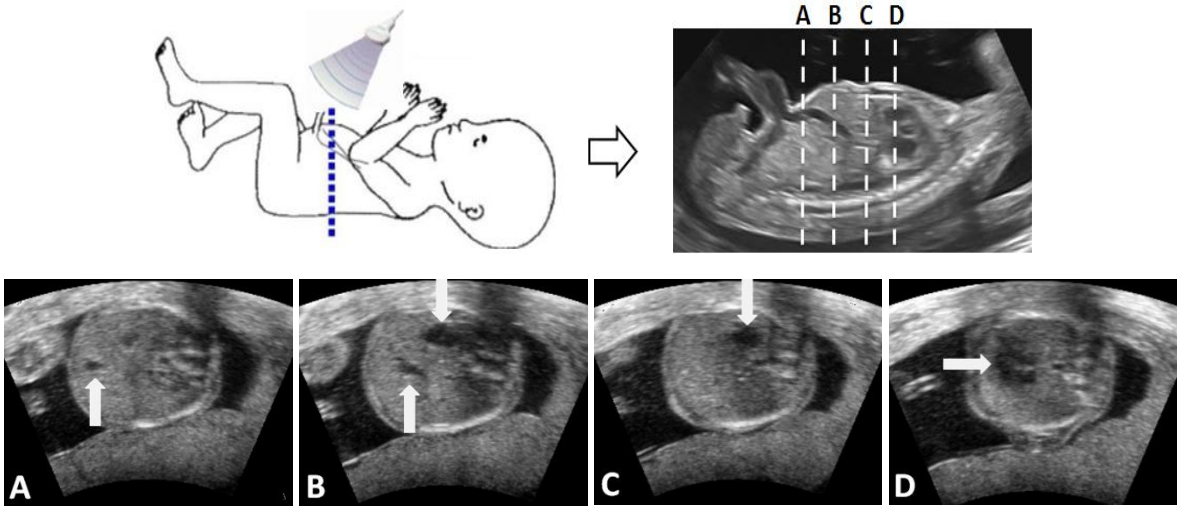


Figure 6.1: Illustration of different slices acquired from a 3D volume at different positions on the fetus. Slice A is too low indicated by umbilical vein (UV) (arrow) that is near the

abdomen wall. Slice B is an optimal slice with stomach bubble (SB) (top arrow) and UV (lower arrow) at correct position. Slice C has only SB (arrow) and Slice D is too high near the heart region (arrow).

6.2.1 Experiments

In this section, we propose an application for plane selection from the fetal abdominal ultrasound volume using our developed machine learning detection method for the stomach and the umbilical vein.

6.2.1.1 Dataset

The 45 fetal abdominal volumes used for this experiment were retrieved from the database of the Oxford Ultrasound Quality Control Unit of the INTERGROWTH-21st project. All ultrasound scans were performed using a Philips HD9 ultrasound machine with a 3-7MHz 3D probe. This work focused on the volumes of singleton pregnancies between 20^{+0} - 27^{+6} weeks of gestation because most of fetal abdominal measurements are performed within this timeframe. The scanned volumes in the database were saved in Philips MVL file format. These files were converted to a RAW file with metadata header using RegisterAndConvert© software (Gooding, 2009) with output resolution of 0.33mm in x, y, and z directions. The average size of a volume was 312x268x244 voxels. Ranges of image planes from each volume were annotated as standard optimal planes after consultation with a trained sonographer. Volumes were divided into three sets without overlap: 10 volumes for training, 5 volumes for the selection of the threshold level and 30 volumes for testing.

6.2.1.2 Method

We employed the detection method proposed in previous chapter for the detection of the two anatomical landmarks (SB and UV) and use the score from the detection to select a range of standardized image planes from the 3D volume.

The flowchart in Figure 6.2 gives the summary of the implementation. Our input volume was regarded as slices of axial planes. The features indicated by the weak classifiers were extracted from the image plane. The maximum score from the sub-windows contained in a particular plane was used to denote the probability of SB and UV detection for that image plane. The final score was calculated by dividing the product of both scores by their sum. The plots showing the scores returned by the method for all image planes in two test volume samples are shown in Figure 6.3 along with the image planes taken at four different positions in the volumes. The range of standard planes was selected from the peak with the global maximum point using an empirically determined threshold value (0.75) for all test volumes.

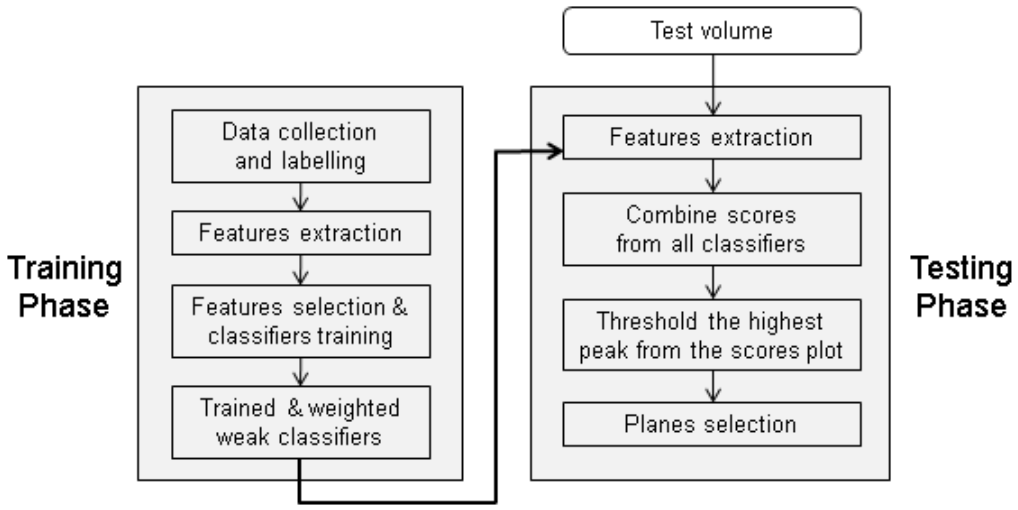
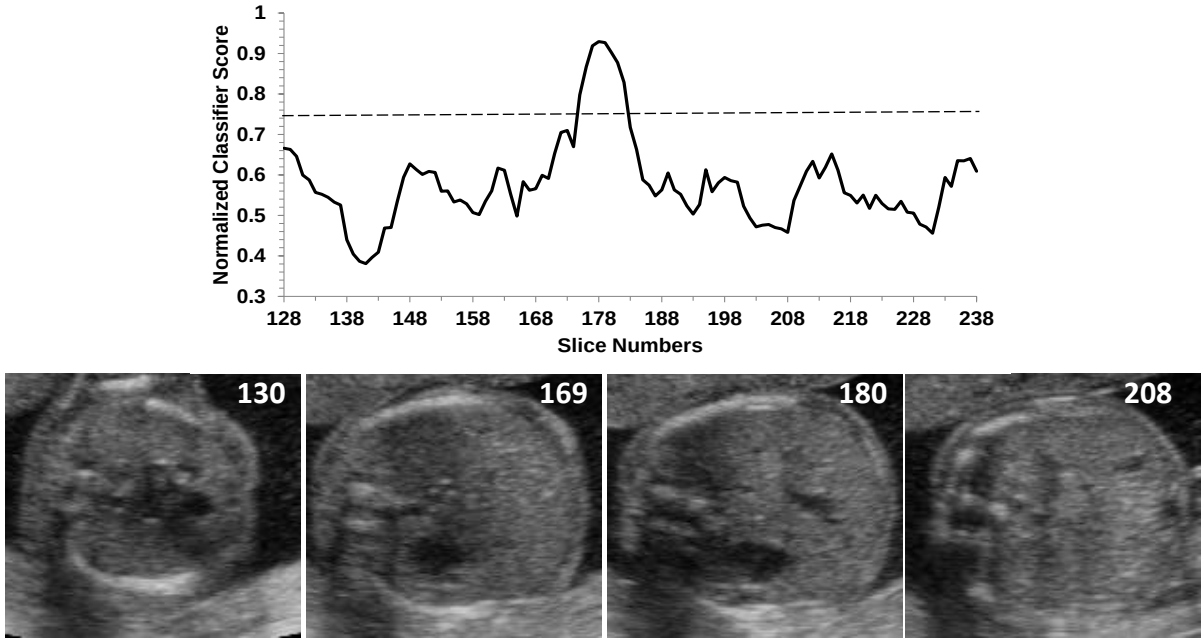
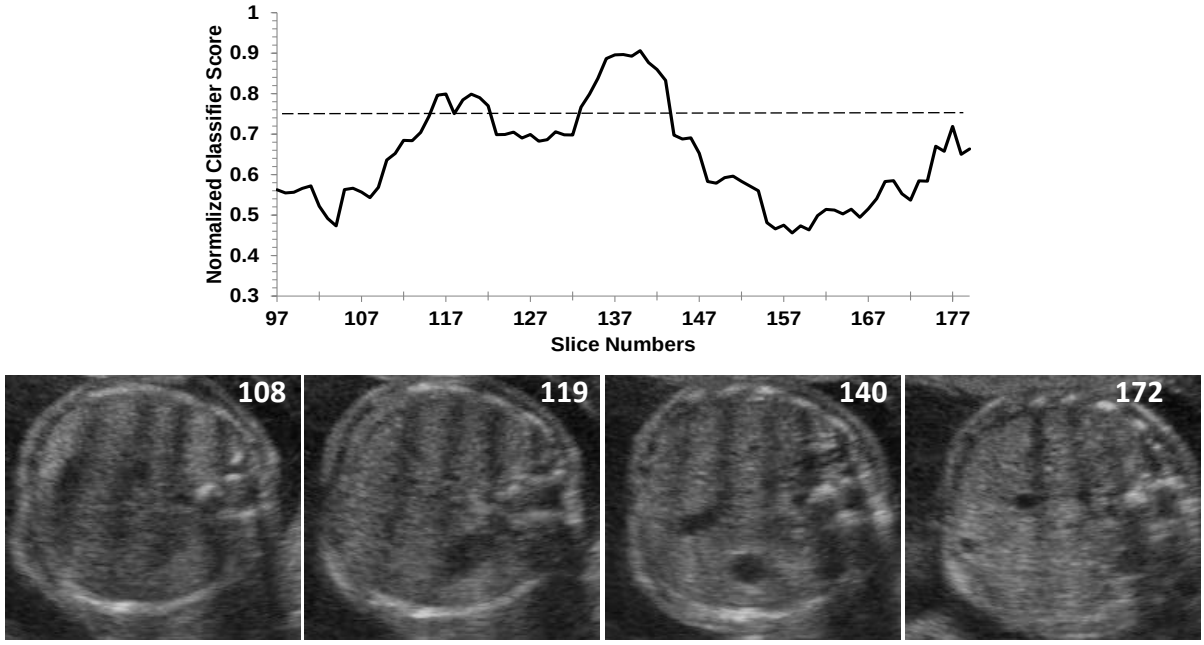


Figure 6.2: Flowchart showing the implementation of the training and testing phase.



(a) Sample Volume 1



(a) Sample Volume 2

Figure 6.3: Graphs showing the normalized classifier scores achieved by the detector for each image plane in two sample volumes. The dotted lines represent the threshold levels. The images shown were taken from four different locations (indicated by the image slice number) in the volume.

6.2.2 Results

The evaluation criteria used to evaluate the performance of the proposed method are:

$$\text{Precision} = \frac{\text{True Positive Planes}}{\text{True Positive Planes} + \text{False Positive Planes}} \tag{6.1}$$

$$\text{Recall} = \frac{\text{True Positive Planes}}{\text{True Positive Planes} + \text{False Negative Planes}} \tag{6.2}$$

The graph in Figure 6.4 shows the results achieved by using the proposed method to select the correct range of slices in 30 test volumes.

6.2.3 Discussion

The automated method showed accurate prediction of the manually labelled planes with an average recall of 93.36% and average precision of 73.32%. Most of the slices were correctly selected by our proposed method as indicated by the high recall percentage. The lower precision value indicates that not all the retrieved planes are correct (false positive). Through closer inspection, we found that most of the planes selected by the automated method contained both the SB and UV but the planes may have not been selected as "the best" by the expert because of a stricter criteria imposed including full visibility of the structures and position of the structure in the image. Some of the examples of these false positives planes are shown in Figure 6.5 and Figure 6.6.

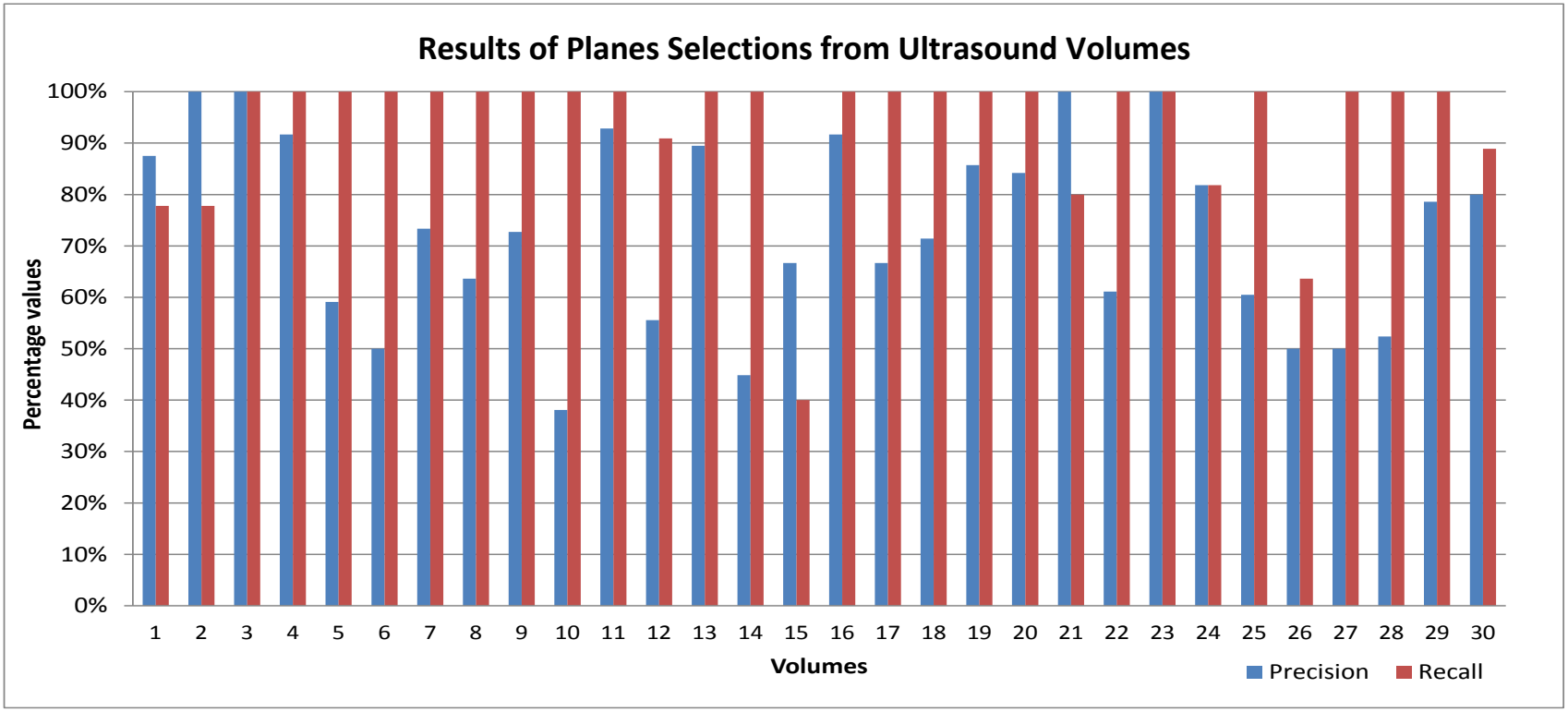


Figure 6.4: Precision and recall values (in percentages) achieved for the selection of standard planes from 30 fetal abdominal volumes.

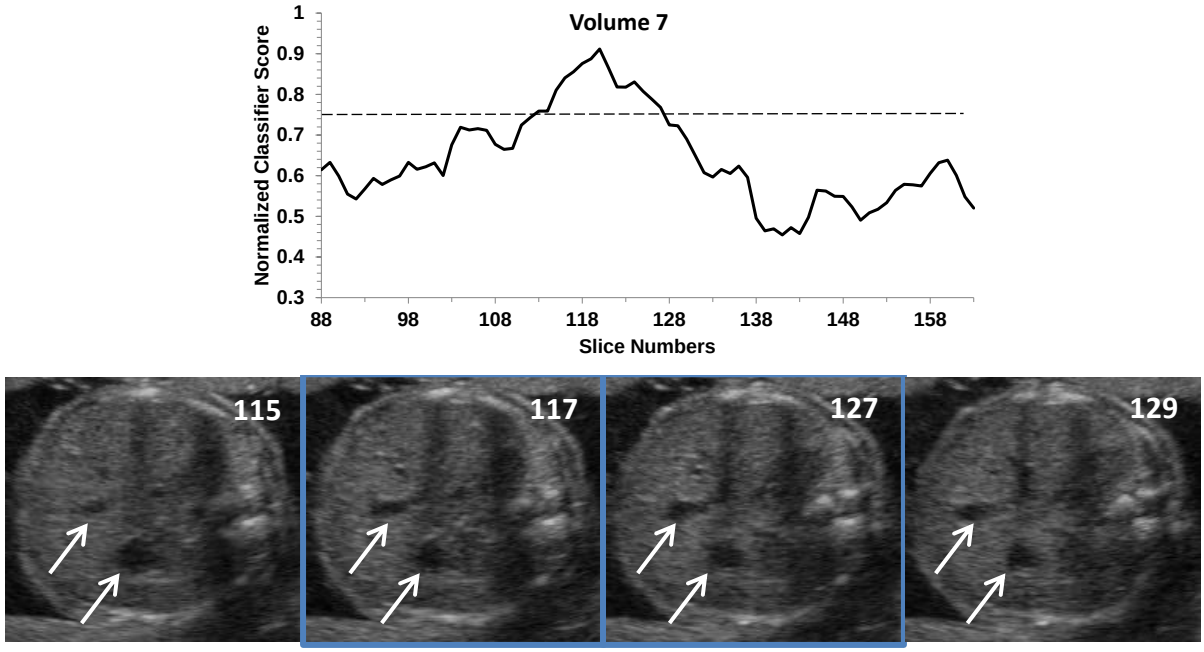


Figure 6.5: Graph plot and image planes for Volume 7. The standard planes defined by the expert were from 117 to 127 (with blue borders). The selected planes by the method were from 115 to 129. The arrows denote the stomach the umbilical vein, respectively.

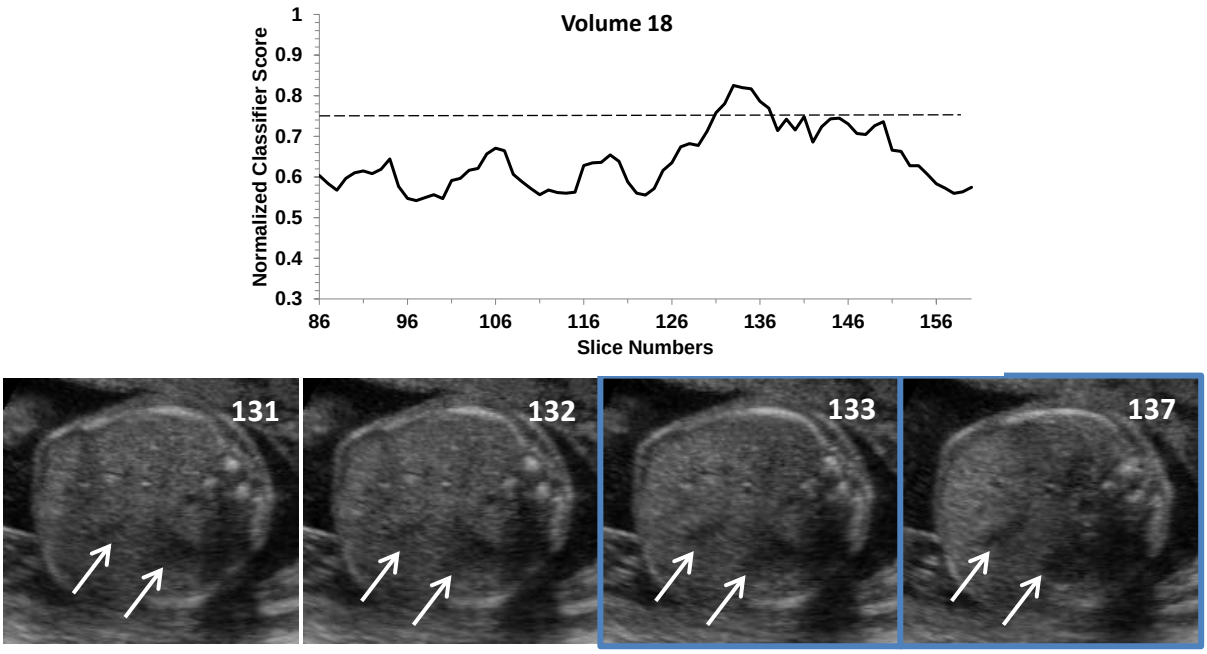


Figure 6.6: Graph plot and image planes for Volume 18. The standard planes defined by the expert were from 133 to 137 (with blue borders). The selected planes by the method were from 131 to 137. The arrows denote the stomach the umbilical vein, respectively.

The proposed method achieved 100% precision and recall values in retrieving standard planes from Volume 3 and Volume 23. The examples of the planes are shown in Figure 6.7.

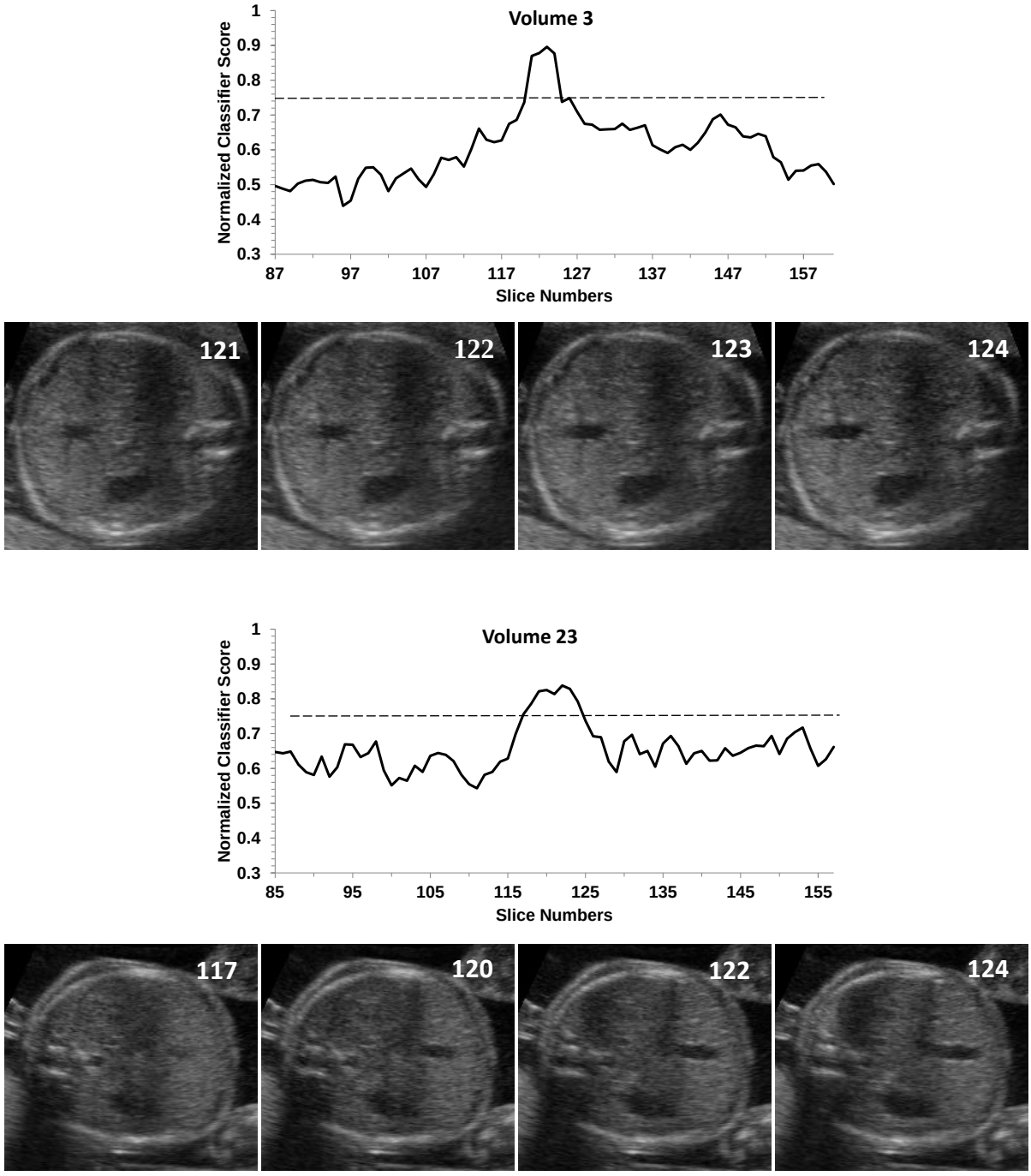


Figure 6.7: Graph plot and image planes from the volume with 100% precision and recall values. The standard planes for Volume 3 and Volume 23 were from plane 121 to plane 124 and from plane 117 to 124, respectively.

Volume 10 recorded the lowest precision value of 38.10%. This is because 13 out of the 21 planes that were selected were False Positive according to the definition used by the expert. The graph and the plane slices for Volumes 10 are shown in Figure 6.8.

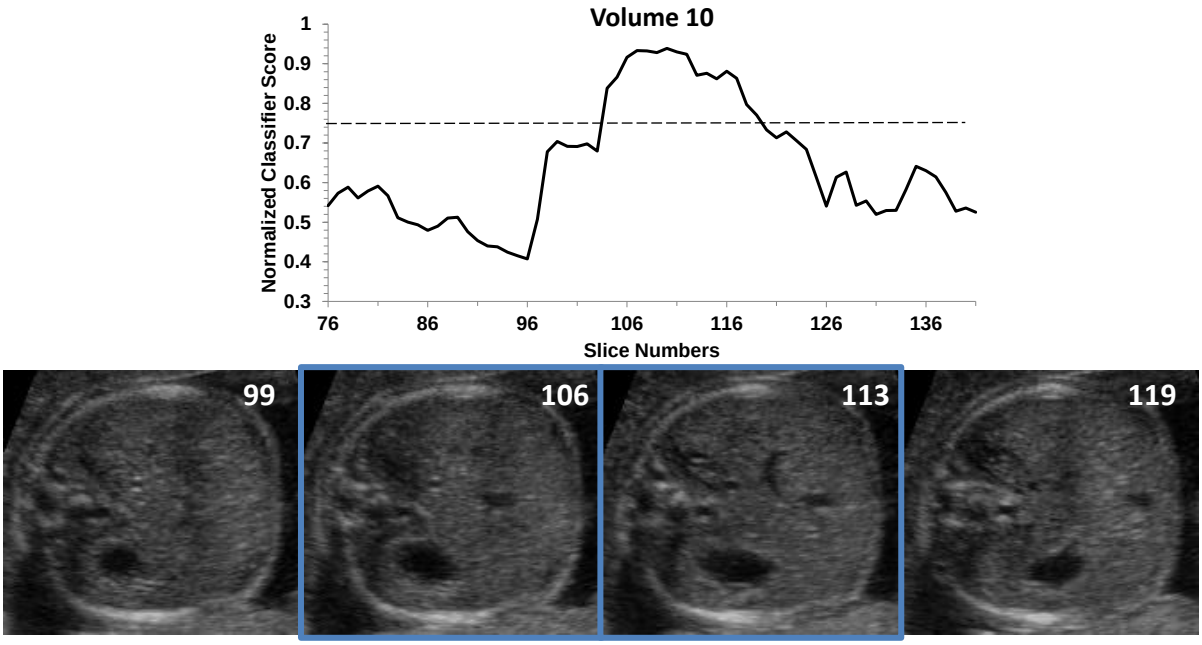


Figure 6.8: Graph plot and image planes from the volume with lowest recall percentage. The standard planes defined by the expert were from plane 106 to plane 113 (blue borders) while the automated method selected Plane 99 to Plane 119.

6.2.4 Conclusions

In this section, we have applied the detection method developed in the previous chapter to the application of selecting diagnostic image planes from fetal abdominal ultrasound volumes. The high recall percentage of 93.36% shows that the method was able to predict most of the planes selected by an expert. Detailed assessment involving more experts to check the inter- and intra-experts agreement would be needed to alleviate its potential to become a tool for improving the efficiency of 3D ultrasound imaging.

Chapter 7 Summary and Future Work

In this thesis, methods to automatically detect two main anatomical landmarks (the stomach and the umbilical vein) within fetal abdominal ultrasound scans had been developed and evaluated. This chapter will summarise the key points of the work presented in this thesis and suggests some possible direction for future studies.

7.1 Summary

The review of literature in Chapter 2 highlighted the need for a fast and reproducible quality assessment of fetal biometry images. Automated detection of anatomical landmarks i.e. the stomach and the umbilical vein in fetal abdominal scan is an important step in making sure that standard image plane has been acquired. The solution to this problem had never been attempted in the image analysis domain before.

In Chapter 3, an original solution for detecting the stomach and the umbilical vein in fetal ultrasound images was designed using a machine learning framework. Haar features extracted from the intensity image were used to train the detector using the AdaBoost learning algorithm. Detection took on an average of 10 seconds for it to be executed using MATLAB 7.6 (code not optimized) on a Pentium Xeon® 3.4 GHz machine. The qualitative and the quantitative results of the methods applied on different clinically relevant gestational age groups demonstrated that such a framework works well across most gestational ages.

Ultrasound images often have low contrast. This motivated the idea to introduce features extracted from local phase images into the detection framework. This work was presented in Chapter 4. The performance of the new versions of detection using local phase features derived using single-scale and multi-scale filters were compared using ROC plots

where the latter features were shown to produce superior result. The multi-scale local phase features were then combined with the intensity features for the classifier training. The high discriminative power of local phase features were proven by its selection as the first weak classifiers for both the stomach and the umbilical vein detector models with large classifier weights. The combination of both intensity and the local phase features were shown to have better detection results compared to using them separately.

An alternative approach to the sliding window method used in the two previous chapters was presented in Chapter 5. A mechanism of identifying candidate locations for the two anatomical objects was introduced using a global feature symmetry map based on local phase. This provides a computationally cheap step over exhaustive image scanning used in the sliding window methods. Different filter scales combinations were evaluated to find the best scales for producing the feature symmetry map. By applying the local classifier (trained using Haar features from the intensity images) on the plausible locations produced using the feature symmetry measure, the so-called hybrid method achieved higher accuracies in the detection of both the stomach and the umbilical vein with a nine-fold increase in the average computational speed.

Two potential application scenarios using the automated detection method were presented in Chapter 6. In the first application, the performance of the computer algorithm in recording the presence and absence of the stomach and the umbilical vein in standard fetal abdominal scan were compared with four expert observers. The computer-experts agreement was found to be comparable with the inter-expert agreement. The second application involved the application of the detection method to the problem of selecting diagnostics 2D plane from a 3D fetal ultrasound volume for fetal biometry measurement. The method showed a good prediction of the planes manually labelled by experts.

7.2 Future Work

This section suggests a number of directions in which the work presented in this thesis could be extended.

7.2.1 Other Anatomical Objects Detection

The method proposed in this work could potentially be applied to the detection of other anatomical objects in ultrasound images. Recently, the method was used in our lab for the detection of choroid plexus in fetal brain images with encouraging results (Namburete et al., 2012). The detection of important landmarks in fetal femoral plane (bone ends) and cephalic plane (thalami and cavum septi pellucidum) could be used for the qualitative assessment of that particular biometric image.

7.2.2 Multi-class Object Detection

A multi-class object detection which would take into account the spatial relationship between the two objects (i.e. the stomach and the umbilical vein) would be interesting to study and might be beneficial for identifying the objects as an entirety. Various multi-class detection schemes had been proposed for the detection of objects in natural images where sometimes features are shared between classes (Torralba et al., 2007, Das et al., 2008, Shotton et al., 2009). Challenges would be to find the right balance between feature sharing and discrimination between classes.

7.2.3 Testing on Data from Other Ultrasound Machines

The dataset used in this study were from Philips HD9 ultrasound machines. It would be interesting to investigate whether the performance of the method is comparable for images taken from other machine. The effect of different acquisition-dependent settings such as

contrast ans spatial resolution on the performance of the method should be studied. Also if the aim of image scoring system is to embed it in ultrasound machine for real-time functionality, it has to be rigorously tested through various machines.

7.2.4 3D Object Detection and Planar Slicing

The proposed detection method can be utilized for object detection in volumetric data by extending the features to 3D. 3D Haar features had been successfully used for the segmentation of the femur in ultrasound volumes (Yaqub et al., 2011) and the detection of key features in brain (Yaqub et al., 2012). 3D object detection would be beneficial for finding standard plane in volumes taken in any fetal orientations by doing planar slicing.

Appendix A Observer Agreement Statistics in Clinical Imaging

Statistical analysis of observer agreement in imaging is generally performed to provide information about the reliability of imaging diagnosis. It is also commonly used to compare the performance of humans and computers (Masmoudi et al., 2009, Tolouee et al., 2011). Among the common standard methods for description of agreement in regard to categorical data are:

1) Percentage of Agreement

2) Cohen's Kappa (κ)

3) Prevalence-Adjusted Bias-Adjusted Kappa (PABAK)

A.1 Percentage of Agreement

Generally, when two observers, A and B, express binary ratings of presence/absence for a particular anatomical object, the results are arranged in a 2 x 2 table as follows:

	Ratings by Observer A		
Ratings by Observer B	**Present**	**Absent**	Total
Present	*a*	*b*	g_1
Absent	*c*	*d*	g_2
Total	f_1	f_2	*N*

Based on the notation in the 2 x 2 table above, the overall proportion of observed agreement, p_o, is calculated as follows:

$$p_o = \frac{a+d}{N} \qquad (A.1)$$

This proportion is usually converted to a percentage. Percentage of agreement is an intuitive approach to measuring agreement but may give a false impression of performance when the number of readings is heavily imbalanced. For example, if the number of positive readings is large relative to the number of negative readings, the agreement in regard to positive readings will dominate the value of p_o.

Therefore, the positive and negative agreements are usually reported as an alternative to the overall agreement in the findings. This will give an indication of the type of decision on which readers disagree. The observed proportion of positive and negative agreement, p_{pos} and p_{neg}, are calculated with the following equations:

$$p_{pos} = \frac{a}{\left(\frac{f_1+g_1}{2}\right)} = \frac{2a}{f_1+g_1} \qquad (A.2)$$

$$p_{neg} = \frac{d}{\left(\frac{f_2+g_2}{2}\right)} = \frac{2d}{f_2+g_2} \qquad (A.3)$$

The advantage of calculation of p_{pos} and p_{neg} is that it reveals any imbalance in the proportion of positive and negative responses.

A.2 Cohen's Kappa (κ)

Although there had been many proposals for alternative approaches to measuring agreement, Cohen's kappa (Cohen, 1960) remains the most commonly used measure. The advantage of the κ coefficient is its adjustment for the proportion of cases in which the observers would agree by chance alone (Fleiss et al., 1969).

The proportions in the total column and in the total row represent the marginal probabilities, which are used as substitute for chance (since the true value of chance is unlikely to be known) (Crewson, 2005). Chance agreement is derived from the observed data, so it will likely change with different observers evaluating the same images. The proportion of chance agreement (p_e) is computed as follows:

$$p_e = \frac{f_1 g_1 + f_2 g_2}{N^2} \qquad (A.4)$$

The value of p_e becomes the correction factor for chance agreement. It is subtracted from the observed agreement (p_o) and from perfect agreement ($N/N = 1$), and the results are then divided to form κ:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (A.5)$$

A κ value of 1.0 represents perfect agreement and a value of zero indicates that there is no agreement. Interpretations of intermediate values are given in Section A.4. Although Cohen's kappa has been widely used as a simple single index of agreement, several authors have pointed out the paradoxes associated with its interpretation (Zwick, 1988, Feinstein and Cicchetti, 1990, Byrt, 1992). It had been suggested that the results of studies of agreement should include the three indices of κ, positive agreement, and negative agreement (Cicchetti and Feinstein, 1990). This is to provide more details about disagreements and the possibility of effects caused by the presence of biases between observers as well as by the true prevalence of the conditions being evaluated.

A.3 Prevalence-Adjusted Bias-Adjusted Kappa (PABAK)

Another index of agreement between two observers can be used as an alternative to Cohen's Kappa. This coefficient known as prevalence-adjusted bias-adjusted kappa

(PABAK), adjusts kappa for the differences in prevalence of the "Present" and "Absent" conditions, and for bias between observer (Byrt et al., 1993). In the process of adjusting for bias and prevalence, it was noted in that paper that the final formula for the index calculation was found to be equivalent to a coefficient that was proposed much earlier (Bennett et al., 1954), known as Bennett's *S* coefficient :

$$PABAK = \frac{\frac{2}{N}\left(\frac{a+d}{2}\right) - 0.5}{1 - 0.5} = 2p_o - 1 \qquad (A.6)$$

In (Zwick, 1988), it was pointed out that other coefficients that were proposed by a number of authors, including RE coefficient (Maxwell, 1977), C coefficient (Janson and Vegelius, 1979) and κ_n coefficient (Brennan and Prediger, 1981), were also found to be equivalent to the Bennett's *S* coefficient.

A.4 Benchmark Scale of Agreement Statistics

Several different benchmark scales had been proposed and used as guidelines for interpreting the magnitude of agreement statistics. Although they were developed to be used with kappa coefficients, they had also been used to communicate the results of other agreement coefficients. (Hartmann, 1977) stated a basic benchmark for kappa values where the proposed that they should exceed 0.6. A more detailed and commonly cited scale was proposed by (Landis and Koch, 1977) and it is shown in Table A.1.

Table A.1: Landis and Koch – Kappa's Benchmark Scale

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

Another benchmark scale was proposed that had only three categories (Fleiss, 1981). The three low value ranges of the Landis-Koch benchmark were collapsed into a single range. This is shown in Table A.2.

Table A.2: Fleiss – Kappa's Benchmark Scale

Kappa Statistic	Strength of Agreement
< 0.40	Poor
0.40 to 0.75	Intermediate to Good
> 0.75	Excellent

Another slightly different benchmark scale which is a modified version of Landis-Koch's scale was proposed by (Altman, 1991) and is illustrated in Table A.3. The only noticeable difference is the first two ranges of values of Landis-Koch's proposal were combined into a single category labelled as "Poor".

Although these benchmarks were produced arbitrarily and not supported by any evidence in the proposal, they served as a useful guideline in numerous studies that used agreement statistics (Constantinidis et al., 1996, Testa et al., 2005, Masselli et al., 2011, Mirjalili et al., 2012).

Table A.3: Altman – Kappa's Benchmark Scale

Kappa Statistic	Strength of Agreement
<0.20	Poor
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Good
0.81 to 1.00	Very Good

Bibliography

[1] ALTMAN, D. G. (1991). *Practical Statistics for Medical Research*, Chapman and Hall.

[2] AMIT, Y. & GEMAN, D. (1999) A Computational Model for Visual Selection. *Neural Computation,* **11** (7), 1691-1715.

[3] ANDERSEN, L. G., ÄNGQUIST, L., ERIKSSON, J. G., FORSEN, T., GAMBORG, M., OSMOND, C., BAKER, J. L. & SØRENSEN, T. I. A. (2010) Birth Weight, Childhood Body Mass Index and Risk of Coronary Heart Disease in Adults: Combined Historical Cohort Studies. *PLoS ONE,* **5** (11).

[4] ANQUEZ, J., ANGELINI, E. D. & BLOCH, I. (2008) Segmentation of Fetal 3d Ultrasound Based on Statistical Prior and Deformable Model. *In:* Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2008, 17-20.

[5] BARKER, D. J. P. (2006) Adult Consequences of Fetal Growth Restriction. *Clinical Obstetrics and Gynecology,* **49** (2), 270-283.

[6] BELAID, A., BOUKERROUI, D., MAINGOURD, Y. & LERALLUT, J. F. (2011) Phase-Based Level Set Segmentation of Ultrasound Images. *IEEE Transactions on Information Technology in Biomedicine,* **15** (1), 138-147.

[7] BENNETT, E. M., ALPERT, R. & GOLDSTEIN, A. C. (1954) Communications through Limited-Response Questioning. *Public Opinion Quarterly,* **18** (3), 303-308.

[8] BERGBOER, N. H., POSTMA, E. O. & HERIK, H. J. V. D. (2006) Context-Based Object Detection in Still Images. *Image and Vision Computing,* **24** (9), 987-1000.

[9] BERNSTEIN, I. M., HORBAR, J. D., BADGER, G. J., OHLSSON, A. & GOLAN, A. (2000) Morbidity and Mortality among Very-Low-Birth-Weight Neonates with Intrauterine Growth Restriction. *American Journal of Obstetrics and Gynecology,* **182** (1), 198-206.

[10] BISHOP, C. M. (1996). *Neural Networks for Pattern Recognition,* New York, Oxford University Press.

[11] BOSCH, A., ZISSERMAN, A. & MUNOZ, X. (2007) Representing Shape with a Spatial Pyramid Kernel. *In:* Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR), 2007, 401-408.

[12] BOUKERROUI, D., NOBLE, J. A. & BRADY, M. (2004) On the Choice of Band-Pass Quadrature Filters. *Journal of Mathematical Imaging and Vision,* **21** (1), 53-80.

[13] BRENNAN, R. L. & PREDIGER, D. J. (1981) Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement,* **41** (3), 687-699.

[14] BRITTO, I. S. W., DE SILVA BUSSAMRA, L. C., ARAUJO JÚNIOR, E., TEDESCO, G. D., NARDOZZA, L. M. M., MORON, A. F. & AOKI, T. (2009) Fetal Lung Volume: Comparison by 2d- and 3d-Sonography in Normal Fetuses. *Archives of Gynecology and Obstetrics,* **280** (3), 363-368.

[15] BURGES, C. J. C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* **2** (2), 121-167.

[16] BYRT, T. (1992) Problems with Kappa. *Journal of Clinical Epidemiology,* **45** (12), 1452.

[17] BYRT, T., BISHOP, J. & CARLIN, J. B. (1993) Bias, Prevalence and Kappa. *Journal of Clinical Epidemiology,* **46** (5), 423-429.

[18] CAMPBELL, S. & WILKIN, D. (1975) Ultrasonic Measurement of Fetal Abdomen Circumference in the Estimation of Fetal Weight. *British Journal of Obstetrics and Gynaecology,* **82** (9), 689-697.

[19] CAO, B., SHEN, D., SUN, J. T., YANG, Q. & CHEN, Z. (2007) Feature Selection in a Kernel Space. *In:* Proceedings of the International Conference on Machine Learning (ICML), 2007, 121-128.

[20] CARNEIRO, G., AMAT, F., GEORGESCU, B., GOOD, S. & COMANICIU, D. (2008a) Semantic-Based Indexing of Fetal Anatomies from 3-D Ultrasound Data Using Global/Semi-Local Context and Sequential Sampling. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2008a.

[21] CARNEIRO, G., GEORGESCU, B., GOOD, S. & COMANICIU, D. (2007) Automatic Fetal Measurements in Ultrasound Using Constrained Probabilistic Boosting Tree. *In:* Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2007, 571-579.

[22] CARNEIRO, G., GEORGESCU, B., GOOD, S. & COMANICIU, D. (2008b) Detection and Measurement of Fetal Anatomies from Ultrasound Images Using a Constrained Probabilistic Boosting Tree. *IEEE Transactions on Medical Imaging,* **27** (9), 1342-1355.

[23] CHALANA, V., WINTER 3RD, T. C., CYR, D. R., HAYNOR, D. R. & KIM, Y. (1996) Automatic Fetal Head Measurements from Sonographic Images. *Academic Radiology,* **3** (8), 628-635.

[24] CHANG, C. H., YU, C. H., CHANG, F. M., KO, H. C. & CHEN, H. Y. (2003) The Assessment of Normal Fetal Brain Volume by 3-D Ultrasound. *Ultrasound in Medicine and Biology,* **29** (9), 1267-1272.

[25] CHANG, C. H., YU, C. H., KO, H. C., CHEN, C. L. & CHANG, F. M. (2006) Predicting Fetal Growth Restriction with Liver Volume by Three-Dimensional Ultrasound: Efficacy Evaluation. *Ultrasound in Medicine and Biology*, **32** (1), 13-17.

[26] CHITTY, L. S., ALTMAN, D. G., HENDERSON, A. & CAMPBELL, S. (1994) Charts of Fetal Size: 3. Abdominal Measurements. *British Journal of Obstetrics and Gynaecology*, **101** (2), 125-131.

[27] CHUM, O. & ZISSERMAN, A. (2007) An Exemplar Model for Learning Object Classes. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[28] CICCHETTI, D. V. & FEINSTEIN, A. R. (1990) High Agreement but Low Kappa: Ii. Resolving the Paradoxes. *Journal of Clinical Epidemiology,* **43** (6), 551-558.

[29] COHEN, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* **20** (1), 37-46.

[30] CONSTANTINIDIS, I., MALKO, J. A., PETERMAN, S. B., LONG JR, R. C., EPSTEIN, C. M., BOOR, D., HOFFMAN JR, J. C., SHUTTER, L. & WEISSMAN, J. D. (1996) Evaluation of 1h Magnetic Resonance Spectroscopic Imaging as a Diagnostic Tool for the Lateralization of Epileptogenic Seizure Foci. *British Journal of Radiology,* **69** (817), 15-24.

[31] CREWSON, P. E. (2005) Reader Agreement Studies. *American Journal of Roentgenology,* **184** (5), 1391-1397.

[32] CROUSE, J. R., HARPOLD, G. H. & KAHL, F. R. (1986) Evaluation of a Scoring System for Extracranial Carotid Atherosclerosis Extent with B-Mode Ultrasound. *Stroke,* **17** (2), 270-275.

[33] DALAL, N. & TRIGGS, B. (2005) Histograms of Oriented Gradients for Human Detection. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005, 886-893.

[34] DARAHEM TEDESCO, G., DE SILVA BUSSAMRA, L. C., ARAUJO JR, E., SCHWACH WERNECK BRITTO, I., MARCONDES MACHADO NARDOZZA, L., FERNANDES MORON, A. & AOKI, T. (2009) Reference Range of Fetal Renal

Volume by Three-Dimensional Ultrasonography Using the Vocal Method. *Fetal Diagnosis and Therapy*, **25** (4), 385-391.

[35] DAS, D., MANSUR, A., KOBAYASHI, Y. & KUNO, Y. (2008) An Integrated Method for Multiple Object Detection and Localization. *In:* Proceedings of the International Symposium on Advances in Visual Computing (ISVC), 2008, 133-144.

[36] DEPRIEST, P. D., SHENSON, D., FRIED, A., HUNTER, J. E., ANDREWS, S. J., GALLION, H. H., PAVLIK, E. J., KRYSCIO, R. J. & VAN NAGELL JR, J. R. (1993) A Morphology Index Based on Sonographic Findings in Ovarian Cancer. *Gynecologic oncology*, **51** (1), 7-11.

[37] DOS SANTOS RIZZI, M. C., JÚNIOR, E. A., NARDOZZA, L. M. M., DINIZ, A. L. D., ROLO, L. C. & MORON, A. F. (2010) Nomogram of Fetal Liver Volume by Three-Dimensional Ultrasonography at 27 to 38 Weeks of Pregnancy Using a New Multiplanar Technique. *American Journal of Perinatology*, **27** (8), 641-647.

[38] DUDLEY, N. (2006) Re: Feasibility and Reproducibility of an Image-Scoring Method for Quality Control of Fetal Biometry in the Second Trimester [4]. *Ultrasound in Obstetrics and Gynecology*, **28** (3), 352.

[39] DUDLEY, N. J. & POTTER, R. (1993) Quality Assurance in Obstetric Ultrasound. *British Journal of Radiology*, **66** (790), 865-870.

[40] EL-NAQA, I., YANG, Y., WERNICK, M. N., GALATSANOS, N. P. & NISHIKAWA, R. M. (2002) A Support Vector Machine Approach for Detection of Microcalcifications. *IEEE Transactions on Medical Imaging*, **21** (12), 1552-1563.

[41] ERIKSSON, J. G., OSMOND, C., KAJANTIE, E., FORSÉN, T. J. & BARKER, D. J. P. (2006) Patterns of Growth among Children Who Later Develop Type 2 Diabetes or Its Risk Factors. *Diabetologia*, **49** (12), 2853-2858.

[42] EVANS, M. (2006). *Prenatal Diagnosis*, New York, United States, McGraw-Hill Professional.

[43] FEINSTEIN, A. R. & CICCHETTI, D. V. (1990) High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, **43** (6), 543-549.

[44] FELSBERG, M. & SOMMER, G. (2001) The Monogenic Signal. *IEEE Transactions on Signal Processing*, **49** (12), 3136-3144.

[45] FERRARI, V., FEVRIER, L., JURIE, F. & SCHMID, C. (2008) Groups of Adjacent Contour Segments for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30** (1), 36-51.

[46] FIELD, D. J. (1987) Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells. *Journal of the Optical Society of America. A, Optics and image science,* **4** (12), 2379-2394.

[47] FLEISCHER, A., ANYAEGBUNAM, A., GUIDETTI, D., RANDOLPH, G. & MERKATZ, I. R. (1992) A Persistent Clinical Problem: Profile of the Term Infant with Significant Respiratory Complications. *Obstetrics and Gynecology,* **79** (2), 185-190.

[48] FLEISCHER, A. C., ROMERO, R., MANNING, F. A., JEANTY, P. & JAMES, A. E. (1991). *The Principles and Practice of Ultrasonography in Obstetrics and Gynecology,* Norwalk, Connecticut, United States, Appleton & Lange.

[49] FLEISS, J. L. (1981). *Statistical Methods for Rates and Proportions,* New York, Wiley.

[50] FLEISS, J. L., COHEN, J. & EVERITT, B. S. (1969) Large Sample Standard Errors of Kappa and Weighted Kappa. *Psychological Bulletin,* **72** (5), 323-327.

[51] FREEMAN, W. T. & ADELSON, E. H. (1991) The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13** (9), 891-906.

[52] FREUND, Y. & SCHAPIRE, R. E. (1997) A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences,* **55** (1), 119-139.

[53] FRIES, N., ALTHUSER, M., FONTANGES, M., TALMANT, C., JOUK, P. S., TINDEL, M. & DUYME, M. (2007) Quality Control of an Image-Scoring Method for Nuchal Translucency Ultrasonography. *American Journal of Obstetrics and Gynecology,* **196** (3), 272.e1-272.e5.

[54] GABBE, S. G., NIEBYL, J. R. & SIMPSON, J. L. (2007). *Obstetrics: Normal and Problem Pregnancies,* Churchill Livingstone/Elsevier.

[55] GLOOR, J. M., BRECKLE, R. J. & GEHRKING, W. C. (1997) Fetal Renal Growth Evaluated by Prenatal Ultrasound Examination. *Mayo Clinic Proceedings,* **72** (2), 124-129.

[56] GOLDSTEIN, S. R. (1991) Embryonic Ultrasonographic Measurements: Crown-Rump Length Revisited. *American Journal of Obstetrics and Gynecology,* **165** (3), 497-501.

[57] GOODING, M. (2009) Unfog4d Register and Convert. 1.2.2 ed. University of Oxford.

[58] GOODING, M. J., RAJPOOT, K., MITCHELL, S., CHAMBERLAIN, P., KENNEDY, S. H. & NOBLE, J. A. (2010) Investigation into the Fusion of Multiple 4-D Fetal Echocardiography Images to Improve Image Quality. *Ultrasound in Medicine and Biology,* **36** (6), 957-966.

[59] GRANLUND, G. & KNUTSSON, H. (1995). *Signal Processing for Computer Vision,* Kluwer.

[60] GRAU, V., BECHER, H. & NOBLE, J. A. (2007) Registration of Multiview Real-Time 3-D Echocardiographic Sequences. *IEEE Transactions on Medical Imaging,* **26** (9), 1154-1165.

[61] GREENSPAN, H., BELONGIE, S., GOODMAN, R., PERONA, P., RAKSHIT, S. & ANDERSON, C. H. (1994) Overcomplete Steerable Pyramid Filters and Rotation Invariance. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1994, 222-228.

[62] HACIHALILOGLU, I., ABUGHARBIEH, R., HODGSON, A. J. & ROHLING, R. N. (2009) Bone Surface Localization in Ultrasound Using Image Phase-Based Features. *Ultrasound in Medicine and Biology,* **35** (9), 1475-1487.

[63] HADLOCK, F. P., DETER, R. L., HARRIST, R. B. & PARK, S. K. (1982) Fetal Abdominal Circumference as a Predictor of Menstrual Age. *American Journal of Roentgenology,* **139** (2), 367-370.

[64] HADLOCK, F. P., DETER, R. L., HARRIST, R. B. & PARK, S. K. (1984) Estimating Fetal Age: Computer-Assisted Analysis of Multiple Fetal Growth Parameters. *Radiology,* **152** (2), 497-501.

[65] HANNA, C. W. & YOUSSEF, A. B. M. (1997) Automated Measurements in Obstetric Ultrasound Images. *In:* IEEE International Conference on Image Processing (ICIP), 1997, 504-507.

[66] HARTMANN, D. P. (1977) Considerations in the Choice of Interobserver Reliability Estimates. *J Appl Behav Anal.,* **10** (1), 103-116.

[67] HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,*, Springer.

[68] HATA, T. & DETER, R. L. (1992) A Review of Fetal Organ Measurements Obtained with Ultrasound: Normal Growth. *Journal of Clinical Ultrasound,* **20** (3), 155-174.

[69] HERMAN, A., MAYMON, R., DREAZEN, E., CASPI, E., BUKOVSKY, I. & WEINRAUB, Z. (1998) Nuchal Translucency Audit: A Novel Image-Scoring Method. *Ultrasound in Obstetrics and Gynecology,* **12** (6), 398-403.

[70] HERMES, L. & BUHMANN, J. M. (2000) Feature Selection for Support Vector Machines. *In:* Proc. ICPR'00, 2000, 716-719.

[71] HSIEH, Y. Y., CHANG, C. C., LEE, C. C. & TSAI, H. D. (2000) Fetal Renal Volume Assessment by Three-Dimensional Ultrasonography. *American Journal of Obstetrics and Gynecology,* **182** (377-379.

[72] HUANG, W., CHAN, K. L., LI, H., LIM, J. H., LIU, J. & WONG, T. Y. (2010) Content-Based Medical Image Retrieval with Metric Learning Via Rank Correlation. *In:* Proceedings of the MICCAI Workshop on Machine Learning in Medical Imaging (MLMI), 2010, 18-25.

[73] INTERGROWTH 21ST (2008). *The International Fetal and Newborn Growth Consortium for the 21st Century Study Protocol,* Oxford University.

[74] IOANNOU, C., SARRIS, I., SALOMON, L. J. & PAPAGEORGHIOU, A. T. (2011) A Review of Fetal Volumetry: The Need for Standardization and Definitions in Measurement Methodology. *Ultrasound in Obstetrics and Gynecology,* **38** (6), 613-619.

[75] IOFFE, S. & FORSYTH, D. A. (2001) Probabilistic Methods for Finding People. *International Journal of Computer Vision,* **43** (1), 45-68.

[76] ISUKAPALLI, R., ELGAMMAL, A. & GREINER, R. (2006) Learning to Detect Objects of Many Classes Using Binary Classifiers. *In:* Proceedings of the European Conference on Computer Vision (ECCV), 2006, 352-364.

[77] JANSON, S. & VEGELIUS, J. (1979) On Generalizations of the G Index and the Phi Coefficient to Nominal Scales. *Multivariate Behavioral Research,* **14** (2), 255-269.

[78] JARDIM, S. M. G. V. B. & FIGUEIREDO, M. A. T. (2005) Segmentation of Fetal Ultrasound Images. *Ultrasound in Medicine and Biology,* **31** (2), 243-250.

[79] JARDIM, S. V. B. & FIGUEIREDO, M. A. T. (2003) Automatic Contour Estimation in Fetal Ultrasound Images. *In:* Proceedings of the IEEE International Conference on Image Processing (ICIP), 2003, 1065-1068.

[80] JIANG, J., ZHENG, S., TOGA, A. W. & TU, Z. (2008) Learning Based Coarse-to-Fine Image Registration. *In:* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[81] KARAVIDES, T., LEUNG, K. Y. E., PACLIK, P., HENDRIKS, E. A. & BOSCH, J. G. (2010) Database Guided Detection of Anatomical Landmark Points in 3d Images of the Heart. *In:* Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2010, 1089-1092.

[82] KOVESI, P. (1996) *Invariant Measures of Image Features from Phase Information.* PhD. Thesis, University of Western Australia.

[83] KOVESI, P. (1997) Symmetry and Asymmetry from Local Phase. *In:* Proceedings of the 10th Australian Join Conference on Artificial Intelligence, 1997, 185-190.

[84] KUNO, A., HAYASHI, Y., AKIYAMA, M., YAMASHIRO, C., TANAKA, H., YANAGIHARA, T. & HATA, T. (2002) Three-Dimensional Sonographic Measurement of Liver Volume in the Small-for-Gestational-Age Fetus. *Journal of Ultrasound in Medicine,* **21** (4), 361-366.

[85] LAMPERT, C. H., BLASCHKO, M. B. & HOFMANN, T. (2008) Beyond Sliding Windows: Object Localization by Efficient Subwindow Search. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[86] LANDIS, J. R. & KOCH, G. G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics,* **33** (1), 159-174.

[87] LAWN, J. E., COUSENS, S. & ZUPAN, J. (2005) 4 Million Neonatal Deaths: When? Where? Why? *Lancet,* **365** (9462), 891-900.

[88] LEUNG, K. Y. E., VAN STRALEN, M., NEMES, A., VOORMOLEN, M. M., VAN BURKEN, G., GELEIJNSE, M. L., TEN CATE, F. J., REIBER, J. H. C., DE JONG, N., VAN DER STEEN, A. F. W. & BOSCH, J. G. (2008) Sparse Registration for Three-Dimensional Stress Echocardiography. *IEEE Transactions on Medical Imaging,* **27** (11), 1568-1579.

[89] LOUGHNA, P., CHITTY, L., EVANS, T. & CHUDLEIGH, T. (2009) Fetal Size and Dating: Charts Recommended for Clinical Obstetric Practice. The British Medical Ultrasound Society (BMUS).

[90] LU, W., TAN, J. & FLOYD, R. (2005) Automated Fetal Head Detection and Measurement in Ultrasound Images by Iterative Randomized Hough Transform. *Ultrasound in Medicine and Biology,* **31** (7), 929-936.

[91] MADABHUSHI, A., FELDMAN, M. D., METAXAS, D. N., TOMASZEWESKI, J. & CHUTE, D. (2005) Automated Detection of Prostatic Adenocarcinoma from High-Resolution Ex Vivo Mri. *IEEE Transactions on Medical Imaging,* **24** (12), 1611-1625.

[92] MASMOUDI, H., HEWITT, S. M., PETRICK, N., MYERS, K. J. & GAVRIELIDES, M. A. (2009) Automated Quantitative Assessment of Her-2/Neu Immunohistochemical Expression in Breast Cancer. *IEEE Transactions on Medical Imaging,* **28** (6), 916-925.

[93] MASSELLI, G., BRUNELLI, R., DI TOLA, M., ANCESCHI, M. & GUALDI, G. (2011) Mr Imaging in the Evaluation of Placental Abruption: Correlation with Sonographic Findings. *Radiology,* **259** (1), 222-230.

[94] MATSOPOULOS, G. K. & MARSHALL, S. (1994) Use of Morphological Image Processing Techniques for the Measurement of a Fetal Head from Ultrasound Images. *Pattern Recognition,* **27** (10), 1317-1324.

[95] MAXWELL, A. E. (1977) Coefficients of Agreement between Observers and Their Interpretation. *The British Journal of Psychiatry,* **130** (1), 79-83.

[96] MCCULLOCH, W. S. & PITTS, W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics,* **5** (4), 115-133.

[97] MCINTIRE, D. D., BLOOM, S. L., CASEY, B. M. & LEVENO, K. J. (1999) Birth Weight in Relation to Morbidity and Mortality among Newborn Infants. *New England Journal of Medicine,* **340** (16), 1234-1238.

[98] MIRJALILI, S. A., MUIRHEAD, J. C. & STRINGER, M. D. (2012) Ultrasound Visualization of the Spinal Accessory Nerve in Vivo. *Journal of Surgical Research,* **175** (1), 11-16.

[99] MOGHADDAM, B., WEISS, Y. & AVIDAN, S. (2007) Fast Pixel/Part Selection with Sparse Eigenvectors. *In:* Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2007, 1-8.

[100] MORRA, J. H., TU, Z., APOSTOLOVA, L. G., GREEN, A. E., AVEDISSIAN, C., MADSEN, S. K., PARIKSHAK, N., HUA, X., TOGA, A. W., JACK JR, C. R., SCHUFF, N., WEINER, M. W. & THOMPSON, P. (2008) Mapping Hippocampal Degeneration in 400 Subjects with a Novel Automated Segmentation Approach. *In:* Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2008, 336-339.

[101] MULET-PARADA, M. & NOBLE, J. A. (2000) 2d+T Acoustic Boundary Detection in Echocardiography. *Medical Image Analysis,* **4** (21-30.

[102] NAMBURETE, A., RAHMATULLAH, B. & NOBLE, A. J. (2012) Nakagami-Based Choroid Plexus Detection in Fetal Ultrasound Images Using Adaboost. *In:* Medical Image Understanding and Analysis (MIUA), 2012.

[103] NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE (2008) Antenatal Care: Routine Care for the Healthy Pregnant Woman. *NICE clinical guideline 62.* London.

[104] NGUYEN, M. H. & DE LA TORRE, F. (2010) Optimal Feature Selection for Support Vector Machines. *Pattern Recognition,* **43** (3), 584-591.

[105] NGUYEN, T. D., KIM, S. H. & KIM, N. C. (2006) Surface Extraction Using Svm-Based Texture Classification for 3d Fetal Ultrasound Imaging. *In:* Proceedings of the International Conference on Communications and Electronics (HUT-ICCE), 2006, 285-290.

[106] NICOLAIDES, K. H., AZAR, G., BYRNE, D., MANSUR, C. & MARKS, K. (1992) Fetal Nuchal Translucency: Ultrasound Screening for Chromosomal Defects in First Trimester of Pregnancy. *British Medical Journal,* **304** (6831), 867-869.

[107] OCHS, R. A., GOLDIN, J. G., ABTIN, F., KIM, H. J., BROWN, K., BATRA, P., ROBACK, D., MCNITT-GRAY, M. F. & BROWN, M. S. (2007) Automated Classification of Lung Bronchovascular Anatomy in Ct Using Adaboost. *Medical Image Analysis,* **11** (3), 315-324.

[108] OPPENHEIM, A. V. & LIM, J. S. (1981) The Importance of Phase in Signals. *Proceedings of the IEEE,* **69** (5), 529-541.

[109] OREN, M., PAPAGEORGIOU, C., SINHA, P., OSUNA, E. & POGGIO, T. (1997) Pedestrian Detection Using Wavelet Templates. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1997, 193-199.

[110] OTT, W. J. (1988) The Diagnosis of Altered Fetal Growth. *Obstetrics and Gynecology Clinics of North America,* **15** (2), 237-63.

[111] PATHAK, S. D., CHALANA, V. & KIM, Y. (1997) Interactive Automatic Fetal Head Measurements from Ultrasound Images Using Multimedia Computer Technology. *Ultrasound in Medicine and Biology,* **23** (5), 665-673.

[112] PAVANI, S. K., DELGADO, D. & FRANGI, A. F. (2010) Haar-Like Features with Optimally Weighted Rectangles for Rapid Object Detection. *Pattern Recognition,* **43** (1), 160-172.

[113] PESCIA, D., PARAGIOS, N. & CHEMOUNY, S. (2008) Automatic Detection of Liver Tumors. *In:* Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2008, 672-675.

[114] PILU, G., REECE, E. A., GOLDSTEIN, I., HOBBINS, J. C. & BOVICELLI, L. (1989) Sonographic Evaluation of the Normal Developmental Anatomy of the Fetal Cerebral Ventricles: 2. The Atria. *Obstetrics and Gynecology,* **73** (2), 250-256.

[115] PRASAD, M. N., BROWN, M. S., AHMAD, S., ABTIN, F., ALLEN, J., DA COSTA, I., KIM, H. J., MCNITT-GRAY, M. F. & GOLDIN, J. G. (2008) Automatic Segmentation of Lung Parenchyma in the Presence of Diseases Based on Curvature of Ribs. *Academic radiology,* **15** (9), 1173-1180.

[116] PRENDERGAST, M., RAFFERTY, G. F., DAVENPORT, M., PERSICO, N., JANI, J., NICOLAIDES, K. & GREENOUGH, A. (2011) Three-Dimensional Ultrasound Fetal Lung Volumes and Infant Respiratory Outcome: A Prospective Observational Study. *BJOG: An International Journal of Obstetrics and Gynaecology*, **118** (608-614.

[117] QING, C., GEORGANAS, N. D. & PETRIU, E. M. (2007) Real-Time Vision-Based Hand Gesture Recognition Using Haar-Like Features. *In:* Proceedings of the IEEE Instrumentation and Measurement Technology (IMTC), 2007.

[118] RAHMATULLAH, B., PAPAGEORGHIOU, A. & NOBLE, J. A. (2011a) Automated Selection of Standardized Planes from Ultrasound Volume. *In:* Proceedings of the International MICCAI Workshop on Machine Learning in Medical Imaging (MLMI), 2011a.

[119] RAHMATULLAH, B., PAPAGEORGHIOU, A. & NOBLE, J. A. (2012a) Image Analysis Using Machine Learning: Anatomical Landmarks Detection in Fetal Ultrasound Image. *In:* IEEE Signature Conference on Computers, Software, and Applications (COMPSAC) *(In press)*, 2012a.

[120] RAHMATULLAH, B., PAPAGEORGHIOU, A. & NOBLE, J. A. (2012b) Integration of Local and Global Features for Anatomical Object Detection in Ultrasound. *In:* Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) *(In press)*, 2012b.

[121] RAHMATULLAH, B., PAPAGEORGHIOU, A. & NOBLE, J. A. (2012c) Multi-Scale Local Phase Features for Anatomical Object Detection in Fetal Ultrasound Image. *In:* Medical Image Understanding and Analysis (MIUA), 2012c.

[122] RAHMATULLAH, B., SARRIS, I., IOANNOU, C., KNIGHT, C., NOBLE, J. A. & PAPAGEORGHIOU, A. (2012d) Automated Fetal Biometry Image Landmark Detection for Confirming Correct Image Planes: Abdominal Circumference. *In:* World Congress on Ultrasound in Obstetrics and Gynecology *(In press)*, 2012d.

[123] RAHMATULLAH, B., SARRIS, I., NOBLE, J. A. & PAPAGEORGHIOU, A. (2011b) A Pilot Study of Automated Image Scoring for Quality Control Purposes in the Context of Multicentre Studies: Abdominal Circumference. *In:* World Congress on Ultrasound in Obstetrics and Gynecology, 2011b.

[124] RAHMATULLAH, B., SARRIS, I., NOBLE, J. A. & PAPAGEORGHIOU, A. (2012e) Automated Standard Plane Selection from Fetal Abdominal Ultrasound Volumes Using a Machine Learning Algorithm. *In:* World Congress on Ultrasound in Obstetrics and Gynecology *(In press)*, 2012e.

[125] RAHMATULLAH, B., SARRIS, I., PAPAGEORGHIOU, A. & NOBLE, J. A. (2011c) Quality Control of Fetal Ultrasound Images: Detection of Abdomen

Anatomical Landmarks Using Adaboost. *In:* 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2011c, 6-9.

[126] RAJPOOT, K., VICENTE, V. V. & NOBLE, J. A. (2009) Local-Phase Based 3d Boundary Detection Using Monogenic Signal and Its Application to Real-Time 3-D Echocardiography Images. *In:* Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2009, 783-786.

[127] REDDY, C. K. & BHUYAN, F. A. (2008) Retrieval and Ranking of Biomedical Images Using Boosted Haar Features. *In:* Proceedings of the IEEE International Conference on BioInformatics and BioEngineering (BIBE), 2008, 1-6.

[128] REINEHR, T., KLEBER, M. & TOSCHKE, M. (2009) Small for Gestational Age Status Is Associated with Metabolic Syndrome in Overweight Children. *European Journal of Endocrinology,* **160** (4), 579-584.

[129] RISNES, K. R., VATTEN, L. J., BAKER, J. L., JAMESON, K., SOVIO, U., KAJANTIE, E., OSLER, M., MORLEY, R., JOKELA, M., PAINTER, R. C., SUNDH, V., JACOBSEN, G. W., ERIKSSON, J. G., SØRENSEN, T. I. A. & BRACKEN, M. B. (2011) Birthweight and Mortality in Adulthood: A Systematic Review and Meta-Analysis. *International Journal of Epidemiology,* **40** (3), 647-661.

[130] ROBBINS, B. & OWENS, R. (1997) 2d Feature Detection Via Local Energy. *Image and Vision Computing,* **15** (5), 353-368.

[131] ROELFSEMA, N. M., HOP, W. C. J., BOITO, S. M. E. & WLADIMIROFF, J. W. (2004) Three-Dimensional Sonographic Measurement of Normal Fetal Brain Volume During the Second Half of Pregnancy. *American Journal of Obstetrics and Gynecology,* **190** (1), 275-280.

[132] ROTH, S., CHANG, T. C., ROBSON, S., SPENCER, J. A. D., WYATT, J. S. & STEWART, A. L. (1999) The Neurodevelopmental Outcome of Term Infants with Different Intrauterine Growth Characteristics. *Early Human Development,* **55** (1), 39-50.

[133] ROYAL COLLEGE OF OBSTETRICIANS AND GYNAECOLOGISTS (2000) Ultrasound Screening for Fetal Abnormalities. London: RCOG.

[134] ROYAL COLLEGE OF OBSTETRICIANS AND GYNAECOLOGISTS (2002) The Investigation and Management of the Small-for-Gestational-Age Fetus. *Guideline No. 31*.

[135] RUEDA, S., KNIGHT, C., PAPAGEORGHIOU, A. & NOBLE, J. A. (2011) Local Phase-Based Fuzzy Connectedness Segmentation of Ultrasound Images. *In:* Medical Image Understanding and Analysis (MIUA), 2011.

[136] RUTTEN, M. J., PISTORIUS, L. R., MULDER, E. J. H., STOUTENBEEK, P., DE VRIES, L. S. & VISSER, G. H. A. (2009) Fetal Cerebellar Volume and Symmetry on 3-D Ultrasound: Volume Measurement with Multiplanar and Vocal Techniques. *Ultrasound in Medicine and Biology,* **35** (8), 1284-1289.

[137] SABOGAL, J. C., BECKER, E., BEGA, G., KOMWILAISAK, R., BERGHELLA, V., WEINER, S. & TOLOSA, J. (2004) Reproducibility of Fetal Lung Volume Measurements with 3-Dimensional Ultrasonography. *Journal of Ultrasound in Medicine,* **23** (3), 347-352.

[138] SALOMON, L. J., BERNARD, J. P., DUYME, M., DORIS, B., MAS, N. & VILLE, Y. (2006) Feasibility and Reproducibility of an Image-Scoring Method for Quality Control of Fetal Biometry in the Second Trimester. *Ultrasound in Obstetrics and Gynecology,* **27** (1), 34-40.

[139] SALOMON, L. J. & VILLE, Y. (2005) Quality Control of Prenatal Ultrasound. *Ultrasound Review of Obstetrics and Gynecology,* **5** (4), 297-303.

[140] SARRIS, I., IOANNOU, C., CHAMBERLAIN, P., OHUMA, E., ROSEMAN, F., HOCH, L., ALTMAN, D. G. & PAPAGEORGHIOU, A. T. (2012) Intra- and Interobserver Variability in Fetal Ultrasound Measurements. *Ultrasound in Obstetrics and Gynecology,* **39** (3), 266-273.

[141] SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. & LEE, W. S. (1998) Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *Annals of Statistics,* **26** (5), 1651-1686.

[142] SCHNEIDERMAN, H. & KANADE, T. (2000) A Histogram-Based Method for Detection of Faces and Cars. *In:* Proceedings of the International Conference on Image Processing (ICIP), 2000, 504-507.

[143] SHEN, D., ZHAN, Y. & DAVATZIKOS, C. (2003) Segmentation of Prostate Boundaries from Ultrasound Images Using Statistical Shape Model. *IEEE Transactions on Medical Imaging,* **22** (4), 539-551.

[144] SHENTON, E. H. (1922) X Rays in Obstetric Practice. *The Lancet,* **199** (5148), 860-861.

[145] SHOTTON, J., WINN, J., ROTHER, C. & CRIMINISI, A. (2009) Textonboost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision,* **81** (1), 2-23.

[146] SIMONYAN, K., ZISSERMAN, A. & CRIMINISI, A. (2011) Immediate Structured Visual Search for Medical Images. *In:* Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2011, 288-296.

[147] SNIJDERS, R. J. M., THOM, E. A., ZACHARY, J. M., PLATT, L. D., GREENE, N., JACKSON, L. G., SABBAGHA, R. E., FILKINS, K., SILVER, R. K., HOGGE, W. A., GINSBERG, N. A., BEVERLY, S., MORGAN, P., BLUM, K., CHILIS, P., HILL, L. M., HECKER, J. & WAPNER, R. J. (2002) First-Trimester Trisomy Screening: Nuchal Translucency Measurement Training and Quality Assurance to Correct and Unify Technique. *Ultrasound in Obstetrics and Gynecology,* **19** (4), 353-359.

[148] SOUKA, A. P. & NICOLAIDES, K. H. (1997) Diagnosis of Fetal Abnormalities at the 10-14-Week Scan. *Ultrasound in Obstetrics and Gynecology,* **10** (6), 429-442.

[149] SUN, Y., KAMEL, M. S. & WANG, Y. (2006) Boosting for Learning Multiple Classes with Imbalances Class Distribution. *In:* Proceedings of the International Conference on Data Mining (ICDM), 2006, 592-602.

[150] TESTA, A. C., AJOSSA, S., FERRANDINA, G., FRUSCELLA, E., LUDOVISI, M., MALAGGESE, M., SCAMBIA, G., MELIS, G. B. & GUERRIERO, S. (2005) Does Quantitative Analysis of Three-Dimensional Power Doppler Angiography Have a Role in the Diagnosis of Malignant Pelvic Solid Tumors? A Preliminary Study. *Ultrasound in Obstetrics and Gynecology,* **26** (1), 67-72.

[151] THOMAS, J. G., PETERS, R. A. & JEANTY, P. (1991) Automatic Segmentation of Ultrasound Images Using Morphological Operators. *IEEE Transactions on Medical Imaging,* **10** (2), 180-186.

[152] TIEU, K. & VIOLA, P. (2004) Boosting Image Retrieval. *International Journal of Computer Vision,* **56** (1-2), 17-36.

[153] TOLOUEE, A., ABRISHAMI MOGHADDAM, H., FOROUZANFAR, M., GITY, M. & GARNAVI, R. (2011) Image Based Diagnostic Aid System for Interstitial Lung Diseases. *Expert Systems with Applications,* **38** (6), 7755-7765.

[154] TORRALBA, A., MURPHY, K. P. & FREEMAN, W. T. (2007) Sharing Visual Features for Multiclass and Multiview Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **29** (5), 854-869.

[155] TURK, M. A. & PENTLAND, A. P. (1991) Face Recognition Using Eigenfaces. *In:* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1991, 586-591.

[156] TZANAKIS, N. E., EFSTATHIOU, S. P., DANULIDIS, K., RALLIS, G. E., TSIOULOS, D. I., CHATZIVASILIOU, A., PEROS, G. & NIKITEAS, N. I. (2005) A New Approach to Accurate Diagnosis of Acute Appendicitis. *World Journal of Surgery,* **29** (9), 1151-1156.

[157] UNICEF (2004) State of the World's Children 2005. New York.

[158] UNICEF & WHO (2004) Low Birthweight: Country, Regional and Global Estimates. New York.

[159] VAN LIMBERGEN, E., VAN DER SCHUEREN, E. & VAN TONGELEN, K. (1989) Cosmetic Evaluation of Breast Conserving Treatment for Mammary Cancer. 1. Proposal of a Quantitative Scoring System. *Radiotherapy and Oncology,* **16** (3), 159-167.

[160] VAPNIK, V. (1995). *The Nature of Statistical Learning Theory*, New York, United States, Springer-Verlag.

[161] VILLE, Y. (2008) 'Ceci N'est Pas Une ÉChographie': A Plea for Quality Assessment in Prenatal Ultrasound. *Ultrasound in Obstetrics and Gynecology,* **31** (1), 1-5.

[162] VIOLA, P. & JONES, M. J. (2004) Robust Real-Time Face Detection. *International Journal of Computer Vision,* **57** (2), 137-154.

[163] WANG, L., LEE, S. L., MERRIFIELD, R. & YANG, G. Z. (2011) Subject-Specific Cardiac Segmentation Based on Reinforcement Learning with Shape Instantiation. *In:* Proceedings of the International MICCAI Workshop on Machine Learning in Medical Imaging (MLMI), 2011, 300-307.

[164] WANG, S., YAO, J., PETRICK, N. & SUMMERS, R. M. (2010) Combining Statistical and Geometric Features for Colonic Polyp Detection in Ctc Based on Multiple Kernel Learning. *International Journal of Computational Intelligence and Applications,* **9** (1), 1-15.

[165] WARSOF, S. L., COOPER, D. J., LITTLE, D. & CAMPBELL, S. (1986) Routine Ultrasound Screening for Antenatal Detection of Intrauterine Growth Retardation. *Obstetrics and Gynecology,* **67** (1), 33-39.

[166] WEI, L., YANG, Y., NISHIKAWA, R. M. & JIANG, Y. (2005) A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications. *IEEE Transactions on Medical Imaging,* **24** (2), 371-380.

[167] WEIWEI, Z., BRADY, J. M., HARALD, B. & NOBLE, J. A. (2011) Spatio-Temporal (2d+T) Non-Rigid Registration of Real-Time 3d Echocardiography and Cardiovascular Mr Image Sequences. *Physics in Medicine and Biology,* **56** (5), 1341.

[168] WHITEHILL, J., LITTLEWORT, G., FASEL, I., BARTLETT, M. & MOVELLAN, J. (2009) Toward Practical Smile Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **31** (11), 2106-2111.

[169] WILCOX, A. (1983) Intrauterine Growth Retardation: Beyond Birthweight Criteria. *Early Human Development,* **8** (189-193.

[170] WU, G., QI, F. & SHEN, D. (2006) Learning-Based Deformable Registration of Mr Brain Images. *IEEE Transactions on Medical Imaging,* **25** (9), 1145-1157.

[171] YAQUB, M., IOANNOU, C., PAPAGEORGHIOU, A., JAVAID, K., COOPER, C. & NOBLE, J. (2010a) Improving Boundary Definition for 3d Ultrasound Quantification of Fetal Femur. *In:* Medical Image Understanding and Analysis (MIUA), 2010a.

[172] YAQUB, M., JAVAID, M. K., COOPER, C. & NOBLE, J. A. (2011) Improving the Classification Accuracy of the Classic Rf Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation. *In:* Proceedings of the International MICCAI Workshop on Machine Learning in Medical Imaging (MLMI), 2011, 184-192.

[173] YAQUB, M., MAHON, P., JAVAID, K., COOPER, C. & NOBLE, J. (2010b) Weighted Voting in 3d Random Forest Segmentation. *In:* Medical Image Understanding and Analysis (MIUA), 2010b.

[174] YAQUB, M., NAPOLITANO, R., IOANNOU, C., PAPAGEORGHIOU, A. & NOBLE, A. (2012) Automatic Detection of Local Fetal Brain Structures in Ultrasound Images. *In:* Proceedings of the International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2012, 1555-1558.

[175] YU, C. H., CHANG, C. H., CHANG, F. M., KO, H. C. & CHEN, H. Y. (2000) Fetal Renal Volume in Normal Gestation: A Three-Dimensional Ultrasound Study. *Ultrasound in Medicine and Biology,* **26** (1253-1256.

[176] YU, J., WANG, Y. & CHEN, P. (2008a) Fetal Ultrasound Image Segmentation System and Its Use in Fetal Weight Estimation. *Medical and Biological Engineering and Computing,* **46** (12), 1227-1237.

[177] YU, J., WANG, Y., CHEN, P. & SHEN, Y. (2008b) Fetal Abdominal Contour Extraction and Measurement in Ultrasound Images. *Ultrasound in Medicine and Biology,* **34** (2), 169-182.

[178] YU, J. H., WANG, Y. Y., CHEN, P. & XU, H. Y. (2007) Two-Dimensional Fuzzy Clustering for Ultrasound Image Segmentation. *In:* International Conference on Bioinformatics and Biomedical Engineering (ICBBE), 2007, 599-603.

[179] YUANZHONG, L., HARA, S. & SHIMURA, K. (2006) A Machine Learning Approach for Locating Boundaries of Liver Tumors in Ct Images. *In:* Proceedings of the International Conference on Pattern Recognition (ICPR), 2006, 400-403.

[180] ZADOR, I. E., SALARI, V., CHIK, L. & SOKOL, R. J. (1991) Ultrasound Measurement of the Fetal Head: Computer Versus Operator. *Ultrasound in Obstetric and Gynaecology,* **1** (3), 208-211.

[181] ZHOU, J., CHANG, S., METAXAS, D. & AXEL, L. (2007) Vascular Structure Segmentation and Bifurcation Detection. *In:* Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2007, 872-875.

[182] ZWICK, R. (1988) Another Look at Interrater Agreement. *Psychological Bulletin,* **103** (3), 374-378.