Automatic Measurements of Femoral Characteristics using 3D Ultrasound Images in Utero

Mohammad Yaqub

Wolfson College

Supervised by Prof. J. Alison Noble, Dr. Kassim Javaid and Prof. Cyrus Cooper

Hilary Term 2011

Degree of Doctor of Philosophy in Biomedical Engineering

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences

University of Oxford

Abstract

Vitamin D is very important for endochondral ossification and it is commonly insufficient during pregnancy (Javaid et al., 2006). Insufficiency of vitamin D during pregnancy predicts bone mass and hence predicts adult osteoporosis (Javaid et al., 2006). The relationship between maternal vitamin D and manually measured fetal biometry has been studied (Mahon et al., 2009). However, manual fetal biometry especially volumetric measurements are subjective, time-consuming and possibly irreproducible. Computerised measurements can overcome or at least reduce such problems. This thesis concerns the development and evaluation of novel methods to do this.

This thesis makes three contributions. Firstly, we have developed a novel technique based on the Random Forests (RF) classifier to segment and measure several fetal femoral characteristics from 3D ultrasound volumes automatically. We propose a feature selection step in the training stage to eliminate irrelevant features and utilise the "good" ones. We also develop a weighted voting mechanism to weight tree probabilistic decisions in the RF classifier. We show that the new RF classifier is more accurate than the classic method (Yaqub et al., 2010b, Yaqub et al., 2011b). We achieved 83% segmentation precision using the proposed technique compared to manually segmented volumes. The proposed segmentation technique was also validated on segmenting adult brain structures in MR images and it showed excellent accuracy.

The second contribution is a wavelet-based image fusion technique to enhance the quality of the fetal femur and to compensate for missing information in one volume due to signal attenuation and acoustic shadowing. We show that using image fusion to increase the image quality of ultrasound images of bony structures leads to a more accurate and reproducible assessment and measurement qualitatively and quantitatively (Yaqub et al., 2010a, Yaqub et al., 2011a).

The third contribution concerns the analysis of data from a cohort study of 450 fetal femoral ultrasound volumes (18-21 week gestation). The femur length, cross-sectional areas, volume, splaying indices and angles were automatically measured using the RF method. The relationship between these measurements and the fetal gestational age and maternal vitamin D was investigated. Segmentation of a fetal femur is fast (2.3s/volume), thanks to the parallel implementation. The femur volume, length, splaying index were found to significantly correlate with fetal gestational age. Furthermore, significant correlations between the automatic measurements and 10 nmol increment in maternal 25OHD during second trimester were found.

Acknowledgement

I would like to thank my supervisors Prof. J. Alison Noble, Dr. Kassim Javaid and Prof. Cyrus Cooper for the help, support and guidance that was given to me. Thanks to your endurance through this project and the patience in giving me. Their diverse knowledge helped me in all directions to accomplish this work. I would like to acknowledge and thank the NIHR BRU for the funding they provided to support this research project. I also would like to thank Dr. Pam Mahon, Dr. Aris Papageorghiou, Dr. Christos Ioannou for their collaboration and contribution in this work. Many thanks to Wolfson College, the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Institute of Biomedical Engineering/Engineering Science, the Nuffield Department of Obstetrics and Gynaecology and many other departments at the University of Oxford for the facilities, tools, training courses, etc, they provided. I also would like to thank all my colleagues in the BioMedIA laboratory for support and kindness.

I would like to thank my parents for the great support and patience they gave me. I would like to thank my best friend Majdi for the unlimited support and help he provided me.

Finally, countless thanks and love to my wife Laila and my little son Zaid. Laila was with me step by step and without her support I would not have produced this thesis.

Thank you all.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Obstetricians, radiologists, neurosurgeons and many other types of clinicians rely on different medical image modalities to diagnose diseases and assess patients health. Manual measurement of structures in medical images is used in common practice to assist this. For instance, obstetricians measure fetal biometry (head circumference, femur length, abdominal circumference, etc) to assess fetal growth and look for abnormalities. While manual measurements from medical images have been used extensively, they are subjective, often time-consuming and potentially poorly reproducible. With the recent acquisition of high dimensional medical images and the increase interest to perform complex measurements on these images, these disadvantages increase. Although the need for automated measurement tools is not questioned, it is very difficult to provide effective and reliable solutions.

This doctoral thesis studies the fetal bone in 3D ultrasound images looking at 1) automated measurement of the fetal femoral characteristics, 2) enhancing the quality of 3D ultrasound images of the fetal femur and 3) assessing the effect of maternal vitamin D status on automated complex fetal femoral biometry.

1.1 Rationale

In the UK, osteoporotic hip fractures 1) cost approximately £2.3 billion per year (Kanis et al., 2000), 2) is associated with 30% mortality at one year and 3) is associated with a loss of independence, e.g., walking and dressing. Therefore, many clinical studies have investigated this problem by looking at different factors which might affect bone growth (Kanis et al., 2000, Dennison et al., 2001, Cooper et al., 2002, Javaid and Cooper, 2002, Prince and Glendenning, 2004, Cooper et al., 2006, Javaid et al., 2006, Mahon, 2007, Mahon et al., 2009). Maternal vitamin D is very important for fetal endochondral ossification required for femoral growth. Vitamin D is commonly insufficient in mothers during pregnancy and it predicts bone mass in later childhood (Javaid et al., 2006). In the assessment of fetal bone growth (Mahon, 2007, Mahon et al., 2009), maternal vitamin D was related to manual fetal femoral biometry. As previously mentioned, manual measurement suffers from a few problems which we target to overcome using automated tools. The main goal of this thesis is to investigate the relationship between fetal and maternal characteristics and fetal femoral characteristics measured automatically to assess fetal bone growth and predicts abnormalities. To achieve this goal, computerised techniques are required to assess the fetal femur by finding several femoral characteristics. Although the main goal of the thesis is purely clinical, a key contribution is to look to develop novel techniques to tackle this problem. Introducing new ideas should hopefully help solve some issues and push this field forward in this application as well as others.

1.2 Contributions

The main contributions of this thesis are three folds:

1. The development of a novel technique to segment and measure fetal femoral characteristics from 3D ultrasound images automatically, accurately, quickly and precisely.
2. The investigation of the relationship between the measured characteristics and fetal and maternal characteristics, e.g., fetal gestational age and maternal vitamin D.
3. The development of a 3D ultrasound image quality enhancement technique based on image fusion to improve fetal femoral definition.

An thorough validation on a reasonable number of cases is used in each case to support the proposed new techniques.

1.3 Thesis Outline

This thesis is organised as follows. Chapter 2 provides a literature review that covers both the clinical and technical sides. From the clinical perspective we describe ultrasound imaging in Section 2.1, bone biology in Section 2.2 and related clinical studies in Section 2.3. A brief description of image segmentation with the focus of medical imaging is shown in Section 2.4. Section 2.5 reviews the classic technique which we build our segmentation method upon (Random Forests). A brief description on medical image fusion is shown in Section 2.6. The method, experimental setup, results and discussion of the proposed segmentation technique are illustrated in detail in Chapter 3. The proposed image fusion technique to enhance fetal femoral definition is presented in Chapter 4. Clinical comparisons with fetal and maternal characteristics is discussed in Chapter 5. Finally, Chapter 6 concludes the whole thesis and discusses a few directions of future work.

Chapter 2

Literature Review

This chapter reviews the literature relevant to this thesis. The review covers both clinical and technical aspects. Section 2.1 provides a brief overview of ultrasound imaging while fetal bone growth is discussed in Section 2.2. We then summarise two clinical studies which motivated this work in Section 2.3. The technical review, Section 2.4, starts with image segmentation focusing on published techniques that were influential to our work. We discuss the segmentation technique we build our work on in Section 2.5. We finally describe image fusion and its motivation especially in medical imaging applications in Section 2.6.

2.1 Ultrasound Imaging

Ultrasound imaging (or sonography) is a relatively inexpensive, fast and radiation-free imaging modality. Ultrasound is excellent for non-invasively imaging and diagnosing a number of medical conditions. It can also be done in real-time which is important for some tasks. The use of ultrasound imaging to assess fetal growth was one of the early applications of ultrasound. Ultrasound is used not only in fetal imaging but also in different domains including assessing heart disease, breast cancer diagnosis, visualising

muscles, tendons, and many internal organs, which helps calculate the size and highlight the structure and any pathological lesions with real-time tomographic images. Ultrasound can be used to image different human structures like the heart, kidneys, liver, pancreas, gall bladder, eye, thyroid, blood vessels. Ultrasound can also be used in image-guided interventions like neurosurgery (King et al., 2010) and to guide breast (Mallapragada et al., 2008) and prostate (Xu et al., 2010) biopsies .

Although ultrasound imaging has several advantages over other image modalities like MRI and CT, it suffers from many problems. For example, ultrasound suffers from acoustic shadowing, attenuation, motion effects and speckle. In addition, anatomical definition is poor when compared with MRI or CT. Due to this, some organs are very hard to be visualised and others cannot be seen. Moreover, the quality of the acquired images highly depends on the sonographer experience. Acquisition reproducibility is a major problem in ultrasound imaging in which intra or inter-sonographer acquisition may highly affect the quality and the details in the acquired images. The problem becomes harder when using ultrasound for imaging fetuses because of the movement of the fetus and differences in maternal fat thickness and volume of amniotic acid. Furthermore, ultrasound images of bones (of interest to this thesis) are highly affected by signal attenuation and acoustic shadowing. This mainly happens because of the calcification of bone which makes most of the ultrasound beam bounce back. Figure 2.1 shows a screen shot of the femur bone for a 34 weeks fetus. The figure also shows a few challenges in ultrasound images. On the other hand, ultrasound imaging is radiation free, inexpensive and real-time, therefore it is still the main imaging modality during pregnancy. Although MRI is safe and can be used during pregnancy, it is not approved

for use below 20 week gestation. Therefore, it cannot be used in early pregnancy and is not cost-effective or practical for abnormality screening.

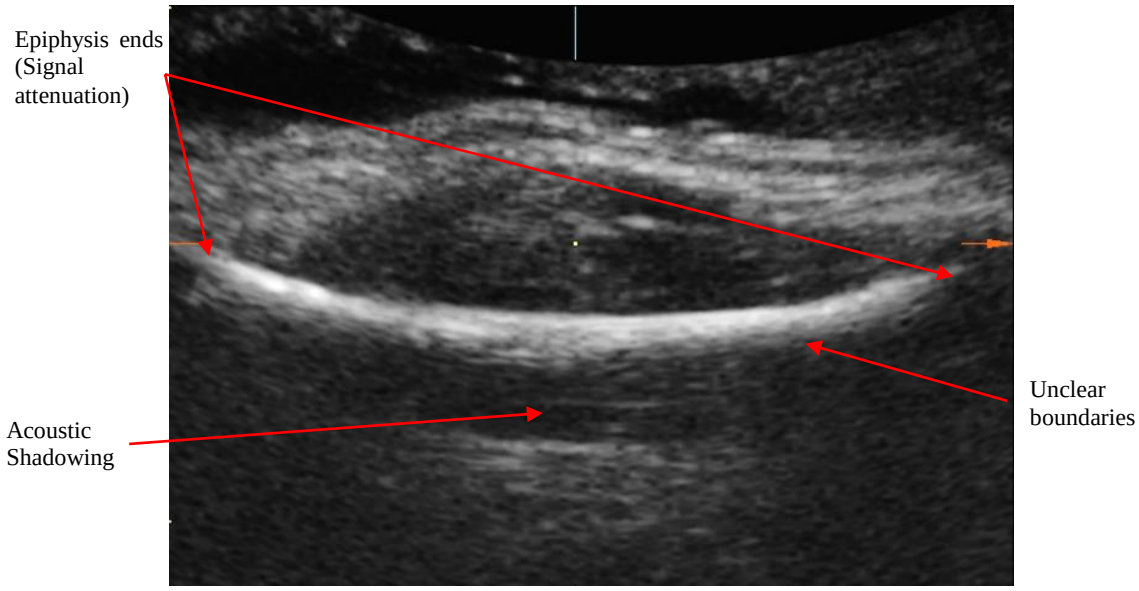


Figure 2.1. Ultrasound image of the fetal femur. The femur is represented by the white linear segment within the thigh tissues. The figure shows a few problems like poor anatomical definition, shadowing from the bottom part of the femur, loss of signal at the two epiphysis ends and low signal to noise ratio (Mahon, 2007).

2.1.1 Disadvantages of 2D Ultrasonography

Although 2D ultrasound is currently the most dominant image modality used during pregnancy, the use of 3D ultrasound is expected to supersede 2D ultrasound in the future (Prager et al., 2010). We already mentioned the main advantages of ultrasound imaging in general. However, 2D ultrasound imaging has a few specific issues. 2D ultrasound imaging suffers from the subjectivity of the conventional exam because the resulting image quality is dependent on the experience and knowledge of the sonographer. Furthermore, it is difficult to localize a thin 2D ultrasound image plane in the organ, and difficult to reproduce a particular region-of-interest at a later time,

making the conventional 2D exam poor for quantitative prospective or follow-up studies. Further, the patient's anatomy or orientation sometimes restricts the imaging angle, resulting in inaccessibility of the optimal 2D image plane necessary for diagnosis (Fenster and Downey, 1996, Prager et al., 2010). 3D ultrasound imaging aims to generate a 3D representation of structures which in turn helps overcome most of 2D ultrasound scanning problems.

2.1.2 3D Ultrasound Acquisition

Prager et al. (2010) described the three common ways to acquire a 3D ultrasound volume. The first group is the free-hand acquisition in which the operator holds an assembly composed of the probe attachment, and manipulates it in the usual manner over the anatomy to be viewed. Although the free-hand 3D scanning approach offers great flexibility, problems of noise and scanning gaps may reduce the image quality, particularly when imaging small structures at high resolution. The second group of acquisition strategy requires sweeping a 1D transducer array via a mechanical device. In this acquisition type, the probe is fixed and 2D ultrasound images are acquired at predefined spatial intervals so that the imaging sequence samples the volume of interest properly, without missing any regions. The 1D transducer produces 2D B-scan images and known locations in 3D space. The last and most challenging image acquisition is by using the matrix array probe. The idea is to have multiple 2D array elements which can fire signal simultaneously. It was first introduced by Volumetrics (Volumetrics Inc., USA) in the 1990's and later, after significant advances in image quality, released by Philips in 2003 in echocardiography. Image acquisition using matrix array has only recently been introduced in obstetrics. Philips healthcare has recently launched a new

matrix array probe (X6-1) for obstetrics use. For more information of 3D ultrasound imaging please see (Fenster and Downey, 1996, Prager et al., 2010).

2.2 Fetal Bone Biology Overview

Endochondral ossification is the process associated with the development of fetal bone (ASBMR, 2006, Cooper et al., 2006, Wikipedia, 2009). In this process, cartilage is replaced by bone. Ossification and the formation of bone happen in the long bones such as the femur and humerus while intra-membranous ossification is the process used to make flat bones such as the mandible and flat bones of the skull. In this doctoral thesis, we are mainly interested in formation of long bones especially the femur. The femur starts to appear histologically in the mid-shaft at 7 to 8 weeks. About two weeks later, it starts to appear radiologically. The diaphysis of the bone is usually completely ossified at birth, but the epiphyses are still cartilaginous and ossification continues in the secondary ossification centre (ASBMR, 2006, Wikipedia, 2009). Figure 2.2 shows the main steps in fetal bone growth.

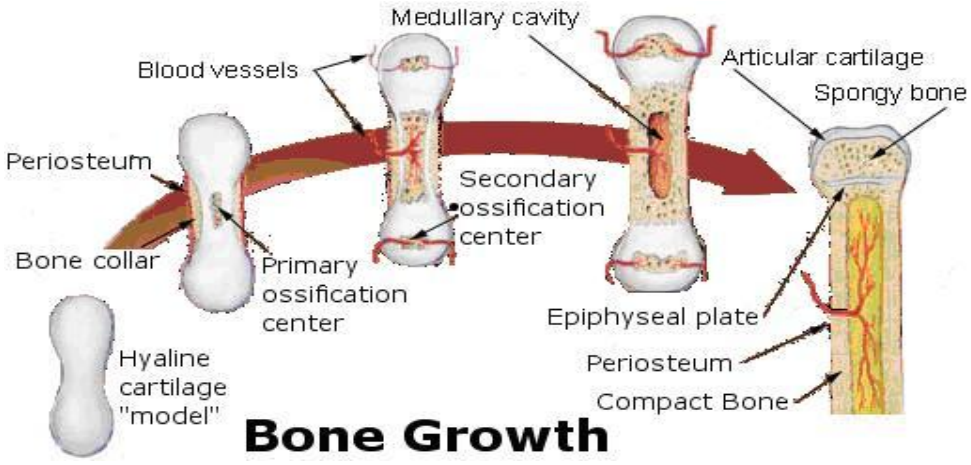


Figure 2.2. The process of Endochondral ossification (Wikipedia, 2009).

Vitamin D is a fat-soluble vitamin and it is one of the body's fundamental vitamins needed for survival. It is mainly needed for bone mineralisation which refers to the process of bone formation. Deficiency in vitamin D may lead to different types of diseases like rickets, osteomalacia and osteoporosis (Rauch and Schoenau, 2002, Prince and Glendenning, 2004, ASBMR, 2006). Several studies have investigated maternal vitamin D status and its effect on bone growth in fetuses and children (Javaid, 2005, Javaid et al., 2006, Gale et al., 2008). In an interesting work by Mahon (2007), the assessment of ultrasound on fetal skeletal development has been investigated. That study is described in more detail in Section 2.3.1.

2.3 Clinical Studies

This doctoral research considered analysis of data from three clinical studies. The Southampton Women's Survey (SWS) data and the Internet Brain Segmentation Repository (IBSR)[1] were used to validate the segmentation techniques developed in Chapter 3. The image fusion work presented in Chapter 4 was validated on data from the Intergrowth-21st repository (Intergrowth-21st Project). We describe the SWS and Intergrowth-21st datasets here but we delay the description of IBSR dataset until Section 3.3.1. This is because it is not an ultrasound dataset and is not an urgent clinical problem to this thesis but it has been used to support our technical contribution.

2.3.1 Ultrasound Assessment Of Fetal Musculo-Skeletal Development (SWS)

A recent clinical study by Mahon (2007) aimed to correlate maternal characteristics to manual measurements of the fetal femur in three Dimensional (3D) ultrasound images. The main goal was to assess how maternal characteristics affect fetal bone growth. 3D

[1] http://www.cma.mgh.harvard.edu/ibsr/

ultrasound scans which capture the fetal thigh and surrounding tissues were used in this study. In the SWS data, 19 week scans capture more information about tissues than the 34 week ones (see Figure 2.3). This is mainly because the femur is small though the sonographer can bound the scanning region by putting a virtual box to acquire a specific structure only (unfortunately, this had not been done when the SWS data was acquired). Scans were acquired on KretzGE Voluson 730® system (Kretztechnik AG, Zipf, Austria) and by a single sonographer. 4–8 MHz 3D transducer was used to acquire ultrasound volumes. For further details regarding this study ethical approval see Appendix A.

In this study, 517 fetuses were scanned at 19 and 34 week gestation. Although the scans were termed 19 and 34 week scans, a window of gestational ages was accepted. This window is [18-21] week scans for the 19 week and [33-36] week scans for the 34 week. Neonatal Dual-energy X-ray Absorptiometry (DXA) scans were acquired within two weeks after birth. Manual measurements on the uppermost fetal thigh were made on ultrasound volumes of these fetuses. The measurements were correlated with maternal body composition and vitamin D status, to determine if maternal characteristics have an effect on fetal musculo-skeletal development. For mothers who have insufficient vitamin D, a splaying effect of the distal femoral metaphysic appeared on 10% of the fetuses. The splaying index was calculated as a ratio between femoral distal cross-sectional area and femoral length, see equation (2.1). Figure 2.4 shows the splaying effect on DXA and ultrasound images while Figure 2.5 shows schematically main measurements performed in this study to quantify the splaying index.

$$\text{Splaying index} = \frac{\text{Distal cross sectional area}}{\text{Femur length}}. \tag{2.1}$$

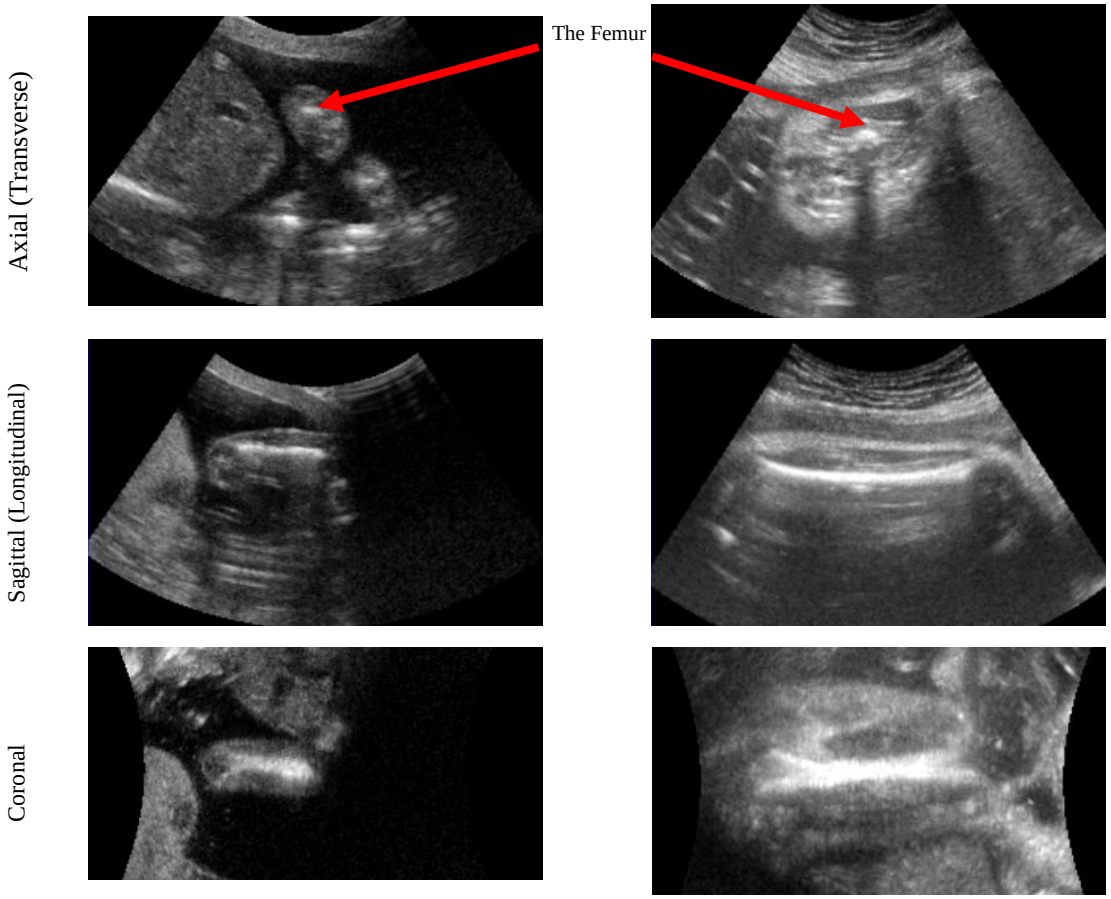
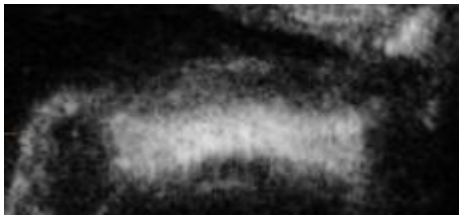


Figure 2.3. 19 (left column) and 34 (right column) weeks 3D ultrasound views. Top row represents the axial (transverse) plane. Mid row represents the sagittal (longitudinal) plane. Bottom row represents the coronal plane.

(a) DXA image for a femur of new born baby



(b) Ultrasound image for 19 weeks fetus

Figure 2.4. Splaying effect from the distal end of the femur (right side of each image). (a) DXA, (b) ultrasound (Mahon, 2007) with permission.

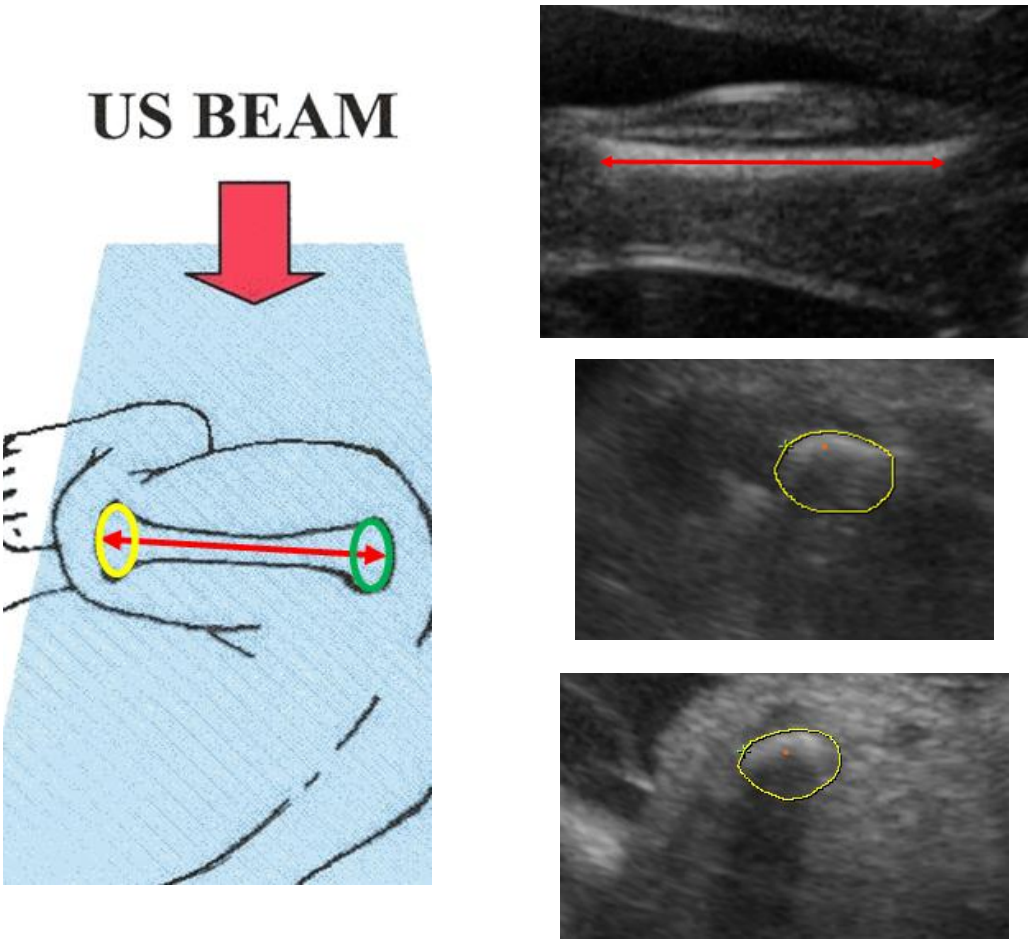


Figure 2.5. The diagram on the left shows three main measurements manually performed on the SWS ultrasound data. Femur length is the *red* arrowed straight line. From the figure on the left side, the *green* elliptical like shape is the proximal CSA while the *yellow* one is the distal CSA. The three images on the right show femur length on a longitudinal plane (top), proximal CSA on a transverse plane (mid), and distal CSA on a transverse plane (bottom). Parts of this Figure is taken from (Mahon, 2007) with permission.

2.3.1.1 Manual Measurements

In Mahon (2007), several manual fetal femoral bone measurements were performed on the 3D ultrasound volumes; for instance, line, area and volume measurements were performed on the fetal femur and thigh. Since our work is interested in quantifying and analysing the fetal femur, we focused on the following measurements 1) Femur Length (FL), 2) Femur Proximal Cross-Sectional Area (Proximal CSA), 3) Femur Distal Cross-

Sectional Area (Distal CSA), and 4) Femur Volume (FV). See Figure 2.5 for a description of these measurements.

2.3.1.2 Results and Conclusions

In Mahon (2007) and a subsequent paper (Mahon et al., 2009), it was found that at 19 week lower maternal vitamin D levels and increased splaying index (splaying index was calculated as distal or proximal CSA divided by femur length) were correlated. This suggests that the splaying index is a useful tool to detect fetuses likely to be at risk from vitamin D insufficiency.

2.3.2 Intergrowth-21st

The main goal of this study is to develop new prescriptive standards to assess normal fetal growth and the nutritional status of newborn infants. The study is performed on eight geographically diverse populations. For more information see (Intergrowth-21st Project)[2] and for ethical approval see Appendix A. We used some of the 3D ultrasound volumes in our studies. The data from this project was used in Chapter 4.

2.4 Image Segmentation

The main goal of this thesis is to develop algorithms and tools to automatically quantify femoral characteristics in ultrasound images and correlate these measurements to maternal characteristics. Image segmentation is the first step toward this goal. Image segmentation is the process of partitioning foreground objects from background in an image (Shapiro and Stockman, 2001). In our case, the goal is to segment the fetal femur quickly and accurately to measure several biometric characteristics of it in order to study their relation to maternal body composition and vitamin D level. In this section,

[2] http://www.medscinet.net/intergrowth/

we briefly review the relevant medical image segmentation literature. We also survey the main techniques developed in fetal ultrasound image segmentation in Section 2.4.2.

2.4.1 Introduction

In general, image segmentation can be performed in 1) two dimensional space (2D image), e.g., natural images; 2) in three dimensional space (3D image), e.g., Magnetic Resonance Imaging (MRI); and 3) in four dimensional space (3D + time), e.g., 3D+t echocardiography. Image segmentation has been used in a variety of practical applications in different domains. In medical imaging, image segmentation is an indispensable tool for clinicians and doctors for diagnosis and treatment. Another major area of applications is biometrics, e.g., fingerprint, face recognition, iris recognition, DNA, etc. Image segmentation is also important for instance in satellite imaging, automatic traffic control systems and many others. For more information see (Haralick, 1983, Pham et al., 2000, Shapiro and Stockman, 2001, Yu Jin, 2001, Noble and Boukerroui, 2006).

Before talking about automatic and semi-automatic segmentation techniques, we consider manual segmentation and measurement in order to appreciate why there is a need for automatic and semi-automatic segmentation. Manual segmentation and measurement are usually done through a mouse-like device for instance an interactive monitor e.g., WACOM CINTIQ 21UX. Although human experience is included during segmentation and measurement, they are usually highly subjective and reproducibility is very critical. In addition, manual processing and delineation is tedious and can take more than 30 minutes per dataset/image especially on 3D or 4D cases depending on the structure size and complexity. In clinical practice, sonographers and radiologist can suffer from Repetitive Stress Injury (RSI) due to the high number of keystrokes required

for segmentation. Therefore, automatic or semi-automatic segmentation techniques help overcome many of these problems.

There are many segmentation techniques developed in the literature but the focus here are on medical image segmentation especially in ultrasound. Many segmentation techniques applied to medical problems are actually adapted from the domain of natural image segmentation. To name a few main categories, histogram-based models, edge-based models, region-based models, deformable models, graph partitioning models, learning-based models, etc (Haralick, 1983, Pham et al., 2000, Shapiro and Stockman, 2001, Noble and Boukerroui, 2006).

Because of the properties of ultrasound data, general segmentation techniques vary in accuracy when applied to ultrasound data. In addition, depending on the scanned structures and on the clinical application, segmentation techniques provide different results on ultrasound data. Since the target in this work is to segment fetal femur in 3D ultrasound, a few segmentation techniques might fail to do the job. For example, because the edges of the femur on the ultrasound image are unclear, edge-based techniques are not able to delineate its border accurately. Moreover, in a few slices, femur structures are connected to other similar tissues, therefore region growing-based techniques might also fail to provide accurate segmentation. Level set, active contour and other semi-automatic techniques require initialisation. Moreover, they typically rely on edge information. Deformable models may also get "stuck" at local minima which require re-initialisation (Carneiro et al., 2008). More generally, in the fields of computer vision and machine learning there has recently been a great interest in the problem of accurate and robust detection and segmentation of visual classes. Learning-based techniques have been intensively used in different domains but they are still relatively

new in medical image analysis. Example of these techniques are Adaboost (Freund and Schapire, 1997), Probabilistic Boosting Tree (Tu, 2005), Random Forests (Breiman, 2001), etc. The Random Forests technique is reviewed in Section 2.5.

Adaboost is an adaptive strong classifier that is built by combining several weak classifiers sequentially (Freund and Schapire, 1997). Although this strong classifier provides accurate results in several medical segmentation problems (Shimizu et al., 2008, Lupascu et al., 2010), it is slow to train and test and it is not easily extendable for multi-class classification. The Probabilistic Boosting Tree (PBT) (Tu, 2005) makes use of Adaboost algorithm and build a tree in which each node is a strong classifier constructed using Adaboost or similar boosting algorithm. Although Adaboost asymptotically converges to the target distribution, it may need to pick hundreds of weak classifiers (Tu, 2005). On the other hand, PBT creates a strong classifier on each node in a tree where the number of week classifiers in each node can be small which makes the creation of a node faster while maintaining good classification accuracy through the use of multiple nodes in a tree. In addition, hierarchal classification provides an easier framework for multi-class classification and it can be implemented in parallel. Finally, Random Forests technique has shown to be comparable accuracy to the pre-mentioned classifiers while being much faster (Breiman, 2001, Caruana and Niculescu-Mizil, 2006). In addition, Random Forests can be used in classification as well as in regression (Breiman et al., 1984, Criminisi et al., 2010). Moreover, Random Forests is inherently multi-class and highly parallelisable which makes it more suitable for real-time applications.

2.4.2 Fetal Ultrasound Image Segmentation

Ultrasound image segmentation is of clinical interest but unfortunately very hard to do accurately and robustly. Noble and Boukerroui (2006) have nicely reviewed the main techniques used in ultrasound segmentation although fetal applications were not covered in depth. Fetal ultrasound image segmentation is very important to assess the growth of the fetus and measurements from the body of the fetus are used to estimate fetal birth weight, height, growth normality and other factors. As indicated in Table 2.1, research groups have developed techniques to segment and measure different regions of interest in the fetal body. For example, a few groups have focused on fetal structures segmentation and detection (Zador et al., 1991, Matsopoulos and Marshall, 1994, Chalana et al., 1996, Hanna and Youssef, 1997, Pathak et al., 1997, Jardim and Figueiredo, 2003, Archip et al., 2005, Jardim and Figueiredo, 2005, Lu et al., 2005, Liu et al., 2006, Carneiro et al., 2007, Carneiro et al., 2008), the femur (Thomas et al., 1991, Jardim and Figueiredo, 2005, Carneiro et al., 2007, Carneiro et al., 2008), the abdominal area (Carneiro et al., 2007, Carneiro et al., 2008, Yu et al., 2008) and other fetal structures (Liu et al., 2006, Carneiro et al., 2007, Carneiro et al., 2008). Work on automatic segmentation of fetal bony structures using ultrasound has only been performed in 2D. Only a few groups (Zador et al., 1991, Chalana et al., 1996, Pathak et al., 1997, Archip et al., 2005, Lu et al., 2005, Carneiro et al., 2007, Carneiro et al., 2008, Yu et al., 2008) have investigated the inter-observer variability. None of the research groups that have targeted fetal ultrasound segmentation has studied intra-observer variability. Thus compared to other areas of ultrasound segmentation, it is relatively under-developed.

Automatic measurement of the fetal femur in ultrasound has not been well-studied. To our knowledge, only three research groups have tried to measure femur length. In 1991, Thomas et al., (1991) developed an automatic segmentation technique based on morphological operators to measure femur length in fetal 2D ultrasound images. After more than a decade, Jardim and Figueiredo, (2003, 2005) developed a method to segment the femur and the head from fetal 2D ultrasound images. The method is based on the evolution of a parametric deformable shape. Recently, Carneiro et al., (2007, 2008) have developed an automatic detection and measurement method of different fetal structures in 2D ultrasound images. They deal with the detection problem as a learning problem. The main idea is to automatically distinguish between the appearance of the object of interest and background by training a constrained probabilistic boosting tree classifier (Tu, 2005). Note that the three mentioned approaches are used to measure fetal femur length only. To our knowledge, we are the first to investigate automatic fetal femoral volume segmentation and femoral shape characterisation. Table 2.1 summarises the techniques developed to segment and measure fetal femur and the ones developed to measure Head Circumference (HC), BiParietal Diameter (BPD), Abdominal Circumference (AC) and a few others. The techniques are listed historically with the femur segmentation ones highlighted.

Table 2.1. Historical listing of the work done in fetal ultrasound segmentation. The table shows the following columns: reference, region of interest, segmentation technique(s), the validation strategy, the error(s), if the work has investigated intra-observer variability and inter-observer variability and if it is semi or fully automatic, respectively.

Reference	ROI	Seg. Tech.	Validation			Error	Intra obs.	Inter obs. (# raters)	auto
			Data set	# cases	Type				
(Thomas et al., 1991)	Femur (Length)	Morphological operators	Real	24	Linear regression with manual measurement	0.8133 (MSE)	N	N	Y
(Zador et al., 1991)	Head (HC, BPD, OFD)	Thresholding and Hough transform	Real	75	Mean difference Manual vs. automatic	-0.36 ± 9.87(HC) 1.87 ± 1.94mm (BPD) 2.82 ± 4.13mm (OFD)	N	Y (3)	Y
(Matsopoulos and Marshall, 1994)	Head (BPD)	Morphological operators	Real	NA	Manual vs. automatic	0.1050%	N	N	Y
(Hanna and Youssef, 1997)	Head (HC & BPD)	Morphological operators & Hough transform	Real	NA	Correlation bet. manual & automatic	0.994 (BPD) 0.985 (HC)	N	N	Y
(Subramanian et al., 1997)	Face & heart	Region growing & split/merge	Real	6 data sets 30 to 50 images each	Qualitative	NA	N	N	N
(Chalana et al., 1996, Pathak et al., 1997)	Head (HC & BPD)	Active contour model	Real	35	Mean absolute difference	1.4% (BPD) 2.9% (HC)	N	Y (6)	N
(Zayed et al., 2001)	Whole fetus	Wavelet & fuzzy c-means	Real	10	Qualitative	NA	N	N	Y
(Jardim and Figueiredo, 2003, Jardim and Figueiredo, 2005)	Femur (contour) Head (HC)	Parametric deformable shape/ML estimation	Synthetic & real	50	Qualitative	NA	N	N	N
(Archip et al., 2005)	Whole fetus	NCut & spectral clustering (supervised & unsuper.)	Synthetic & real	5	Hausdorff & mean	Sup. (mm): 3.68(H) 1.88 (mean) Unsup. (mm): 3.56(H) 1.76 (mean)	N	Y (2)	Y
(Lu et al., 2005)	Head (HC & BPD)	K-means & Iterative Hough transform	Real	217	Mean signed & abs. diff.	0.12% (BPD) -0.52% (HC)	N	Y (8)	N

Reference	ROI	Seg. Tech.	Validation			Error	Intra obs.	Inter obs. (# raters)	auto
			Data set	# cases	Type				
(Liu et al., 2006)	Whole fetus	Morphological operators & fuzzy c-means	Real	NA	Qualitative	NA	N	N	Y
(Yu et al., 2008)	Abdominal (circumference)	Edge det., fuzzy c-means, Hough transform & snake	Synthetic & real	150 (real)	Manual vs. automatic (mean abs. Diff.) manual vs. automatic (accuracy in GA estimation)	See reference	N	Y (9)	Y
(Anquez et al., 2008)	Whole fetus	Statistical prior/Level-set	Synthetic & real	4 3D datasets but only one was quantitatively analyzed	Manual vs. automatic classification error	72%	N	N	Y
(Carneiro et al., 2007, Carneiro et al., 2008)	Different fetus organs	Constrained probabilistic boosting tree	Real	Large database of expert annotated 2D images	Hausdorff dist. Average dist. Hadock reg. (GA) Pearson corr.	See references	N	Y (5)	Y

2.5 Random Forests

Random Forests is a discriminative classifier/regressor which consists of several decision trees that are constructed in a random manner (Figure 2.6). The technique has become popular in the computer vision and machine learning communities. Although the work by Breiman (Breiman, 2001) was the first to introduce the name of Random Forests, Amit and Geman were the first to propose the idea of randomised trees (Amit and Geman, 1997) in the machine learning community. In the literature, Random Forests is also referred to as Randomised Trees, Random Decision Forests and other names. Random Forests mainly gained interest in the computer vision domain after the work of Lepetit et al. (2006). A straightforward extension from 2D to 3D to segment medical images was first published in 2009 by Lempitsky et al. (2009). In image

segmentation domain, Random Forests is mainly used as a supervised classification model but it can be used in regression (Breiman et al., 1984) and as an unsupervised learner (Shotton et al., 2008).

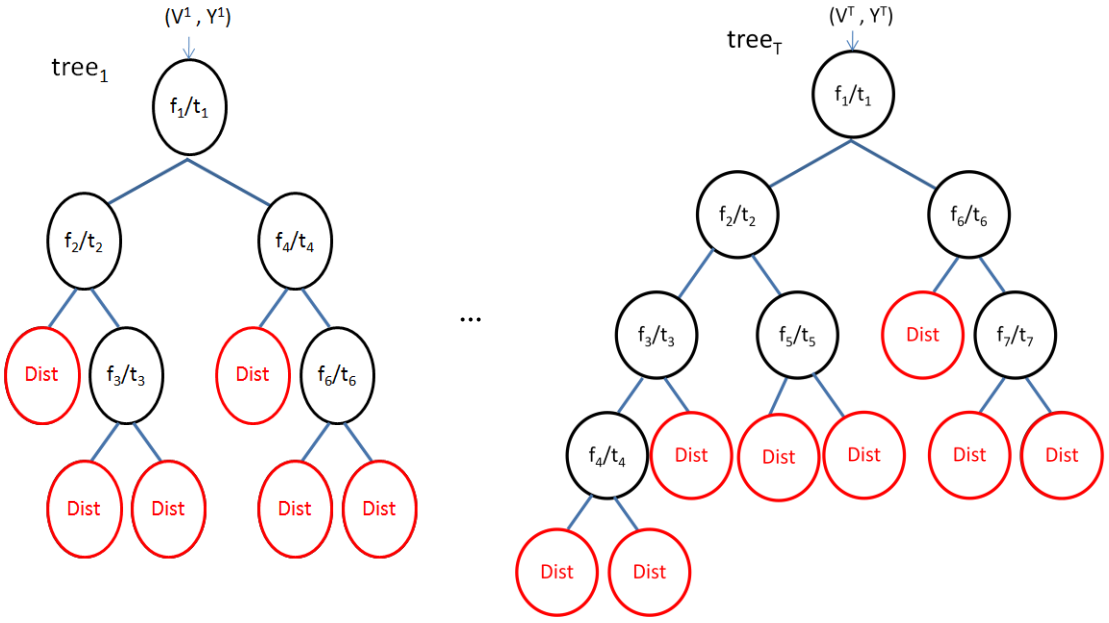


Figure 2.6. Random Forests of *T* trees. Only two trees are show for convenience. Black nodes are split nodes while red nodes are decision leaves (distributions).

Over-fitting and the lack of generalisation are the main problems in decision tree classifiers therefore the Random Forests classifier alleviates these issues by 1) creating tree nodes in a random manner and 2) combining multiple decisions from many trees to produce the final decision. The Random Forests technique has been shown to produce a more accurate classification than conventional decision trees (Pei et al., 2007), and comparable accuracy to Support Vector Machines (SVM) (Bosch et al., 2007, Statnikov and Aliferis, 2007), boosting and other classification techniques (Caruana and Niculescu-Mizil, 2006). On the other hand, the Random Forests technique has several advantages over other classification techniques like SVM and somehow boosting. Generally, the Random Forests approach,

- is computationally efficient in classification;
- generates probabilistic decision (soft decision);
- is inherently a multi-class classifier;
- combines several types of features and handles a huge feature space efficiently;
- rather overcomes over-fitting[3] due to the randomness in feature selection and bagging[4] during training;
- is easily extendable to parallel implementation due to the independence in tree creation;
- can be used for clustering (unsupervised learning).

Although the Random Forests technique has proven its applicability in several classification problems, it still suffers from several problems, particularly

- over-fitting in some datasets especially when the data is highly imbalanced[5], in noisy classification problems or if the training data is small,
- incapability to intelligently handle large numbers of irrelevant features in some classification problems especially when dealing with high dimensional data like medical images (this is investigated in Chapter 3),
- the final decision is usually biased toward weaker trees (this is also studied in Chapter 3).

[3] Over-fitting usually happens when the model is not generalisable because it is fitted to work well on the training data but not on the testing data.
[4] The concept of bagging is illustrated in Section 2.5.2.
[5] The idea of imbalanced trees is illustrated in Section 3.3.2. For more information see (Chen et al., 2004, Thomas et al., 2006).

2.5.1 Random Forests for Medical Image Segmentation

The use of the Random Forests technique for medical image segmentation and detection is relatively new. To our knowledge, four main technical publications have been published in the field of medical image segmentation (Lempitsky et al., 2009, Yi et al., 2009, Geremia et al., 2010, Yaqub et al., 2010b) and two in the field of medical image detection (Criminisi et al., 2009, Criminisi et al., 2010)[6].

Lempitsky et al. (2009) used the standard binary Random Forests to automatically delineate the myocardium in 3D ultrasound of adult hearts. Rectangular and position features were used to build 20 trees. They used 30 candidate features at each tree node which is relatively small. This may significantly affect the segmentation accuracy by decreasing the chance of randomly choosing good features. Yi et al. (2009) segmented three brain tissues from 20 MRI volumes using the traditional Random Forests technique after bias field correction. Intensities, gradients and positions features were utilised. Their method showed relatively good results compared with other techniques but implementation details were vague. Geremia et al. (2010) used the conventional Random Forests technique to segment multi-channel brain MR images for analysis of multiple sclerosis. Intensities and rectangular features were used to build 30 trees. They used 950 candidate features to try at each tree node during training which is the largest compared to other published work. In the case of object detection in medical imaging, Criminisi et al. (Criminisi et al., 2009, Criminisi et al., 2010) used the standard Random Forests technique to automatically detect several organs in CT volumes by finding a 3D bounding box around each organ.

[6] This is before 1/1/2011

2.5.2 Training the Random Forests Classifier

Training a Random Forests classifier proceeds by building several random decision trees. A decision tree is basically a tree where each node is a decision stump. In other words, each node contains a question (a classifier) and according to the answer of this question a split happens to either left or right. Leaf nodes usually contain an outcome of answering a series of questions. The Random Forests technique builds several decision trees in a random nature. Figure 2.7 describes the training stage. The grey area in the figure is important to compare with the proposed technique (developed in Chapter 3).

There are two sources of randomness in the training stage. One is the selection of training examples (Bagging (Breiman, 1996)) for each tree and the other in the feature selection to create a node in any tree. Bagging or Bootstrap aggregating generates several new training sets from the main training set by randomly selecting instances from the main set with replacement. Breiman discusses these two random processes in detail (Breiman, 1996, Breiman, 2001). From the pseudo-code in Figure 2.7 notice that training of each tree is performed recursively on a bootstrapped set from the original training set *V*.

Input:

- Images + Ground Truth (GT is segmented images)
- Parameters:
  - T: number of trees
  - n': number of features to randomly select and test at each tree node where n'<<n and n is the number of features in the feature pool
  - L: max tree depth
  - e: minimum information gain.

Output:

T trees where every node in the tree is a (weak) classifier (feature + threshold). Leaf nodes are probabilistic decisions of how much that leaf to belong to $c_1, c_2 \ldots, c_{Labels}$.

From the images and GT get the training examples V:((v_1, y_1), (v_2, y_2)... (v_K,y_K)) where v_i is the i^{th} voxel and y_i is its class label: $y_i \in \{c_1, c_2 \ldots, c_{Labels}\}$

Create T groups from the training examples V where each group has similar number of training examples as V. Creation is performed by randomly selecting instances from V with replacement (Bagging (Breiman, 1996)). The results are V^1, V^2, ... V^T where V^i is a vector containging K training examples and is used to train the i^{th} tree.

- For t = 1 to T
  TrainDecisionTree(V^t,n',L,e)

Funtion TrainDecisionTree(V, n', L, e)

(1) If MaxTreeDepth(L) Then
- Create a distribution on V (Leaf node) and Exit

End If

(2) Best_Classifier = null

(3) For i = 1 to n'
- Select a random feature f_i without replacement
- Find the best threshold th for f_i over V
- If Best_classifier is worse than (f_i,th) Then
  Set Best_classifier=(f_i, th)
  End If

End For

(4) Classify V on the chosen Best_classifer (split V to V_{left} & V_{right})

(5) If $\Delta E(V_{left}, V_{right}) < e$ Then
- Create a distribution on V (Leaf node)
- Exit

End If

(6) Create a node on the tree that contains the Best_classifer then split to left and right

(7) TrainDecisionTree(V_{left},n',L,e)

(8) TrainDecisionTree(V_{right},n',L,e)

End TrainDecisionTree

Figure 2.7. Pseudo-code for the Random Forests training stage.

Figure 2.8 gives a schematic presentation of node creation in a tree within the Random Forests framework. To create a node in a tree, *n'* features are randomly selected from the feature pool. For each feature, a "good" threshold is found to define a classifier. The classifier that maximises the expected gain of information is chosen (Lepetit and Fua, 2006, Shotton et al., 2008). Node creation is represented by the grey area in Figure 2.7. The expected gain of information (ΔE) can be measured as,

$$\Delta E = -\frac{|V_{left}|}{|V|} E(V_{left}) - \frac{|V_{right}|}{|V|} E(V_{right}), \tag{2.2}$$

$$E(v) = -\sum_{c=1}^{Labels} \frac{|V : class = c|}{|V|} \log_2 \left(\frac{|V : class = c|}{|V|} \right). \tag{2.3}$$

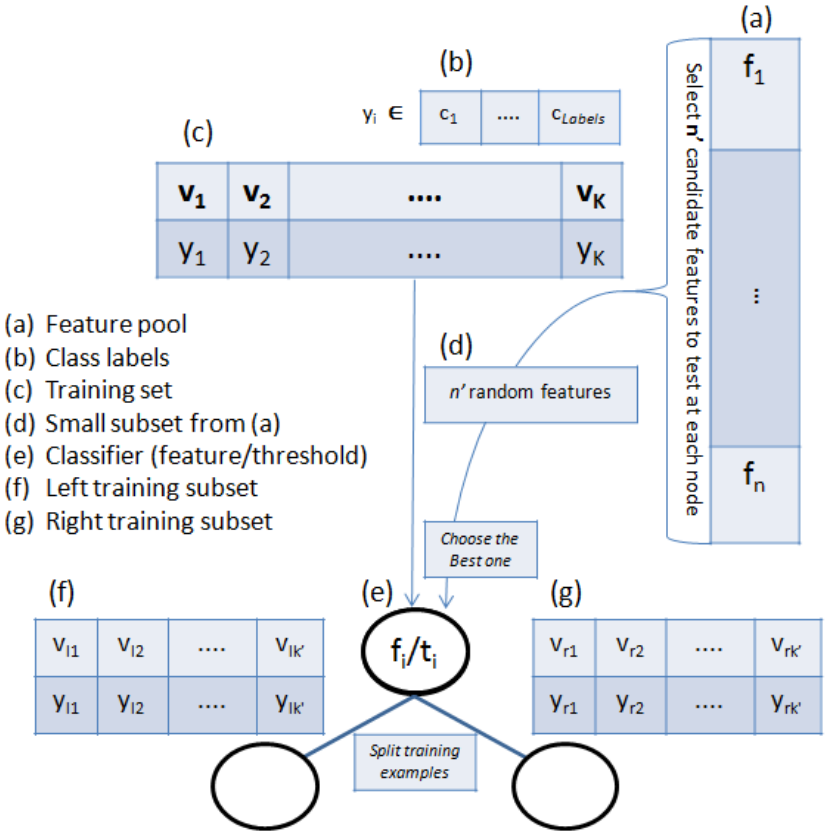


Figure 2.8. Detailed description of creating a node in Random Forests.

Here V_{left} and V_{right} are left and right subsets of the training set V after a split in a tree node happens. For instance, Suppose that we have 100 training examples where 40 are from class c_1 and 60 from c_2. If a classifier splits the 100 examples to 30 (originally $20 \in c_1$, $10 \in c_2$) and 70 (originally $20 \in c_1$, $50 \in c_2$) then ΔE is -0.783. If another classifier splits the 100 training examples to 45 (originally $35 \in c_1$, $10 \in c_2$) and 55 (originally $5 \in c_1$, $50 \in c_2$) then ΔE is -0.597. This implies that the latter classifier is better to classify the 100 examples. This example is illustrated in Figure 2.9.

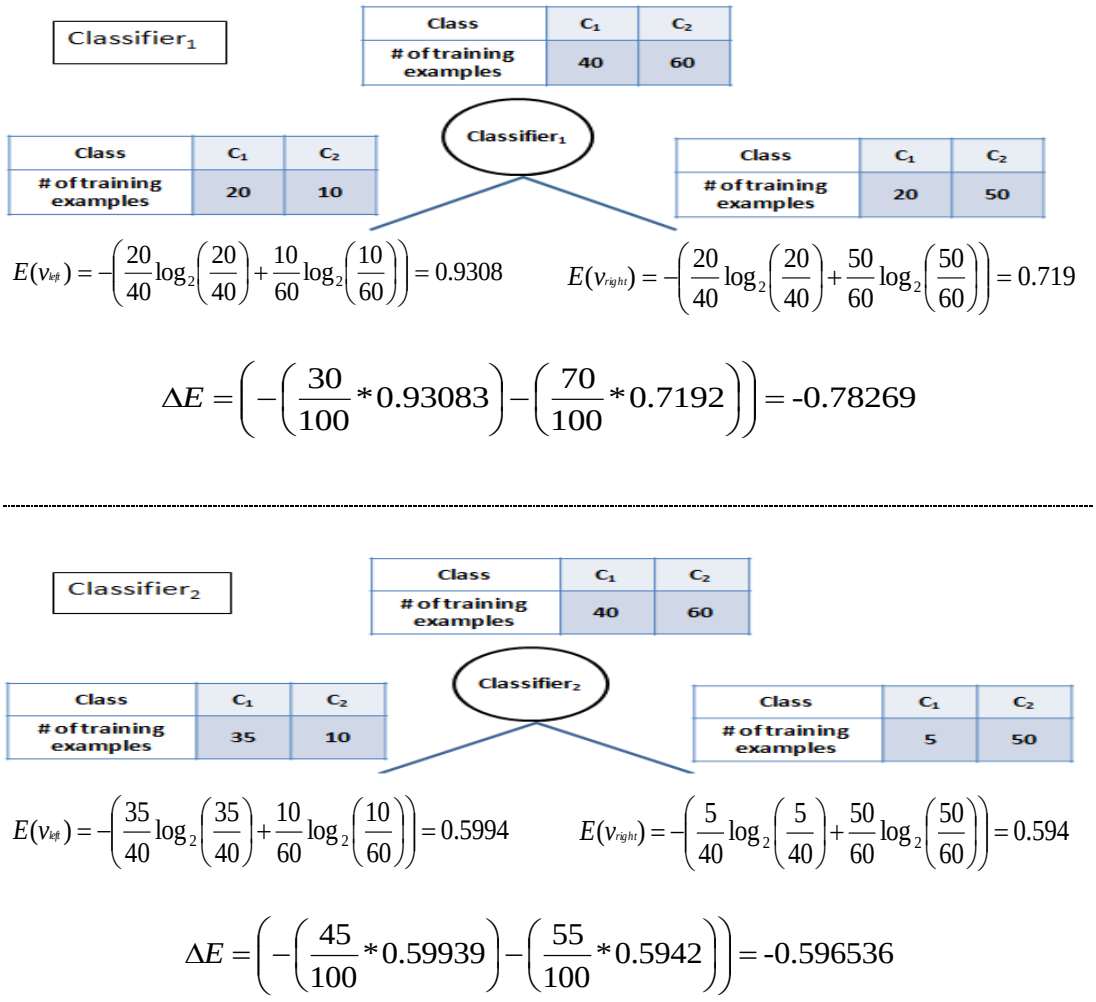


Figure 2.9. An example of Shannon Entropy *E(v)* and the gain of information *ΔE* caused by two classifiers on an arbitrary example. The example contains 100 training points where 40 points belong to one class and 60 to the other class.

For each feature the threshold with the maximum ΔE should be found. In reality this is a time consuming process since many candidate thresholds have to be tested to find the best one. Therefore, faster ways are desirable. For example, a small number of thresholds can be tested. The user may need to specify the number of thresholds to try for each feature. These thresholds can be randomly, normally or uniformly distributed (Breiman, 2001, Viola and Jones, 2004, Lepetit and Fua, 2006).

The recursive creation of a tree is depicted in Figure 2.10. The recursive creation of a tree stops if either 1) the maximum tree depth is reached or 2) no gain of information is achieved if a node splits to left or right (ΔE<e, see equation (2.2)). No gain of information at a node happens either because a small number of training examples reached it or almost all training examples belong to one class. If one of the two stopping conditions holds a leaf node is created. The leaf node contains a distribution of the training examples that reached it for every class, see equation (2.4). For instance, in a three class classification problem (c_1, c_2 and c_3), if 200 training examples reach a leaf node and if 40 training examples belong to class c_1 and 140 belong to class c_2 and 20 belong to class c_3 then this leaf node should contain a distribution of {40,140,20}. Finally, this distribution is used during the testing stage to classify unseen examples.

$$Dist(v(leafNode)) = Hist\left(\forall c_j \in \{c_1, \ldots c_{Labels}\} \ \left| v \in c_j \right|\right), \qquad (2.4)$$

here *v(leafNode)* are the training examples *v* reached that leaf node. $\left| v \in c_j \right|$ is the number of training examples which have class label c_j. *Hist* is the histogram where its bins are the class labels ($c1, \ldots, c_{Labels}$) and each is represented by the frequency of the training examples for each label that reached that leaf.

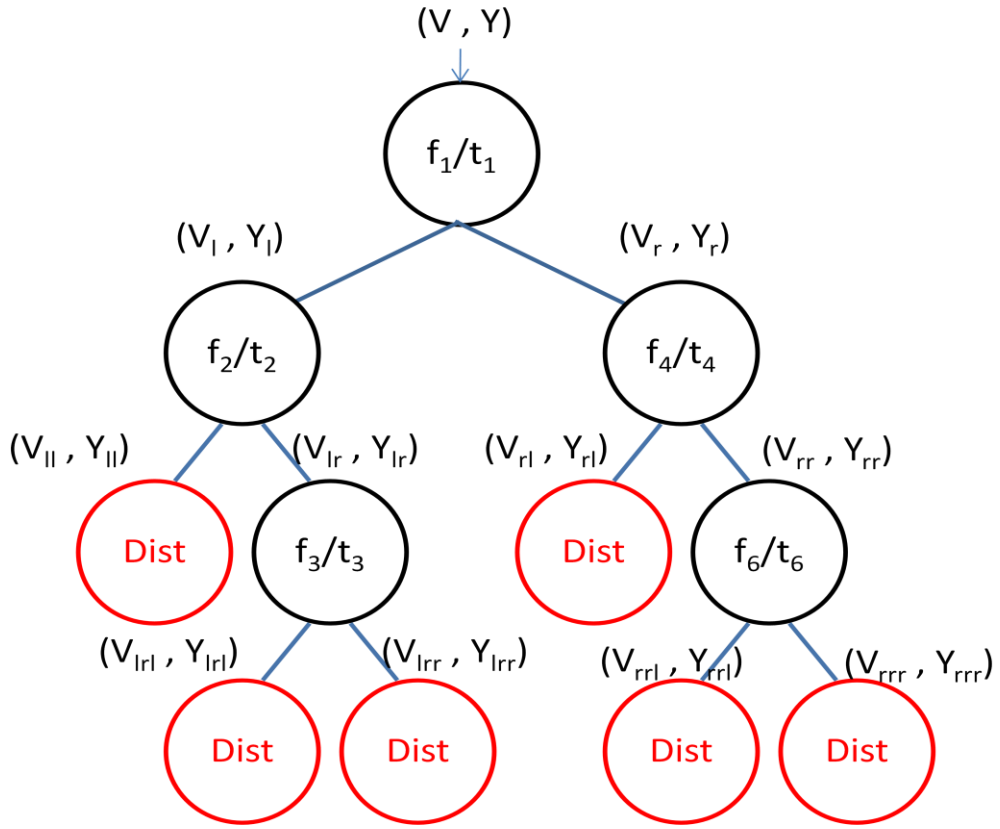


Figure 2.10. Training a tree in a Random Forests framework. Training examples *V* and their class labels *Y* keep splitting until they reach a leaf node (no more gain is achieved if a split happens) or the maximum tree depth is reached.

After training is done, *T* trees are created. Each decision node in any tree contains a classifier which is represented by a feature and its threshold. Each leaf node contains a probabilistic decision for each class.

2.5.3 Testing the Random Forests Classifier

Testing an unseen example (pixel/voxel from an image) requires the test example to traverse each tree independently until a leaf node of each tree is reached. The distribution, calculated during training (equation (2.4)), for the reached leaves from all trees are combined to give the final probabilistic decision. In the literature (Breiman, 2001, Lepetit and Fua, 2006, Schroff et al., 2008, Shotton et al., 2008, Lempitsky et al., 2009, Yi et al., 2009, Criminisi et al., 2010, Geremia et al., 2010), averaging the

probabilistic decisions from all trees is commonly performed to generate a probabilistic decision for each class of an unseen example *v* in a Random Forests framework,

$$Dist(v, RF) = \frac{1}{T}\sum_{t=1}^{T} Dist(v(leafNode).$$

$$\forall c_j \in \{c_1, ...c_{Labels}\} \quad p(c_j \mid v, RF) = Norm\left(Dist(c_j, v, RF)\right) \tag{2.5}$$

Finally, each unseen example will have a probability to belong to each class. The class of a maximum probability is usually chosen to label that example (Breiman, 2001, Shotton et al., 2008). The testing stage is summarised in Figure 2.11. Similarly, grey areas are highlighted to compare with the proposed technique.

```
Input:
• Images → Pixels/Voxels V:((v1, ?), (v2, ?)... (vK,?))
Output:
• Segmented images (Labels)
```

```
(1)For t = 1 to T
     For k = 1 to K
      Dist(k,t) {c1,c2 ..., cLabels} =
      TestDecisionTree(root(treet),vk)
     End For
   End For
(2)For k = 1 to K
    Dist(k){c1,c2 ..., cLabels} = average(Dist(k, ∀ trees)
    {c1,c2 ... cLabels})
    Labels(k)=Max(Dist(k){c1,c2... cLabels})
   End For
```

```
Function TestDecisionTree(node,v)

  If LeafNode(node) Then
    Stop & Return node.Dist
  End If
  If Decision(node.Classifier(v))=Left
    TestDecisionTree(node.LeftNode,v)
  Else
    TestDecisionTree(node.RightNode,v)
  End If
```

Figure 2.11. Pseudo-code for Random Forests testing stage.

2.5.4 Main Parameters in the Random Forests Technique

Classification using Random Forests requires several parameters to be set. These parameters are mainly for the training stage and the main ones are:

- The number of decision trees (T): In the Random Forests framework usually using more trees increases accuracy until the accuracy saturates. The choice of T is data and problem dependent. Although increasing the number of trees usually provide better classification accuracy, it linearly increases the training and testing time. We investigate this issue in Section 3.3.3.
- The maximum tree depth (L): In the top-down creation of a tree, a leaf node is created if no more gain is achieved or the maximum tree depth is reached. In theory, if this parameter is set to infinity and no split gain is considered, this will give 100% accurate training by having only one training example in each leaf. As a result, accuracy in classifying a new test case will be poor because of over-fitting. Therefore, L must be chosen carefully as a trade-off between generalisability and over-fitting.
- Number of candidate features to try at each node (n'): This is a very important parameter because it affects the classification power of each constructed tree within the forest. In this thesis, we further investigate the choice of this parameter in Section 3.2.1.
- The splitting criterion and its threshold to stop growing: In decision trees, a decision has to be made at each node to pick a classifier and to split the training examples to left and right branches. In the literature, information gain or an entropy-based criterion have been widely used (Breiman, 2001, Lepetit

and Fua, 2006, Schroff et al., 2008, Shotton et al., 2008, Criminisi et al., 2009, Criminisi et al., 2010).

- Other parameters: There are a few other parameters to be set like the minimum accepted gain of information (*e*) and the number of thresholds to try for each feature. The former affects the height of trees within the forest while the later affects the strength of features.

In the literature, all these parameters were set experimentally by either qualitative or quantitative analysis.

2.6 Medical Image Fusion

In the context used here, image fusion is the process of combining multiple aligned images. The idea can be understood better if you think of the human eyes. Each eye sees a scene from a slightly translated and rotated angle and the human brain reconstruct one image of the scene. This process is important since it enables the observer to see objects from different angles in one reconstructed view. In medical imaging perspective, reconstructing an image out of several ones of the same object where each image is captured from a slightly different angle, should provide a combined image that contains meaningful information from different sources. In our work, we already mentioned the problems in ultrasound images especially when it is used to image bony structures. Therefore, we believe that image fusion can enhance the analysis of fetal bone structures in 3D ultrasound. The idea is to 1) acquire several 3D ultrasound volumes of the same structure, i.e., the femur, 2) align the volumes and 3) fuse the information in one reconstructed 3D image.

3D image fusion has provided good improvement in adult and fetal echocardiography. It has been used, for instance, to enhance boundary definition for adult heart chambers especially the left ventricle (Soler et al., 2005, Grau et al., 2007, Rajpoot, 2009, Rajpoot et al., 2009b, Smigielski et al., 2010) and to align and fuse multiple 4D fetal echocardiography images (Gooding et al., 2010). In addition, fusion can be used to extend the field of view by stitching multiple images with some overlap (Grau et al., 2007, Rajpoot, 2009). To our knowledge, we are the first to investigate image fusion of 3D ultrasound volumes of a bone structure (the femur).

There are a few ways to combine aligned images including Mean intensity, Max intensity, Median, Mean-Shift intensity and Wavelet-based (Rajpoot, 2009, Gooding et al., 2010). The Mean intensity fusion is a simple technique in which the mean of each aligned image is computed pixel by pixel. This technique is straightforward and fast. Although it reduces speckle noise, it smoothes out the edges which does not fit well to our application since we would like to strengthen the femur boundary. The maximum (Max) intensity fusion on the other hand, enhances image detail but the Signal to Noise Ratio (SNR) may be significantly reduced. The wavelet-based image fusion has shown better results compared to the others (Rajpoot, 2009). Therefore, we adapted and modified the wavelet-based technique to fuse multiple fetal femur 3D ultrasound volumes. For more information see Chapter 4.

Chapter 3

Volumetric Segmentation using Smart Fast-Weighted Random Forests

In the previous chapter we described clinical motivation for analysing 3D fetal femoral ultrasound images. We described how Mahon (2007) manually analysed 3D fetal femoral ultrasound images to measure several characteristics from the fetal femur. In this chapter, we develop automatic techniques to segment and measure the fetal femur. Although the main aim is to automatically segment and measure fetal femoral characteristics from 3D ultrasound, we keep the methodological content of this chapter independent from the application, e.g., segmentation of the fetal femur in 3D ultrasound. Section 3.1 introduces the proposed technique and Section 3.2 demonstrates the method. These two sections are independent of the application. Section 3.3 goes on to describe the experimental setup in which we describe the datasets, validation methodology, implementation details and the automatic measurements. Note that in

Section 3.3, we do not only use the SWS dataset (Mahon, 2007) but also MRI brain datasets from the Internet Brain Segmentation Repository[7]. Section 3.4 provides a thorough evaluation of the segmentation results on both datasets. Finally, we discuss and conclude our contributions in Section 3.5.

3.1 Introduction

Random Forests (RF) which is also referred to as Randomised Trees, Random Decision Forests and other names is an ensemble classifier or regressor that consists of multiple decision trees. For further discussion of the Random Forests classifier see Section 2.5, page 20. Although the Random Forests technique is computationally efficient, inherently multi-class, able to handle huge feature space efficiently, and easily parallelisable, it has not been widely used for medical image segmentation. To our knowledge, only a handful of conference papers have been published that use the classic technique to do medical image segmentation (Lempitsky et al., 2009, Yi et al., 2009, Geremia et al., 2010, Yaqub et al., 2010b). We believe there are some issues in the way that the technique works when applying it to do medical image segmentation in 3D or 4D. The problem mainly arises from two things. First, the size of the acquired medical images increases rapidly. Since Random Forests is a discriminative classifier which relies on image features to classify objects, its success is tightly related to the discrimination power of such features. In high dimensional data, a huge number of features can be computed but usually a significant number of these are irrelevant. Therefore, the existence of "bad" features may drive the technique to provide a poor classification. Second, although the concept of equal opportunity (e.g., an equal vote from each tree in the Random Forests classifier) may be highly welcomed in different

[7] http://www.cma.mgh.harvard.edu/ibsr/

decision making problems, we believe it is not the best to use in this classifier since randomness is well used during the training stage which varies the strength of the trees in the forest.

In this work, we propose a novel segmentation technique based on the Random Forests classifier and demonstrate the technique on two different medical image segmentation problems. The main contributions are 1) a solution to overcome the problem of having a huge feature space which may contain a significant number of "bad" features and 2) a novel weighted mechanism to combine tree decisions to reduce any bias toward weaker trees. This chapter focuses purely on 3D segmentation and issues in doing this efficiently.

3.2 Method

This section starts by highlighting general problems with the Random Forests technique. We then describe the feature sets we use in our work. There are two main novel contributions to the standard Random Forests technique which lead to a better classification accuracy namely, offline feature selection in the training stage and weighted voting in the testing stage. We describe these two in detail next.

3.2.1 Problem Description

Most state-of-the-art medical imaging devices provide 3D images e.g., MRI, CT, ultrasound, etc, that represent the object of interest (like human or fetal structures). The need to analyse such images is increasingly huge. Segmenting such structures accurately is very important for clinicians. Therefore, researchers are trying to develop robust techniques to deal with this problem.

Although the Random Forests technique has been shown to provide a good framework to segment 2D images (Schroff et al., 2008, Shotton et al., 2008), straightforward extension to higher dimensions needs be dealt with carefully. One major issue to consider is the high dimensional feature space which leads to a larger number of features as well as more complex features to calculate. In the standard Random Forests technique, n' features from the feature pool of size n are randomly selected. One feature from the n' randomly selected subset is chosen to create a node in a tree. Notice that the choice of the number of randomly selected features at each node (n') is critical. If n' is large enough then the probability to have a good feature within the n' features is high. Some researchers have experimentally set n' to $\sqrt{n}$ (Breiman, 2001). In 3D images scenario, n can be a very large number (e.g., $n = 10^8$). If $n' \sim = n$ then this requires higher computation which may take days or even weeks on a standard PC to train a Random Forests classifier to segment 3D images. Decreasing the value of n' will speed up the training process but will give lower classification accuracy because of the lower chance of choosing a good classifier at each tree node. One can argue that training time is not important compared to testing time. This is true but the techniques developed so far to segment medical images using standard Random Forests use only a very small value of n'. For instance, in a recent work (Geremia et al., 2010) set n'=950 and roughly speaking used two 3D rectangles of up to 32^3 voxels each. This gives approximately 32^6 ($>10^9$) permutations of features. This gives a low probability to have a good feature within the n' ones which may lead to a poor classification accuracy. Figure 3.1 shows a histogram of Cuboid3D and Haar3D feature sets (see next section for further discussion on feature sets). From Figure 3.1 note that a small number of features are relatively "good" while many features are "bad". The performance of these features vary and a large portion of

these features have low scores. In this work, we propose a more accurate selection of features during training while preserving the random nature of the technique. This is discussed in details in section 3.2.3.

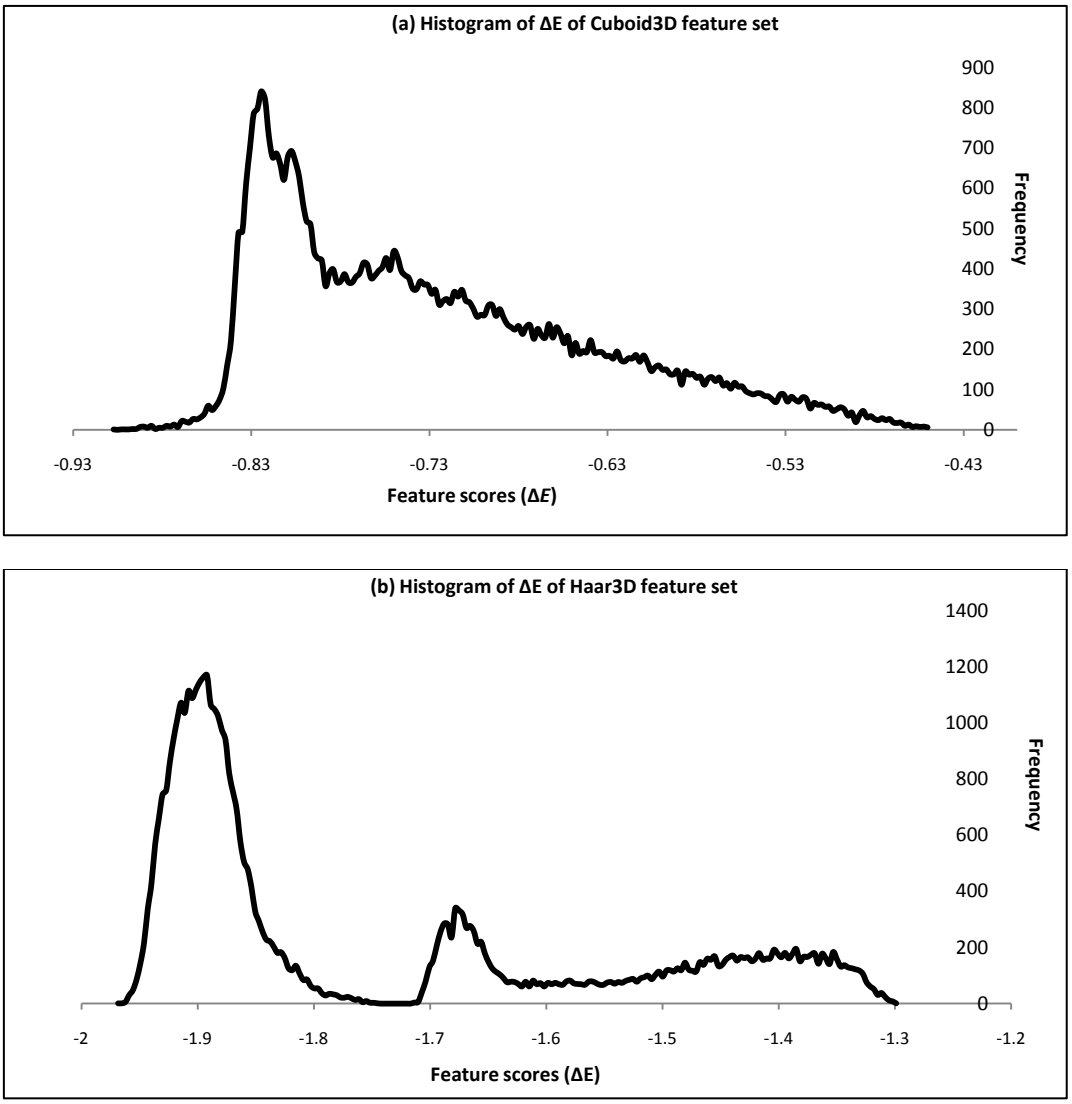


Figure 3.1. Histograms of feature scores (ΔE) for two different feature sets. (a) shows a histogram of Cuboid3D feature set scores (ΔE) generated from classifying two classes (femur and background) in 3D ultrasound images. Parameters for the rectangles are: maximum 3D rectangle dimensions is (4,4,4) and maximum 3D search area around the 3D rectangle is (x, y and z between [-4,4]). Number of features are 4^3*9^3=46656 while for instance the number of features which has $\Delta E \geq$-0.6 is 5171. (b) shows a histogram of Haar 3D feature set scores (ΔE) generated from classifying four classes (WM, GM, CSF and background) in MR brain volumes. Parameters for the Haar3D are 15^3 is the maximum cube size which generates up to 44687 feature while for instance the number of features which has $\Delta E \geq$-1.45 is 6369.

Another major issue to investigate in the traditional Random Forests technique is the way the classification of decision trees are combined during testing. Since each tree is built in a random manner, the decisions at leaves vary from one tree to another. Figure 3.2 shows segmentation results for five individual trees in one forest. Notice that each tree provides a slightly different segmentation result depending on the classifiers inside each tree. As a result, we propose a weighted voting procedure where trees should contribute unevenly to the Random Forests final decision. This is further discussed is Section 3.2.4.

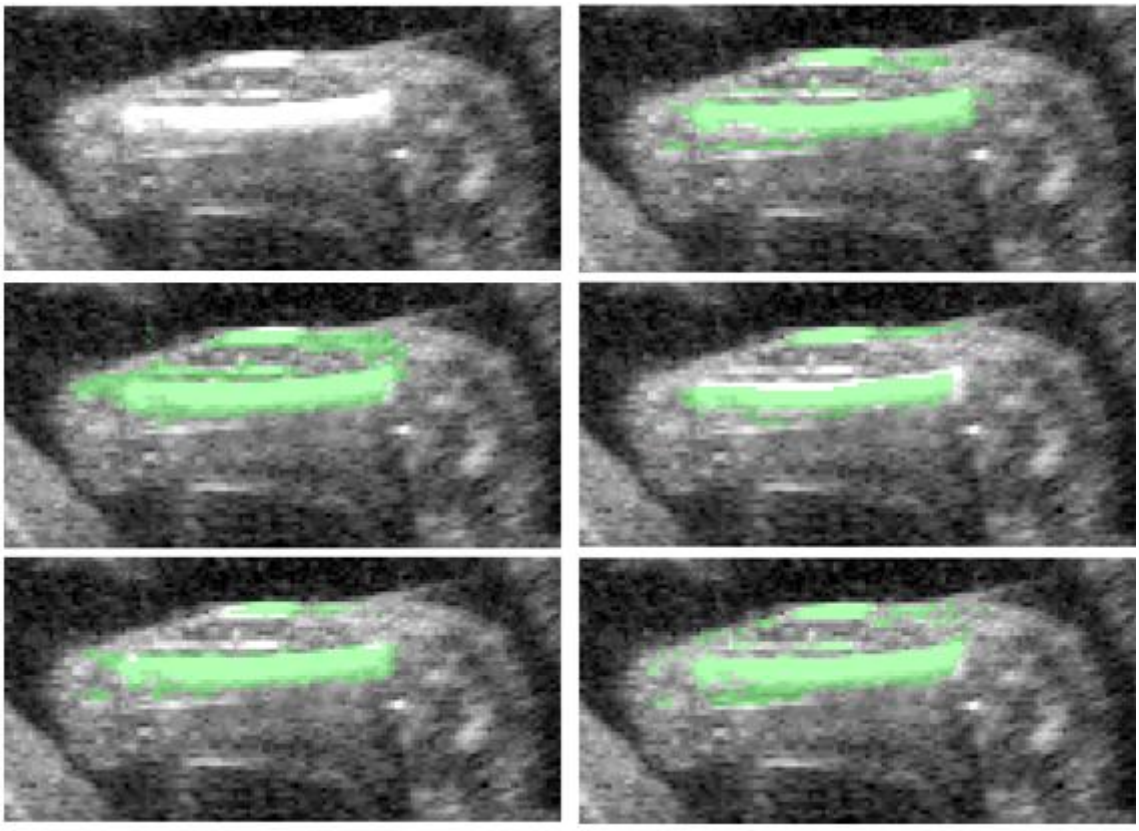


Figure 3.2. Segmentation accuracy by different trees. Top left is a 2D fetal femur slice. The remaining five subfigures are the individual segmentation results of 5 separate trees.

Researchers have mainly used the equal voting mechanism (average) to combine the probabilistic decisions as in equation (2.5) (Breiman, 2001, Lepetit and Fua, 2006, Schroff et al., 2008, Shotton et al., 2008, Criminisi et al., 2009, Criminisi et al., 2010, Geremia et al., 2010). This type of voting is acceptable if the number of trees is significantly large so that the final decision will not be biased toward specific trees. In practice, the number of trees is limited because increasing the number of trees will linearly increase training and testing times. In the medical image segmentation domain, researchers have used a relatively small number of trees (e.g., Geremia et al. (2010) used 30 trees). In this case, we argue that it is important to have a more intelligent way to combine trees decisions.

3.2.2 Feature Sets

Segmentation using Random Forests is formulated as a supervised discriminative model. This type of model starts from the training data to create a classifier. A binary classifier is meant to discriminatively distinguish background from foreground. Each node in a tree is a classifier which consists of a feature and a threshold. Features are descriptors that are extracted from the training data. Several features can be extracted from the training images. For instance, image intensities, intensity gradient magnitudes and edge orientation, curvatures, etc. More complex features can also be used which require more computation, e.g., filter-banks, Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005, Schroff et al., 2008), Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), etc. The choice of features is mainly an application dependent issue. Therefore, in this sub-section we outline the set of features we employed as the basis for our specific segmentation problems (i.e., segmenting the fetal femur in 3D ultrasound images and segmenting brain structures in MR images).

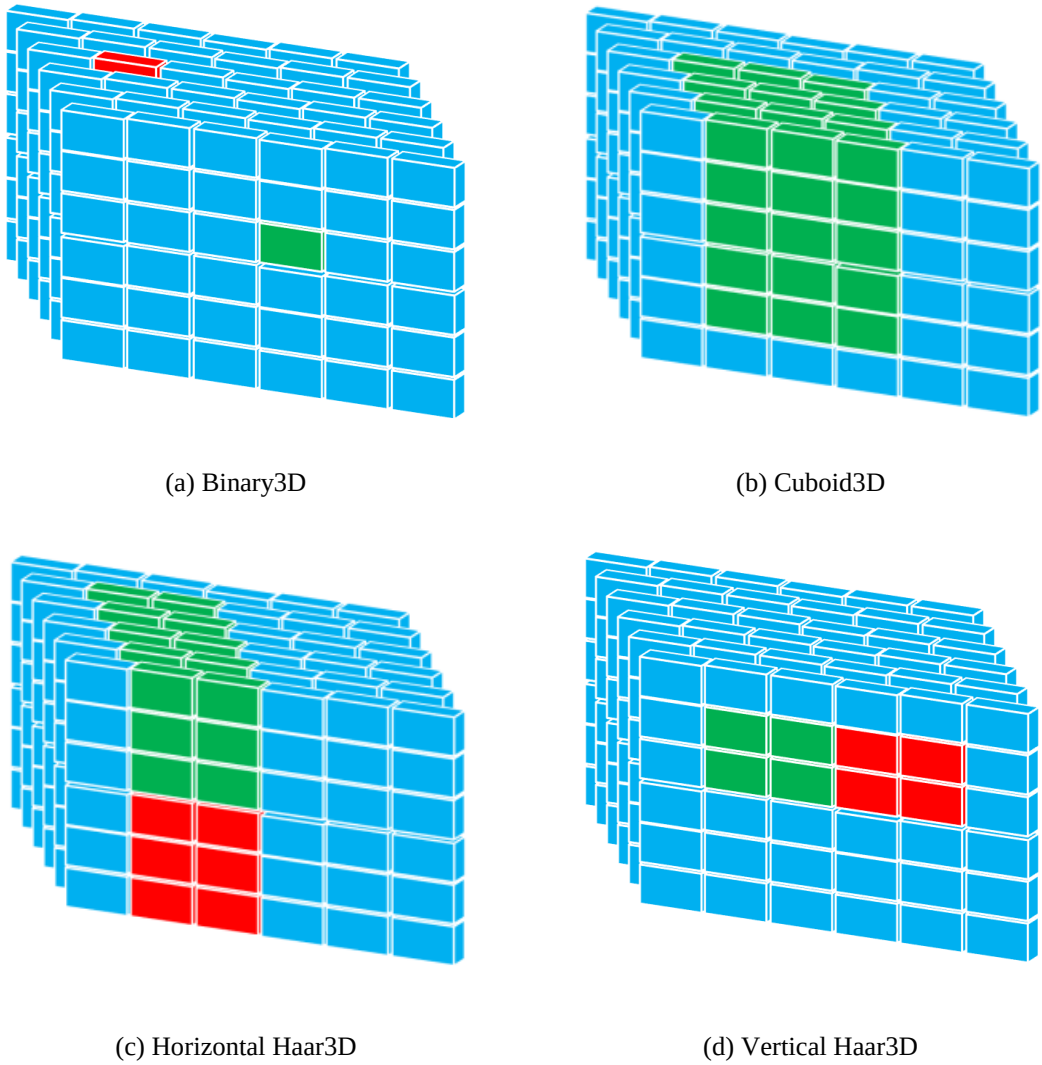
In our work, several feature sets are constructed for each image. We use the phrase "feature set" to denote the group of features of the same type but with different dimensions and locations around a Voxel Of Interest (VoxOI). Unary3D, Binary3D, Cuboid3D (Shotton et al., 2008), Haar3D (Oren et al., 1997, Caruana and Niculescu-Mizil, 2006, Carneiro et al., 2008), Position3D (Lempitsky et al., 2009) and Averaged Cuboid3D (Average3D) feature sets are computed. Figure 3.3 illustrates some of these feature sets.

A Unary3D feature is the intensity value of a random voxel within a random size window around VoxOI. A Binary3D feature is the sum, difference or absolute difference of two random voxel intensities in a random window around VoxOI (Shotton et al., 2008). A Cuboid3D feature is the sum of all voxel intensities in a random size rectangular cuboid starting from a random coordinate around the VoxOI. These feature sets have shown good performance in natural image segmentation (Schroff et al., 2008, Shotton et al., 2008). We extend the 2D integral images (Carneiro et al., 2008) to 3D to find a 3D rectangular sum within a 3D image efficiently as follows

$$Integ(x,y,z)=\sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z}I(i,j,k), \tag{3.1}$$

$$\begin{aligned} &3D-rect-sum\left((x_2,y_2,z_2),(x_1,y_1,z_1)\right)= \\ &\quad Integ(x_2,y_2,z_2)+Integ(x_1,y_1,z_2)+Integ(x_1,y_2,z_1)+Integ(x_2,y_1,z_1) \\ &-Integ(x_1,y_2,z_2)-Integ(x_2,y_1,z_2)-Integ(x_2,y_2,z_1)-Integ(x_1,y_1,z_1), \end{aligned} \tag{3.2}$$

where I is a 3D image and *Integ* is the 3D integral image calculated from I. *3D-rect-sum*$((x_2,y_2,z_2),(x_1,y_1,z_1))$ is the sum of all voxel intensities of a cuboid whose upper top left coordinate is (x_1,y_1,z_1) and lower bottom right coordinate is (x_2,y_2,z_2).



(a) Binary3D

(b) Cuboid3D

(c) Horizontal Haar3D

(d) Vertical Haar3D

Figure 3.3. Examples of the feature sets. The *green* color voxels are summed and subtracted from the summed *red* voxels.

Since Cuboid3D features depend on the size of rectangular cuboid which is biased to its dimensions, we propose an Averaged Cuboid3D which is the 3D rectangular sum divided by the number of voxels in the rectangular cuboid. The Averaged Cuboid3D feature set is actually a Unary3D feature set of the down-sampled image. Haar3D features (one of the wavelet family) are used to capture edge regions (Oren et al., 1997). Haar features are mainly the difference between two or more regions. Each region is usually represented by a Cuboid3D feature. We used 2-blocks and 3-blocks horizontal

and vertical Haar features. For more information see (Carneiro et al., 2008). Finally, Position3D features are used to capture information about the spatial location of voxels. This feature set helps discard many regions that have similar intensity and edge information to the object of interest (e.g., in our case allowing to distinguish the femur from other bone structures). To make the position feature set work, the object of interest has to be approximately aligned in all images.

3.2.3 Fast Random Forests Training

Notice that the gray area in Figure 2.7 (point number (3)) represents the selection of a classifier from a feature pool. We already pointed out two important issues with the current procedure. First, since the feature pool contains many "bad" features, see Figure 3.1, the probability to select a good feature is low. This affects the creation of all trees where many nodes will poorly classify the data. One solution could be to increase the value of n' to maximise the chance of having a good feature within the n' ones but this requires much higher computations. Second, if a Random Forests classifier consists of T trees and each tree has an average height H then the loop in the gray area will be executed approximately $(2^H\text{-}1)\times n'\times T$ times. This is a very time consuming process during training especially if the value of n' is large enough to guarantee the existence of a good feature within the n' ones. This is the main reason for using a small value of n' in all previous work (Criminisi et al., 2009, Lempitsky et al., 2009, Yi et al., 2009, Criminisi et al., 2010, Yaqub et al., 2010b).

We propose an offline feature selection step in which we exhaustively pre-train all features on a well known gold standard (e.g., carefully manually segmented volume). The values of ΔE for all features are recorded as a global performance metric of each feature. A subset of features from each feature set (Binary3D, Cuboid3D, Haar3D, etc)

whose ΔE is greater than a specific value ($\Delta E Threshold(feature\ set)$) are retained. The choice of such a threshold is feature set and application dependent, and therefore was chosen empirically. The feature selection pseudo-code is shown in Figure 3.4. From the global score of each feature (ΔE), a smaller pool of features of size m (lookup table of size m) is chosen from the original pool (where $m<<n$). The scores are normalised to a specific range since each feature set has a slightly different range.

One can imagine the offline feature selection step, that creates a lookup table of "good" features, as training one node in a tree where n' is n. The training examples for this step can be a large enough subset of the whole training examples which has a well-defined ground truth. It is important that this subset captures all variation in the data. This process is time-consuming but it is required to be done only once for each dataset. For example, for a specific MRI scanner the offline feature selection step on brain images with their manual segmentation is performed once and the lookup table can then be used to efficiently train as many classifiers as possible. This means that this step is scanner (image source) specific and object-of-interest specific.

```
Input:
• Images + Ground Truth (GT represents the segmented images)
• ΔEThreshold
(1)For each feature set (FS_i)
(2) For each Feature f_k in FS_i
    • Apply feature (f_k) to the training examples
    • Find the best threshold th for f_k
    • If ΔE(f_k, th)> ΔEThreshold(i) Then
       Add (k, f_j, ΔE) to a lookup table such that k is the
       key and (f_j, ΔE) is the item
     End If
    End For
(3)Normalise ΔE for all features in FS_i
  End For
```

Figure 3.4. Offline feature selection step.

We propose to use the global score for each feature to modify the way the classic RF works. The first modification is to create a smaller feature pool (of size m) of "good" features from the main pool (which is of size n). This enhances the choice of the chosen random feature and eliminates "bad" ones. The second modification we propose is in the training of features during the node creation process. Remember that during node creation, several candidate features (we call it m' to distinguish it from n' in the classic RF) are chosen from the smaller pool to pick the best one. Basically, we can find the best feature in two ways. The first solution is to train on all the m' features as in the classic RF. This keeps the training time of the proposed technique similar to the classic RF. The other solution is not to train on the m' features but to choose the best feature out of m' depending on the pre-computed ΔE which we find during the creation of the smaller feature pool (see Figure 3.4). The chosen feature will then be applied to the training examples to construct a classifier. This means that m' lookup table comparisons are needed and only one feature is computed at each tree node instead of n' in the standard Random Forests technique. This makes the total number of features to compute approximately $[(2^H - 1) \times 1 \times T]$ instead of $[(2^H - 1) \times n' \times T]$ in the traditional Random Forests framework i.e., saving a factor of n' where n' is typically >100. The reliance of the pre-computed ΔE to choose the best feature significantly speeds up the training process which enables us to run the training process several times with different parameters (e.g., n', maximum tree depth, number of trees, etc...) and different settings (e.g., depth vs. breadth creation of trees). Figure 3.5 shows how the proposed technique can select and train a classifier. This pseudo-code is a replacement to the gray area in Figure 2.7. To summarise, we propose two modifications to the classic RF:

```
(1)Load the lookup table
(2)Select m' random keys from the lookup table
(3)Choose the feature f_k that has the maximum global score (ΔE)
   from the m' random ones
(4)Apply the chosen feature f_k to the training examples
(5)Find the best threshold th for f_k
(6)Set Best_Classifier =(f_k, th, ΔE)
```

Figure 3.5. The proposed selection of a classifier to create a node in a tree.

1. Smaller pool of features that contains "good" features is created and used in training.
2. In the node creation process we could do either a or b below
    a. create a node as in the classic RF. This is important to study effect of the first modification alone. We call this the Smart Random Forests (SRF).
    b. select the best feature of the *m'* candidate features at each node depending on the pre-computed information gain (*ΔE*). We call this the Smart Fast Random Forests (SFRF).

Note that we abbreviate the classic Random Forests as RF.

Notice that in point (6) in Figure 3.5 we retain not only the best classifier (feature and threshold) but also *ΔE*. We use this value later in testing as a global measure of how good each node in each tree. For more information see the next section. The remainder of the training stage follows the steps of the traditional Random Forests technique.

3.2.4 Weighted Random Forests Testing

During the testing stage, an unseen example traverses each tree starting from the root until it reaches a leaf node. The final probabilistic decision is combined from the *T*

probabilistic distributions in the *T* trees. It is clear from Figure 3.2 that each tree in a forest contributes differently to the final decision. This has mainly happened because of the random selection and creation of nodes. This also shows that classifiers vary in their classification accuracy. Therefore, we propose a weighted voting mechanism to calculate a probabilistic decision for each unseen example based on the strength of each tree to classify a specific unseen example.

Remember that each node does not only contain a classifier but also a score (ΔE) of how well that classifier did during the training stage. An unseen example follows a path starting from the root of the tree until it reaches a leaf node. In that path, it visits approximately *H* (average tree height) classifiers on each tree until it reaches a leaf node. We used the mean of ΔE computed from the *H* classifiers that exist on that path as a measure of how good the tree is at classifying the test example. This can be formulated in equation (3.3) and is depicted in Figure 3.6.

$$Score(v, tree_t) = \frac{1}{H_t} \sum_{h=1}^{H_t} \Delta E_h (node_h \; at \; tree_t). \tag{3.3}$$

After an unseen example (*v*) is classified on *T* trees, the *T* scores are normalised so they sum to one:

$$Norm\big(Score(v, tree_t)\big) = \frac{1}{T-1}\left(1 - \frac{\frac{1}{H_t}\sum_{h=1}^{H_t} \Delta E_h (node_h \;\; at \;\; tree_t)}{\sum_{i=1}^{T}\left(\frac{1}{H_i}\sum_{h=1}^{H_i} \Delta E_h (node_h \;\; at \;\; tree_i)\right)}\right). \tag{3.4}$$

The normalised weight to classify an unseen example *v* is then embedded in the probabilistic classification so that equation (2.5) that is used in the classic Random Forests technique becomes:

$$\forall c_j \in \{c_1,...c_{Labels}\} \;\; p(c_j \mid v, RF) = \sum_{t=1}^{T} Norm\big(Score(v, tree_t)\big) p(c_j \mid v, leaf\,(tree_t)). \tag{3.5}$$

The complete testing pseudo-code is illustrated in Figure 3.7. The main changes between Figure 3.7 and Figure 2.11 are shaded in both figures. We add the letter "W" to the abbreviation of the variations of Random Forests. As a result, we have the classic Random Forests (RF), Smart Random Forests (SRF), Smart Fast Random Forests (SFRF), Weighted Random Forests (WRF), Smart Weighted Random Forests (SWRF) and Smart Fast Weighted Random Forests (SFWRF).

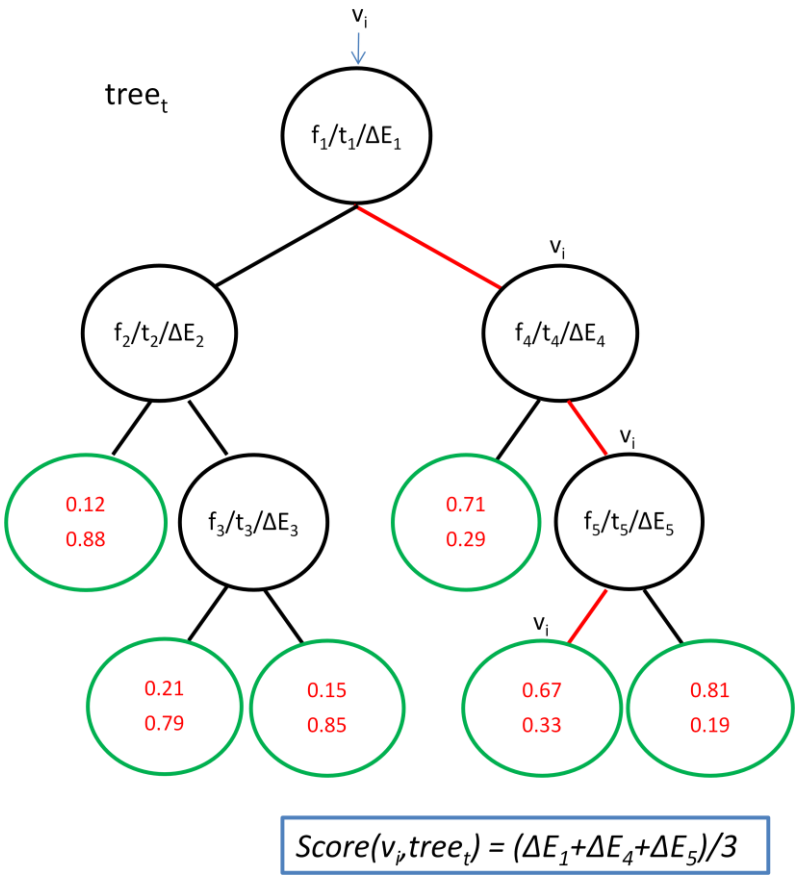


Figure 3.6. An example of a trained tree in a forests. Testing an unseen example v_i at the $tree_t$ starts from the root. In this example, *red* edges represent the path v_i followed (f_1, f_4 and f_5) until it reached the leaf node which contains the probabilistic distribution for two classes (0.67, 0.33). The strength (score) of this tree to classify this unseen example is calculated below the tree.

```
Input:
• Images → Pixels/Voxels V:((v1, ?), (v2, ?)... (vK,?))
Output:
• Segmented images (Labels)
(1)For t = 1 to T
     For k = 1 to K
     [Score(k,t),Dist(k,t){c1,c2 ..., cLabels}f] =
     TestDecisionTree(root(treet),vk,0)
     End For
   End For
(2)Normalise Score(t,k) as in eaquation (3.4)
(3)For k = 1 to K
    Dist(k){c1,c2 ..., cLabels} = (Norm(Score(k, ∀ trees ))×
    Dist(k, ∀trees) {c1,c2 ... cLabels})
    Labels(k)=Max(Dist(k){c1,c2... cLabels})
   End For
```

```
Function TestDecisionTree(node,v, Score)

  If LeafNode(node) Then
    Stop & Return (Score/node.Level) and node.Dist
  End If
  Score = Score + Node.ΔE
  If Decision(node.Classifier(v))=Left
    TestDecisionTree(node.LeftNode,v, Score)
  Else
    TestDecisionTree(node.RightNode,v, Score)
  End If
End TestDecisionTree
```

Figure 3.7. Pseudo-code for proposed testing stage.

Finally, to validate our proposed weighted voting testing we compared it to 1) the classic RF equal voting from all trees (averaging) (Schroff et al., 2008, Shotton et al., 2008, Criminisi et al., 2009, Lempitsky et al., 2009, Yi et al., 2009, Geremia et al., 2010) and 2) the equal voting from the most confident trees (Criminisi et al., 2010). The key idea is to use the certainty of the distribution at leaf nodes to assess if a leaf has a confident soft decision or not to do this. In this thesis, we use a modification of the validation approach proposed by Criminisi et al. (2010). We use the entropy $E(v,tree_t)$

of the probabilistic distribution at leaves as a certainty measure, see equation (2.3). This measure is normalised (equation (3.6)) and used to weight all trees (equation (3.7)).

$$Norm\left(E(v,tree_t)\right)=\frac{1}{T-1}\left(1-\frac{E(v,tree_t)}{\sum_{i=1}^{T}\left(E(v,tree_t)\right)}\right), \tag{3.6}$$

$$\forall c_j \in \{c_1,\ldots c_{Labels}\} \quad p(c_j \mid v, RF)= \sum_{t=1}^{T} Norm\left(E(v,tree_t)\right)p(c_j \mid v, leaf\ (tree_t)). \tag{3.7}$$

3.2.5 Optional Post-processing

This step is application-specific. In the next section, we describe the two datasets we used in validation. Post-processing was only needed in the first dataset (3D ultrasound fetal femur) and is mainly applied to reject regions with similar local shape and intensity distribution to the object of interest. Although Random Forests classifier provides good classification accuracy, it employs a discriminative model that captures local similarities. As a result, any structure which looks similar in intensity distribution and local shape can be regarded as the object of interest. A position feature set is added to reject such regions. Unfortunately, some femur like structures are close to the femur and therefore position features may not be able to distinguish between the two (e.g., the femur is connected to tibia via the knee ligaments). To accommodate this, the largest 3D connected component was automatically selected (the femur). This step was not used for the brain MRI experiment and does not affect the accuracy of segmentation of identified objects.

3.3 Experimental Setup

3.3.1 Dataset

Two datasets were used to validate the proposed technique. The first dataset is 51 fetal femur ultrasound volumes from the SWS dataset (Mahon, 2007). For further description of the SWS dataset see Section 2.3.1. Note that although 51 volumes were used in quantitative analyses in this chapter, the remaining 399 volumes were used in the clinical comparisons, see Chapter 5. In this dataset, volume dimensions are approximately 70×70×140 voxels with a (0.5×0.5×0.5) mm^3 voxel spacing. Fetal femurs were manually segmented from the 3D ultrasound and used in the validation. The manual segmentation of the fetal femur from the 51 ultrasound volumes was performed by an expert (M. Y.) who underwent extensive training by an experienced superintendent sonographer (P. M.). The manual segmentation was also assessed by a clinician (C. I.). The manual segmentation was performed on a dedicated machine with an interactive monitor (WACOM CINTIQ 21UX). The software used to do the manual segmentation is MITK (the Medical Imaging Interaction Toolkit MITK, version 0.12.2, (c) 2003-2008 German Cancer Research Center, Division of Medical and Biological Informatics).

The second dataset is brain MR images from the Internet Brain Segmentation Repository (IBSR)[8]. We have mainly chosen this dataset so we can compare our proposed technique to state-of-the-art ones on a public database. The dataset consists of 20 T1-weighted MR brain images of normal subjects. The manual segmentation of these volumes are available. The MR images were manually segmented to four objects namely White Matter (WM), Gray Matter (GM), Cerebro-Spinal Fluid (CSF) and

[8] http://www.cma.mgh.harvard.edu/ibsr/

background. Two 1.5 Tesla MRI scanners have been used (10 scans on Siemens Magnetom MR System and 10 scans on General Electric Signa MR System). Volume size is 256×256×[60-64] voxels with 1×1×3.1 mm^3 voxel spacing. MRI brain segmentation is still an open problem to solve accurately and its clinical motivations are uncountable (Balafar et al., 2010).

3.3.2 Validation Methodology

Experiments comparing the RF, SRF, SFRF, WRF, SWRF, SFWRF are performed. We only report the results of the 3D techniques although results from 2D Random Forests were used in Chapter 5. The validation was performed on the two datasets (fetal femoral ultrasound and brain MRI). Although cross-validation is a sensible approach to use to validate a supervised classification technique, it is extremely time consuming to use for 3D medical image segmentation. In addition, since the number of training examples is huge (e.g., approximately $1.4*10^7$ in the fetal femoral ultrasound dataset and $8*10^7$ in the MRI dataset), splitting the data to two groups (training and testing) is a reasonable validation.

In the first experiment, from the 51 manually segmented cases, the offline feature selection step was performed on 10 3D fetal femoral ultrasound images. Training was done on another 20 volumes. Results are reported on segmenting the femur from the remaining 21. In the second experiment, MR images were first corrected of bias field using the N4 algorithm (Tustison et al., 2010). Since there were 20 cases only, 5% of the training voxels were randomly selected from the 20 volumes for the feature selection step. Training was performed on 10 MR volumes while testing results are reported on the remaining 10 volumes.

If the number of training examples for each class in a Random Forests framework are highly different, this can lead to a highly imbalanced trees which in turn affects the classification accuracy. This is the case for both datasets. Therefore we randomly sub-sampled from the majority classes to approximately match the size of the minority class (Chen et al., 2004, Thomas et al., 2006).

We use precision, recall, dice similarity and Jaccard index as measures of how accurate the automatic segmentation techniques are. We also utilise the precision/recall curve to investigate the best posterior threshold to use in our experiments. Although these four measures find the similarity between the manually segmented object of interest and the automatically segmented ones, they have slightly different goal. Precision measures the similarity between the correctly segmented object (the intersection between the manual and automatic segmentation) to the automatically segment one, see equation (3.8). Recall measures the similarity between the correctly segmented object to the ground truth, see equation (3.9). Dice similarity coefficient and Jaccard index measure the similarity between the correctly segmented object to both the ground truth and the automatically segmented object, see equations (3.10) and (3.11). If both the ground truth and the automatically segmented object have the same size then precision, recall and dice coefficient are the same otherwise they are different. On the other hand, Jaccard index finds the ratio between the correctly segmented object with respect to both segmentation (the union between the manual and automatic segmentation), see equation (3.11).

$$precision = \frac{|GT \cap Auto|}{|Auto|} = \frac{|correctly\ Segmented|}{|Auto|}, \tag{3.8}$$

$$recall = \frac{|GT \cap Auto|}{|GT|} = \frac{|correctly\ Segmented|}{|GT|}, \tag{3.9}$$

$$dice = \frac{2 \times |GT \cap Auto|}{|GT| + |Auto|} = \frac{2 \times |correctly\ Segmented|}{|GT| + |Auto|}, \tag{3.10}$$

$$Jaccard = \frac{|GT \cap Auto|}{|GT \cup Auto|} = \frac{|correctly\ Segmented|}{|GT \cup Auto|}. \tag{3.11}$$

In all these equations, |.| represents the cardinality of a set which is the number of elements in a set. Also, *GT* represents the ground truth which is the manually segmented object of interest while *Auto* represents the automatically segmented object of interest. All these measures lie between 0 and 1 where the larger the value the more accurate and in turn the better.

3.3.3 Implementation Details

Although choosing the right training parameters is critical, it is problem and data dependent. Therefore, it is hard to choose the optimal parameters in training that work best in testing. As a result, researchers tend to empirically choose the parameters. In our experiments, the main training parameters are 1) the number of trees (*T*), 2) the maximum tree depth (*L*), and 3) the number of features to test during node creation (*n'* in the classic Random Forests and *m'* in the proposed Random Forests). In the offline feature selection step, the main parameters are *ΔEThreshold* for each feature set which we set to keep approximately 10-15% of the original pool of the feature sets. This means that $m \sim = 0.125 \times n$. All parameters were set experimentally and they are shown in Table 3.1 for both experiments. Notice that the maximum tree depth (*L*) was lowered for the ultrasound fetal femoral segmentation experiment. This has been done to avoid probable over-fitting. This binary classification problem tends to produce better results on new cases by using a tree depth $L = 11$, see Figure 3.8.

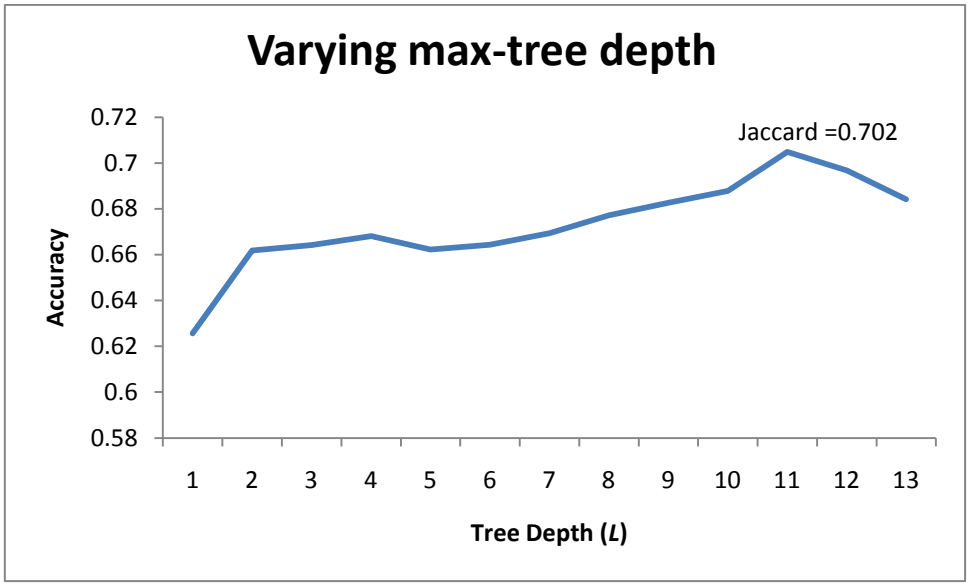


Figure 3.8. Mean Jaccard measure as a function of max-tree depth (L) with fixed forest size T = 20.

To find the best number of trees (T), we ran a small experiment on 20 fetal femoral ultrasound volumes. The goal was to train several classic RFs with a different number of trees and to visually compare the segmentation accuracy to choose a good value for T. Figure 3.9 shows an example of a 2D slice out of a 3D volume segmented using the classic Random Forests technique that is trained on different number of trees. We found that using T=20 gave a visually good segmentation. Consequently, we trained several forests with different number of trees between 1 and 20 trees to find where the accuracy saturates. Figure 3.10 shows the relationship between the accuracy (Jaccard index) and the number of trees in the forest. Although the accuracy tends to saturate at T=17, we used T=20 in our experiments to make sure that we were not at the edge of the value of T. We have not done this test to the MRI dataset but in the literature T~=20 has been used in related publications (Criminisi et al., 2009, Yi et al., 2009, Criminisi et al., 2010, Geremia et al., 2010).

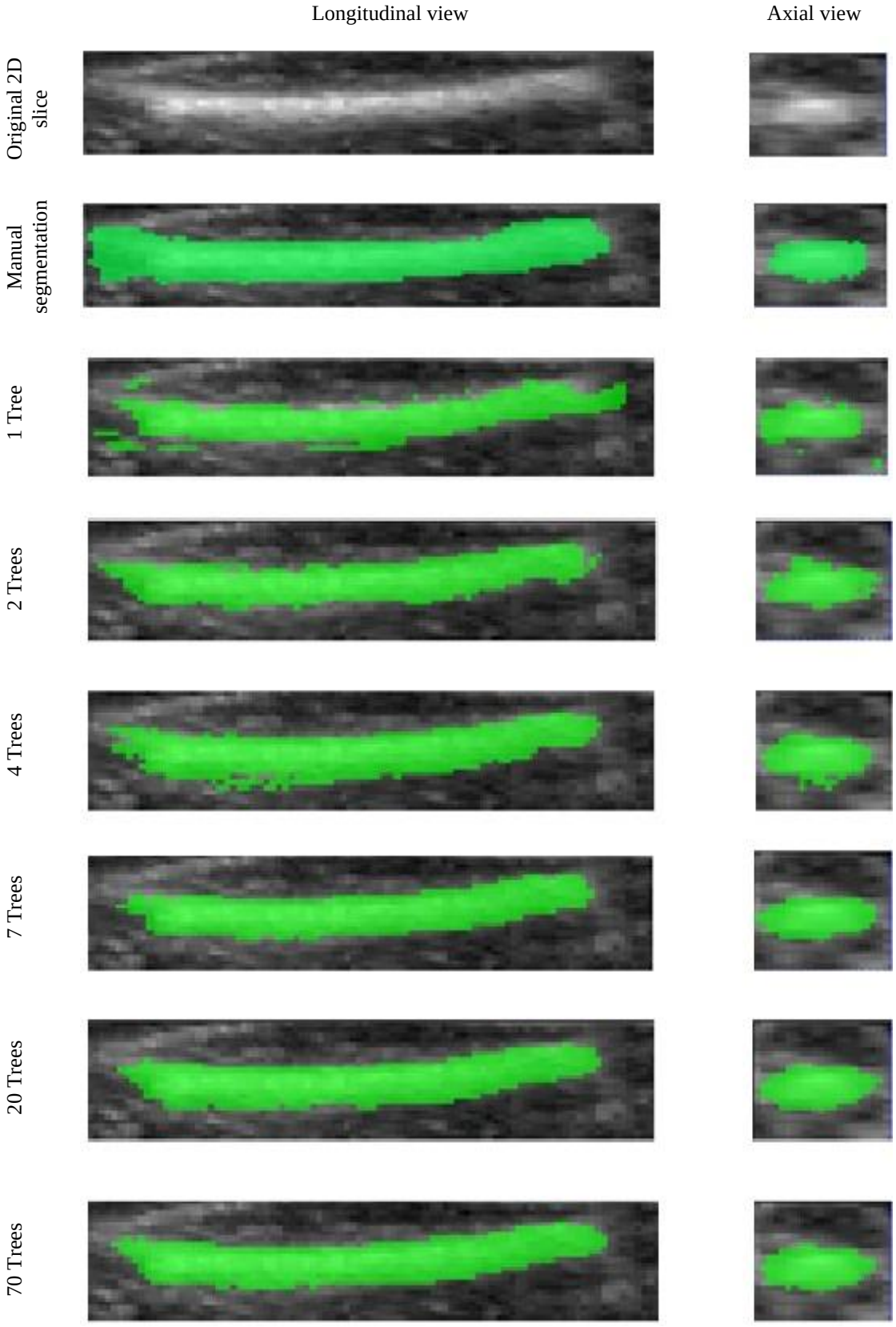


Figure 3.9. 2D slices that show the segmentation of the fetal femur in 3D ultrasound using the classic Random Forests which is trained on different number of trees.

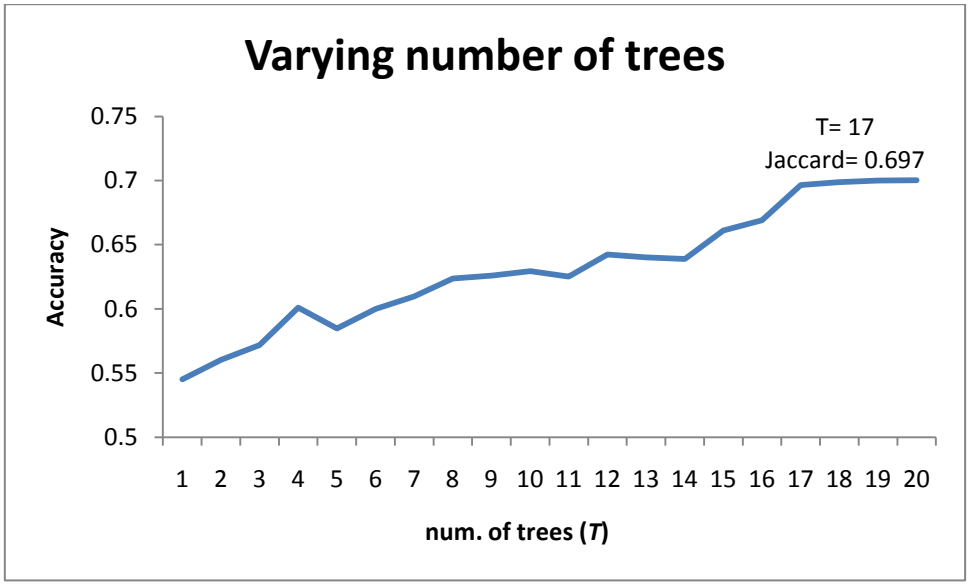


Figure 3.10. Mean Jaccard measure as a function of forest size (T) with fixed max-tree depth L = 11.

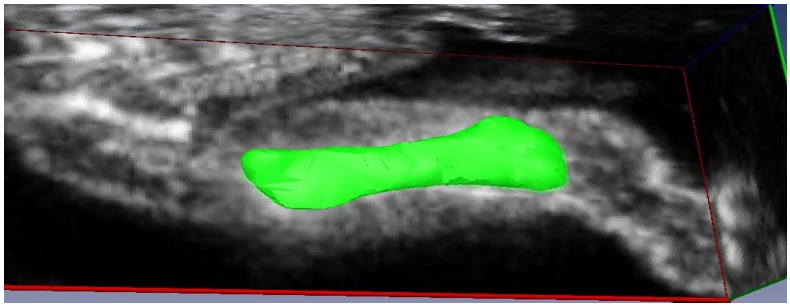
Table 3.1. The main parameters in both experiments (fetal femoral segmentation MRI brain)

	T	L	n'	m'
Ultrasound fetal femur	20	11	100	100
MRI brain	20	15	200	200

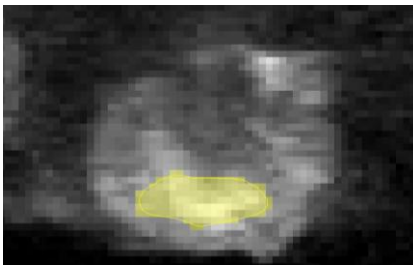
Since training and testing each tree is independent, parallel implementation is possible. Therefore, we trained and tested trees in parallel. Parallelism usually drops the running time by a factor of min(T, number of processors). For instance, if T is 20 and number of processors is 10 then approximately 10X speed can be achieved. In our experiments we used 10 processors. Finally, the code was implemented in the C# language.

3.3.4 Automatic Fetal Femoral Measurements

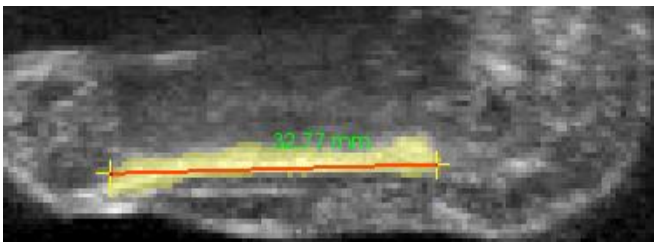
From the previous steps, a 3D segmented object-of-interest is automatically found. For the MRI brain dataset we are only interested in a quantitative comparison with the manual segmentation while in the fetal femoral dataset we are also interested in the clinical comparison with fetal and maternal characteristics, see Chapter 5. Therefore, the fetal femur in the 450 ultrasound volumes was segmented and several measurements from it were automatically generated. Figure 3.11 (a) shows an example of a 3D segmented femur. Several measurements can be automatically found from the segmented femur. The main target of these measurements is to correlate them with vitamin D level in mothers and fetal gestation age. The measurements we calculated were Femur Volume (FV), Femur Length (FL), three femur Cross-Sectional Area (CSA), six splaying indices and six splaying angles.



(a) The whole fetal femur segmented (FV)



(b) The fetal femur cross-sectional area (CSA)



(c) The fetal femur length (FL)

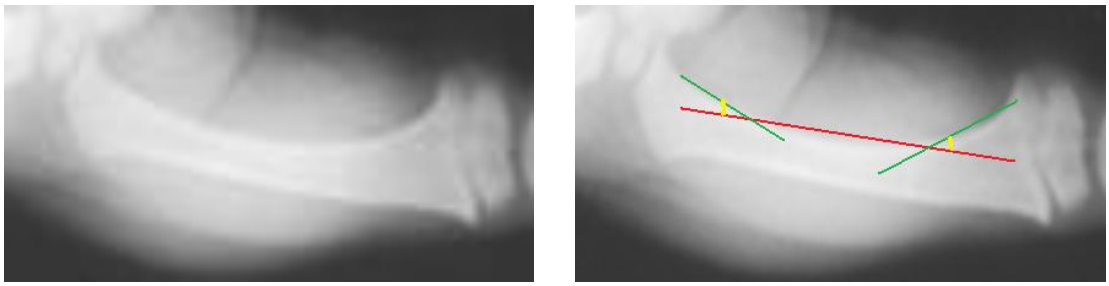
Figure 3.11. Automatic fetal femoral measurements. (a-c) represent the automatically measured fetal femur volume, cross-sectional area and length respectively. These measurements were automatically found from the segmented femur. For instance (c) shows how femur length is calculated as the distance between both left and right femur ends of the segmented femur from a 2D longitudinal view.

Femur volume was calculated by multiplying the number of segmented voxels by voxel size. Voxel size is the voxel dimension in x-axis × voxel dimension in y-axis × voxel dimension in z-axis. Femur length was calculated as the maximum length of a straight line of pixels from the longitudinal view. Femur cross-sectional area, splaying index and splaying angle were more tricky since two measurements can be found from the two sides of the femur. For each one of these measurements we calculated them from both sides and we then reported the maximum (Max), average (Avg) and minimum (Min) and of the two to be used in clinical comparison (see Chapter 5). To measure cross-sectional area from each side, we found the maximum area of the femur from each side calculated from the transverse view. Variable names of the three cross-sectional areas are MaxCSA, AvgCSA and MinCSA. Two ways were used to measure the splaying index. The first splaying index is the same as in (Mahon, 2007), i.e., PSI=CSA/FL. The second splaying index is slightly different where the denominator is the femur length squared to make the index unitless, i.e., $MSI=CSA/FL^2$. Notice that six splaying indices are generated since we have three CSAs. As a result, MaxPSI, AvgPSI, MinPSI, MaxMSI, AvgMSI and MinMSI are reported. Figure 3.11 shows a geometric representation of the femur volume, cross-sectional area and length in a fetal ultrasound image.

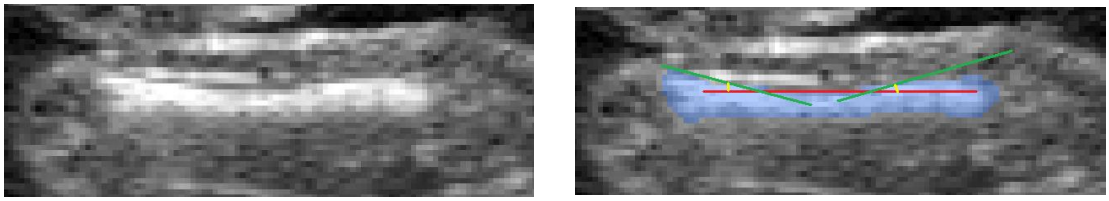
We have also used splaying angles in the clinical comparison. A splaying angle is the angle between the tangent of the femur epiphysis upper border and the tangent of the mid-shaft upper border. Two angles can be found from the left and right epiphyses. The geometric description of such angles is shown in Figure 3.12. To calculate the left and right splaying angles we need to find two lines, namely the upper border of the mid-shaft and the upper borders of both epiphysis. These lines are approximated and found

from a 2D longitudinal view of the segmented border of the femur. The 2D slice was chosen such that it contains the largest femur where femur length is usually measured. The first line is the mid-shaft line which is the straight line that approximates the mid-shaft upper border, see the *red* line in Figure 3.12 (a & b). Finding this line is relatively easy since the upper border of the mid-shaft is mainly straight. Therefore the mean of the gradient of the border is used to find the slope of that line ($Slope_1$). The second line is the epiphysis line which is an approximation of the epiphysis upper border, see *green* lines in Figure 3.12 (a & b). Finding this line is tricky because the shape of the upper border of the epiphysis is very irregular. In addition, this part of the femur has less accurate segmentation because of its poor definition compared to the mid-shaft. We also used the mean of the gradient of the epiphysis border to find the slope of that line ($Slope_2$). The inverse tangent of the difference of the slopes of the two lines are used to estimate this angle, i.e., $\theta = tan^{-1}(Slope_1 - Slope_2)$.

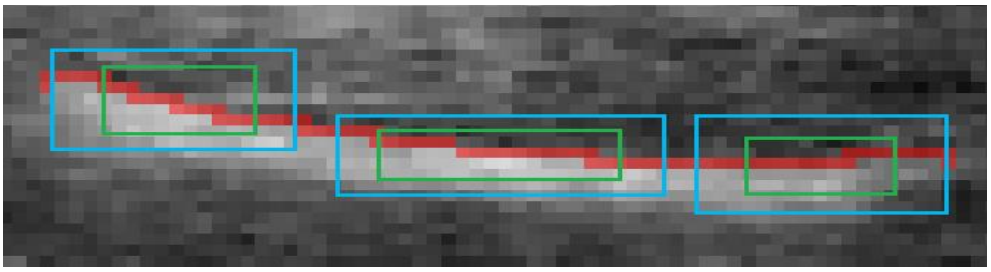
The only remaining thing to decide is which part of the upper border of the femur belongs to the upper border of the mid-shaft and which belongs to upper border of the left and right epiphyses. We have tried two methods (flexible and strict) to find which part is which. Both methods are visually described in Figure 3.12 (c) by two different colour boxes. The blue boxes include more upper border of the femur within each region while the green ones restrict the technique to a smaller region. The mean gradient of the border within each box is considered the slope of the line. Both methods only vary of the range of femur upper border to approximate a line but both certainly provide slightly different angles, see Figure 3.12 (d).



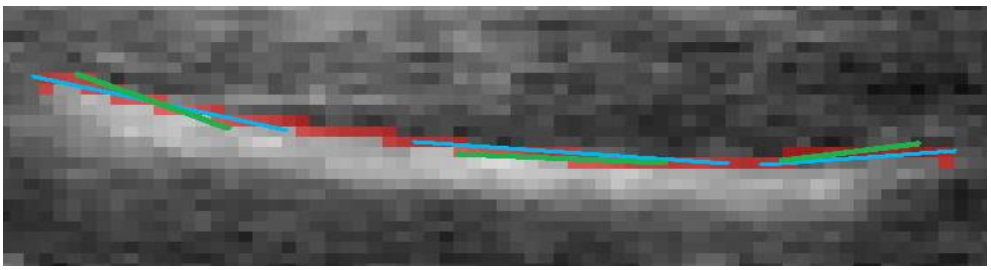
(a) DXA scan of a neonatal (left) and the two splaying angles (right)



(b) 19 weeks gestation ultrasound scan (left) and two splaying angles (yellow) calculated from the border of the segmented femur (blue)



(c) Two methods to find the border that belongs to the mid-shaft and each epiphysis. The two methods are represented by blue (method1) and green (method2) boxes.



(d) Approximate lines using the two methods. Blue represents method1 and green represents method2.

Figure 3.12. Geometric representation of the splaying angles and how they are measured. (a & b) *Red* and *green* lines are approximations of the upper border of the femur since it is usually not straight. *Red* line represents the straight line of the mid-shaft of the femur while *green* lines represent straight lines of both ends. *Yellow* angles represent the calculated angles between the *red* and the *green* lines. (a) shows a DXA scan for a neonatal femur. (b) shows a fetal femur ultrasound scan. (c & d) describe the two methods of finding the borders that belong to the mid-shaft and each epiphysis. The *red* line in (c & d) represents the automatically segmented upper border of the femur. *Blue* and *green* boxes in (c) show the two regions from which the lines are approximated from the *red* border.

3.4 Results

We first investigated the selection of the best posterior threshold. We then report the segmentation results using the classic Random Forests (RF), the smart classic Random Forests (SRF), the smart fast Random Forests with offline feature selection and equal voting (Smart Fast Random Forests (SFRF)), and the smart fast Random Forests with offline feature selection and weighted voting (Smart Fast-weighted Random Forests (SFWRF)).

3.4.1 The Fetal Femur Segmentation in 3D Ultrasound

3.4.1.1 Choosing The Best Posterior Threshold

We first investigated the best posterior threshold to use since the Random Forests technique generates a soft decision. Figure 3.13 shows the precision/recall curve at which the small *red* point represents the point where we believe the best trade-off between precision and recall exists (posterior threshold = 0.64).

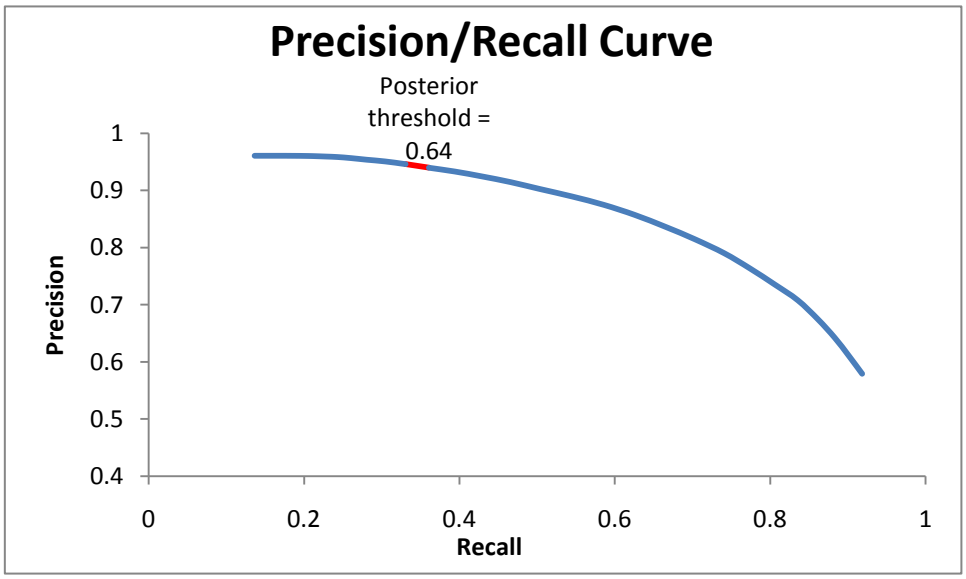


Figure 3.13. Precision/Recall curve to find the best posterior threshold.

3.4.1.2 Selected Features

We investigated the feature sets in the forest to report the frequency of selected features at different tree levels. Figure 3.14 (a) shows the frequency of the selected features at different tree levels using the classic RF. Note that Unary features are the most frequently occurring while 3-blocks Haar3D features are the less. On the other hand, Figure 3.14 (b) shows the frequency of the selected features using the SFRF/SFWRF in which Binary3D is the most dominant feature set and similarly the 3-blocks Haar3D feature set is the least.

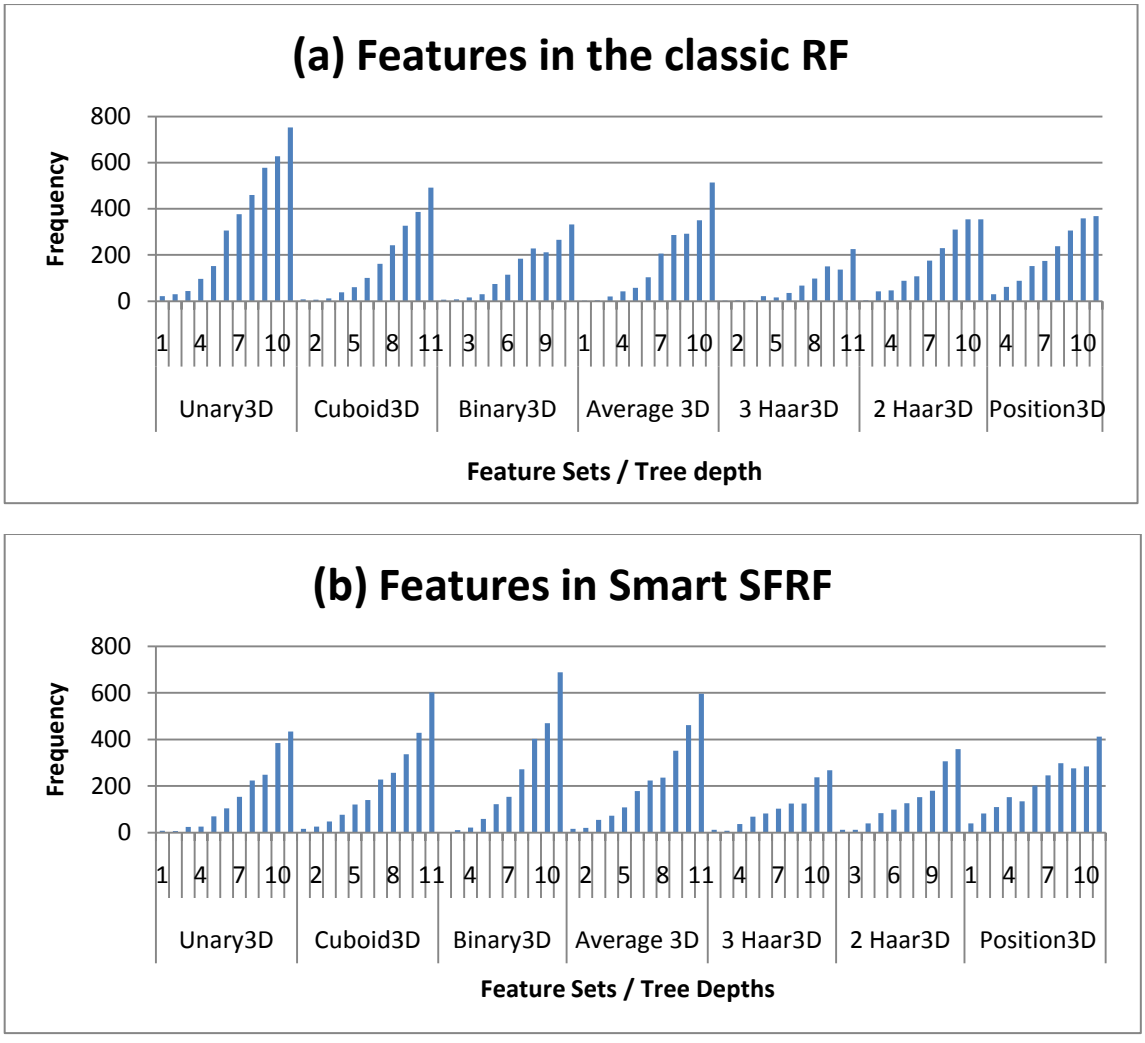


Figure 3.14. Histogram of the selected features at different tree levels within the forest. (a) shows the selected features in the classic RF while (b) shows the selected features in the Smart RF.

3.4.1.3 Testing Performance in the Proposed RF

Since we use 10 volumes in the offline feature selection step, 20 to train on and 21 to test, it is important to investigate how the segmentation accuracy varies if we change the training examples for both the offline feature selection and the actual training of the classifier. We explore this issue we have performed 3-fold cross validation on the 30 volumes used in the offline feature selection and the actual training by altering the choice of 10 volumes in the offline feature selection and the choice of 20 volumes in the fast RF training. In the three runs we tested the accuracy of the same 21 volumes. We report in Figure 3.15 the dice similarity measure on the 21 volumes. The figure illustrates that changing the pre-training/training datasets does not significantly affect the testing accuracy as long as the variability in the pre-training and training datasets is large enough. Notice also how the different runs agree as accuracy increases. W believe this is because easier cases represent better quality images and as a result better accuracy.

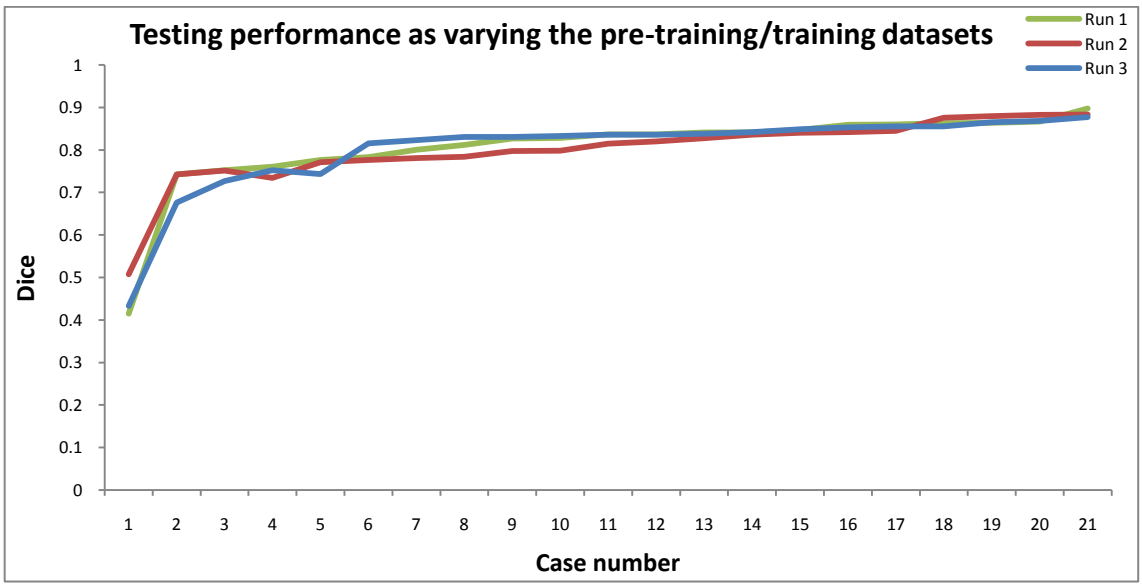


Figure 3.15. Dice similarity on volumes (sorted in ascending order according to Run 1). Each line represents the accuracy of the 21 test volumes by varying the offline feature selection and the actual training volumes.

3.4.1.4 Qualitative Comparison

We performed several comparisons in this experiment to validate the proposed techniques and compare them with the conventional technique. Visual comparison between the classic Random Forests, Smart Fast Random Forests and Smart Fast-Weighted Random Forests is demonstrated in Figure 3.16. From Figure 3.16 (b), the classic RF usually produces inaccurate segmentation and it fails to stop when some femur-like structures are connected to it. SFRF (Figure 3.16 (c)) provides better segmentation accuracy although the epiphyses border was not correctly segmented. SFWRF in Figure 3.16 (d), on the other hand, shows better segmentation accuracy compared to SFRF and superior enhancement over the conventional RF. In addition, Figure 3.17 shows visual segmentation comparison of several 2D slices from the same ultrasound volume.

To study the effect of reducing the size of the feature pool such that it contains the "good" features we visually compare the segmentation on a 2D slice using the classic RF and Smart RF in Figure 3.18. In addition, Figure 3.19 shows the effect of weighted voting by visually comparing the classic RF and the Weighted RF (WRF). Quantitative comparisons are described in detail next.

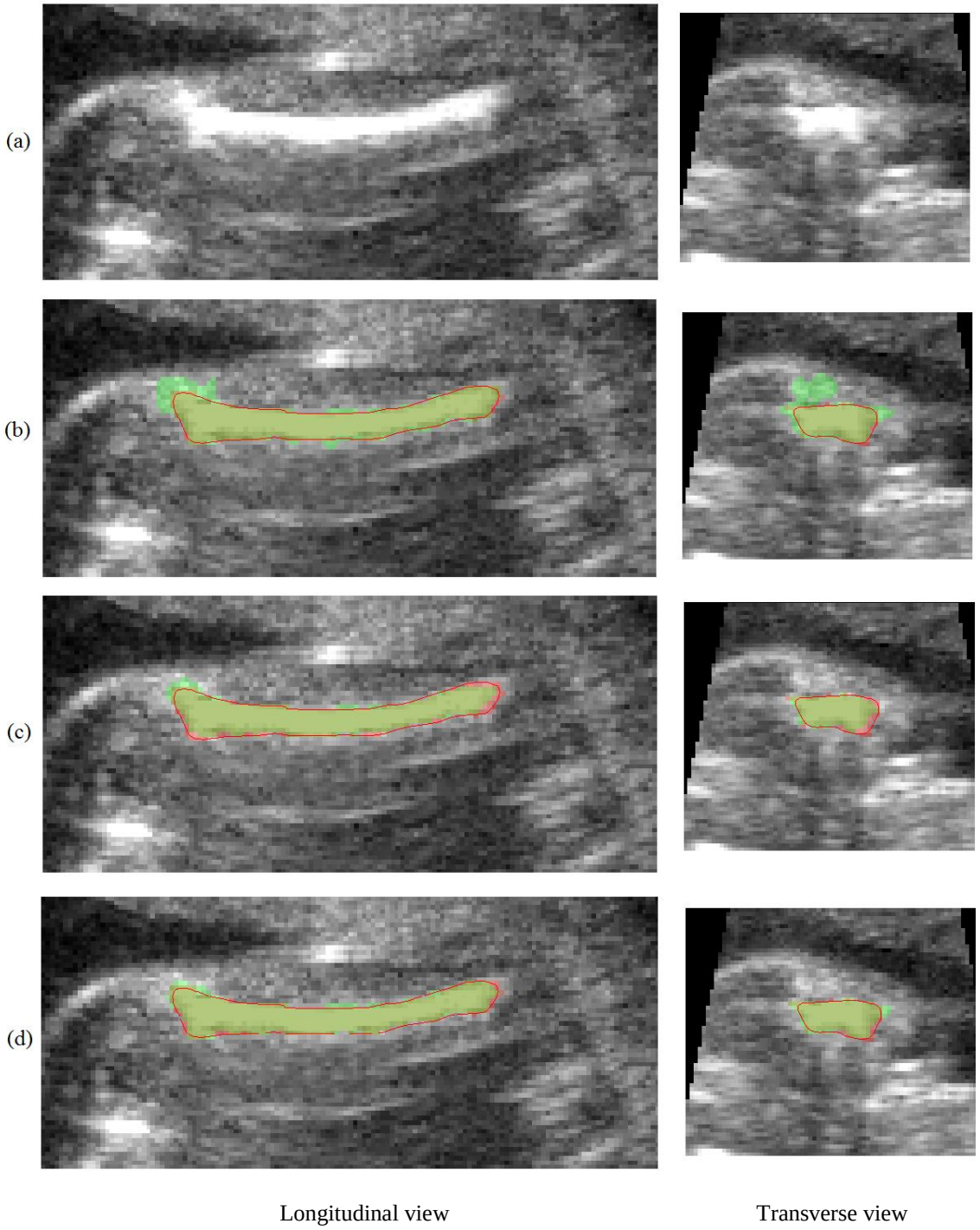
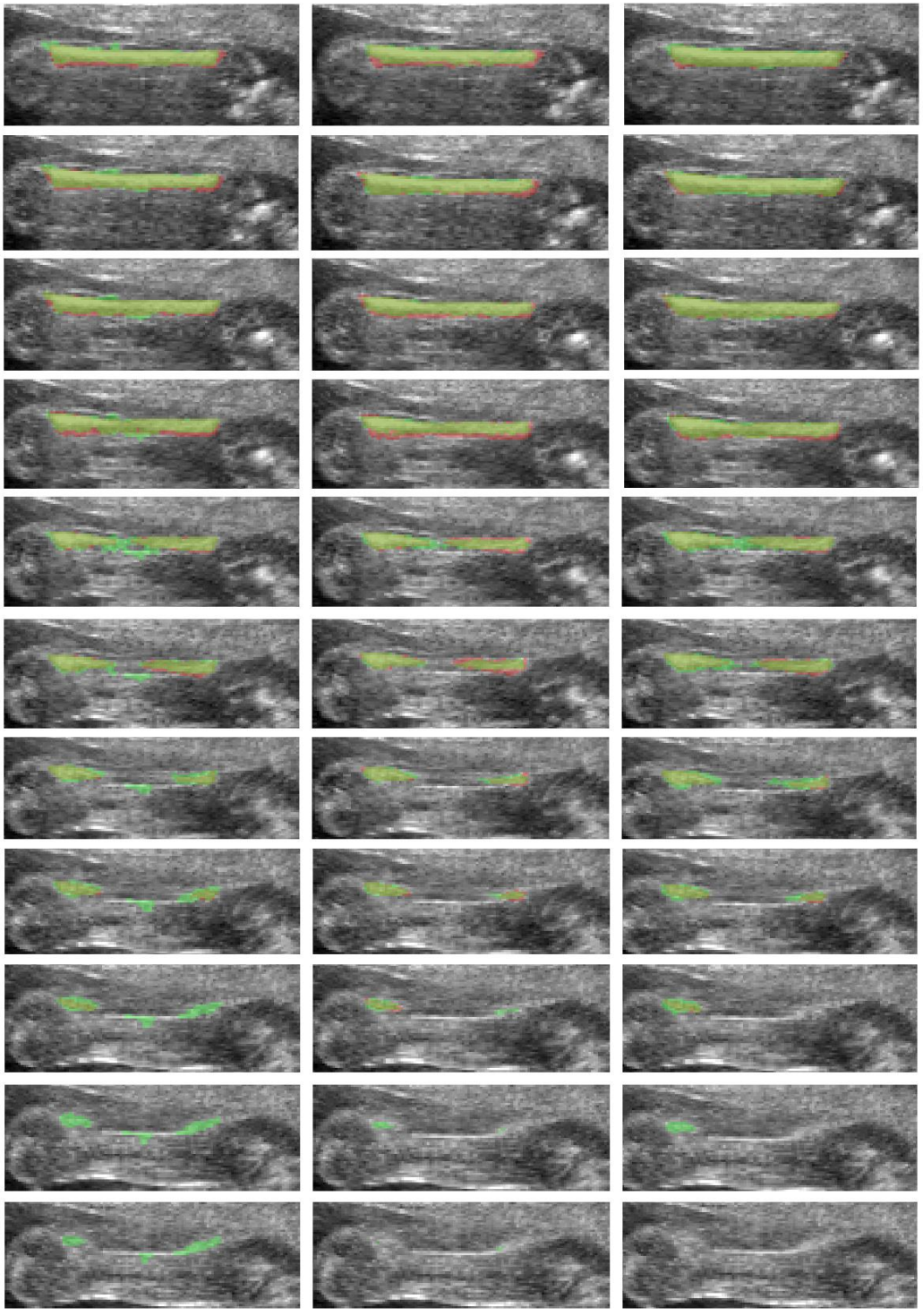


Figure 3.16. Visual segmentation results on an ultrasound fetal femur. (a) shows a 2D ultrasound slice from the 3D volume. (b-d) show the segmentations from classic Random Forests, Smart Fast Random Forests and Smart Fast-Weighted Random Forests respectively. In (b-d), the *red boundary* is an approximation of the manual segmentation; *red* voxels represent false negative which show the voxels that were manually segmented by an expert but were not segmented by the technique; *green* voxels represent false positive which show voxels that were segmented by technique but not by the expert; the yellowish voxels represent the true positive which are the correctly segmented voxels.



(a) The classic RF (b) Smart Fast RF (c) Smart Fast-Weighted RF

Figure 3.17. Visual segmentation results on several slices of a 3D ultrasound fetal femur using (a) the classic Random Forests, (b) the Smart Fast Random Forests and (c) the Smart Fast-Weighted Random Forests. In all 2D slices, *red* voxels represent false negative which show the voxels that were manually segmented by an expert but were not segmented by the technique; *green* voxels represent false positive which show voxels that were segmented by technique but not by the expert; the yellowish voxels represent the true positive which are the correctly segmented voxels.

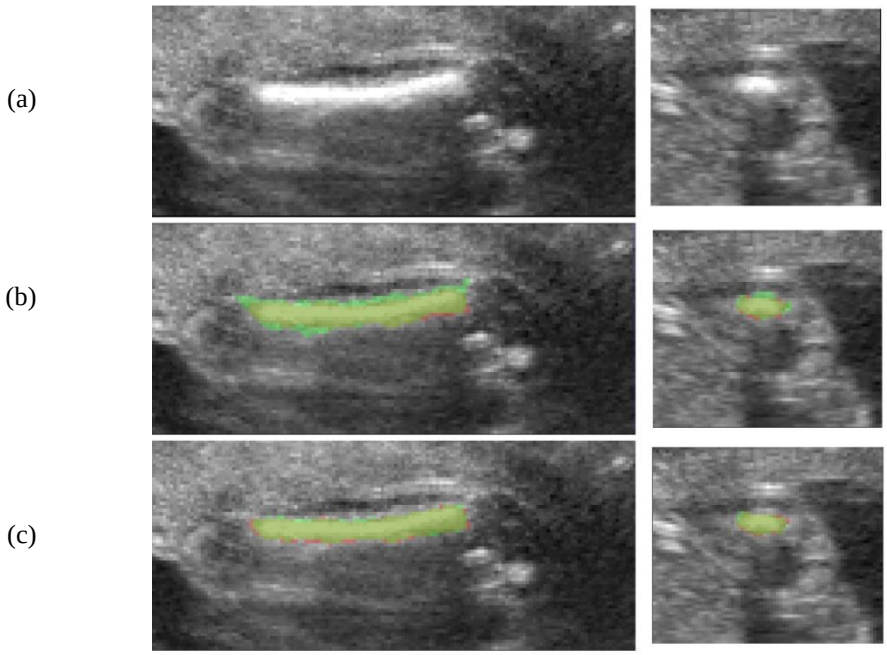


Figure 3.18. The effect of the offline feature selection. Visual segmentation results on an ultrasound fetal femur. (a) shows a 2D ultrasound slice from the 3D volume. (b & c) show the segmentations from classic Random Forests and Smart Random Forests respectively. See the previous figure for a description of colours.

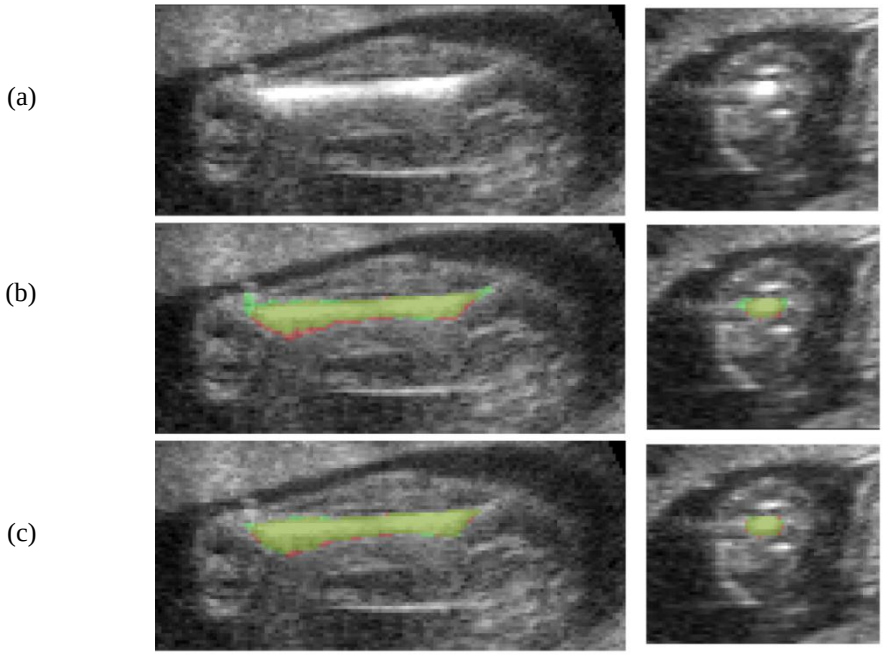


Figure 3.19. The effect of the weighted voting. Visual segmentation results on an ultrasound fetal femur. (a) shows a 2D ultrasound slice from the 3D volume. (b & c) show the segmentations from classic Random Forests and Weighted Random Forests respectively. See the previous figure for a description of colours.

3.4.1.5 Volume Comparison

Volume comparisons between manually segmented femoral volumes and automatically segmented ones are shown in Figure 3.20. Conventional Random Forests shows poor correlation with the manual segmentation while Smart Fast Random Forests shows a slight improvement over the classic Random Forests. On the other hand, Smart Random Forests and Smart Fast-weighted Random Forests are almost the same and they both show superior correlations compared to the other two methods.

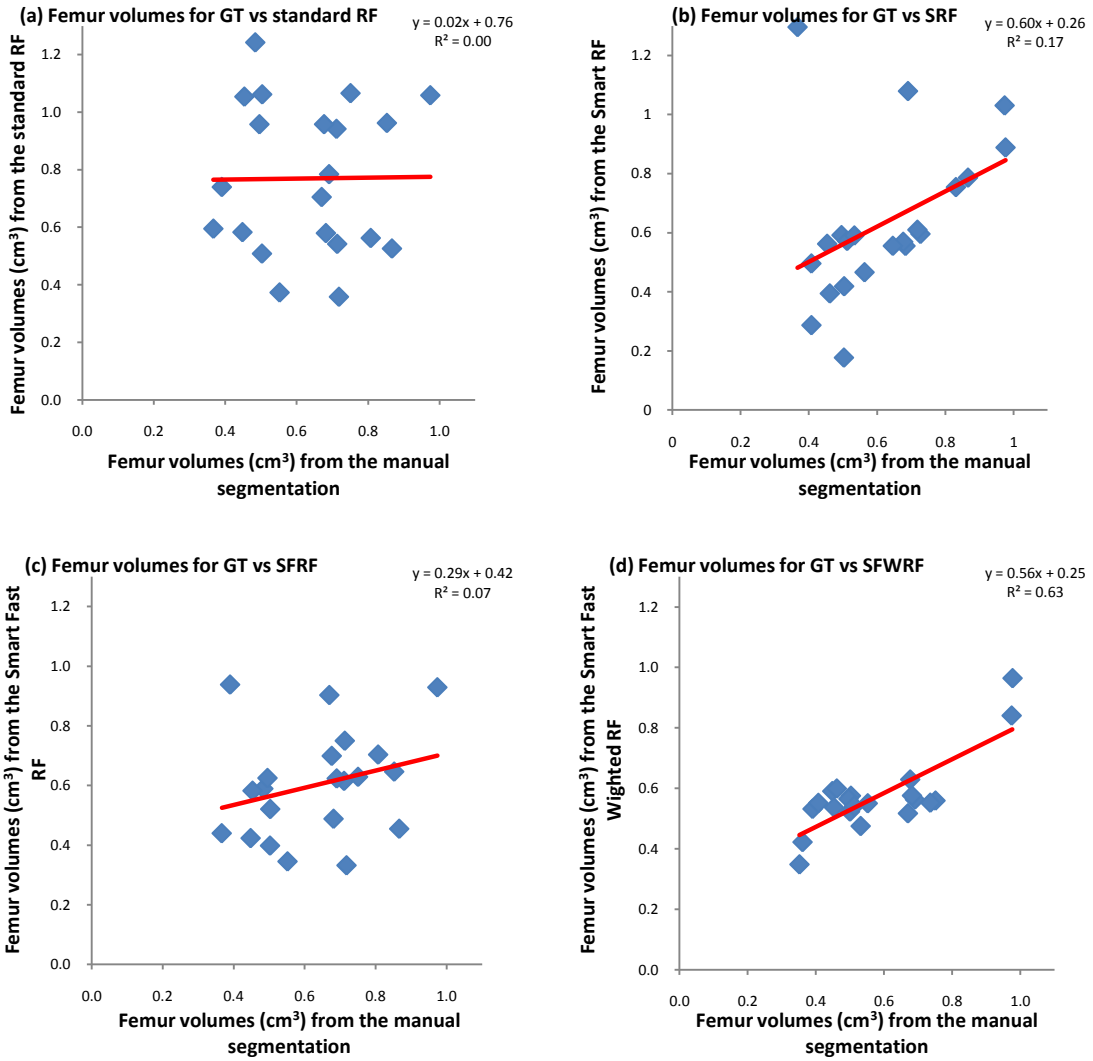


Figure 3.20. Linear regression analysis of 21 femoral volumes between the manually segmented (GT) and (a) the conventional Random Forests (RF), (b) the smart conventional Random Forests (SRF), (c) the smart fast Random Forests with even vote (SFRF) and (d) the smart fast-weighted Random Forests (SFWRF).

3.4.1.6 Quantitative Comparison

The mean and standard deviation of precision, recall, dice coefficient and Jaccard index from all testing cases are reported in Table 3.2. In such experiments, a better measure is usually captured by a mean close to one and a small standard deviation. We compare in Table 3.2 the classic Random Forests (RF), the classic Random Forests with Entropy-based weighting Random Forests (Entropy weighted RF, see equation (3.7)), the classic Random Forests with weighted voting in testing (WRF), the Smart Fast Random Forests (SFRF), the Smart classic Random Forests (SRF), the Smart Fast-Weighted Random Forests (SFWRF) and the Smart weighted Random Forests (SWRF). The listing in Table 3.2 is sorted in ascending order by the Jaccard index. Note that the SWRF is the best with respect to precision, dice and Jaccard indexes but with very slight improvement over the SFWRF. However, SFWRF has the advantage of much faster training over the SWRF. On the other hand, the mean recall of the conventional RF is better than in all proposed RFs. This is mainly because conventional Random Forests usually over-segment the femur such that the intersection between Random Forests segmentation and the manual one is nearly equal to the manual segmentation. Therefore, according to equation (3.9) the difference between the numerator and the denominator are small which makes the recall close to 1. Notice also how the proposed weighted voting is slightly better than the entropy-based weighting. As we mentioned, SWRF is the most accurate among the remaining because it simply 1) uses the reduced feature pool of "good" features, 2) retrain a node the selected good features, and 3) weight trees decision based on the strength of each path in each tree.

Finally, offline feature selection, training and testing times are reported in Table 3.3. Testing time is reported for segmenting one volume. Although the classic Random

Forests classifier does not require an offline feature selection step, its training time is significantly higher than in the proposed fast RFs. On the other hand, all proposed weighted Random Forests requires slightly more testing time than the classic Random Forests since it computes trees' weights for each test voxel.

Table 3.2. The mean ± standard deviation of precision, recall, dice similarity and Jaccard index over 21 3D fetal femur ultrasound cases. Comparisons are shown for classic Random Forests, Entropy-weighted Random Forests, Weighted Random Forests, Smart Fast Random Forests, Smart Random Forests, Smart Fast-Weighted Random Forests and Smart Weighted Random Forests.

	Precision	Recall	Dice	Jaccard
RF	0.68±0.18	0.89±0.11	0.75±0.11	0.61±0.13
Entropy weighted RF	0.79±0.13	0.77±0.13	0.76±0.11	0.63±0.12
WRF	0.78±0.12	0.79±0.13	0.77±0.11	0.64±0.13
SFRF	0.79±0.14	0.81±0.11	0.78±0.08	0.65±0.10
SRF	0.82±0.11	0.80±0.1	0.80±0.09	0.68±0.11
SFWRF	0.83±0.09	0.82±0.08	0.81±0.04	0.70±0.06
SWRF	0.85±0.05	0.81±0.09	0.82±0.07	0.71±0.08

Table 3.3. Times to compute Feature scores, training and testing for the classic Random Forests, Entropy-weighted Random Forests, Weighted Random Forests, Smart Fast Random Forests, Smart Random Forests, Smart Fast-Weighted Random Forests and Smart Weighted Random Forests on the 3D fetal femur ultrasound dataset.

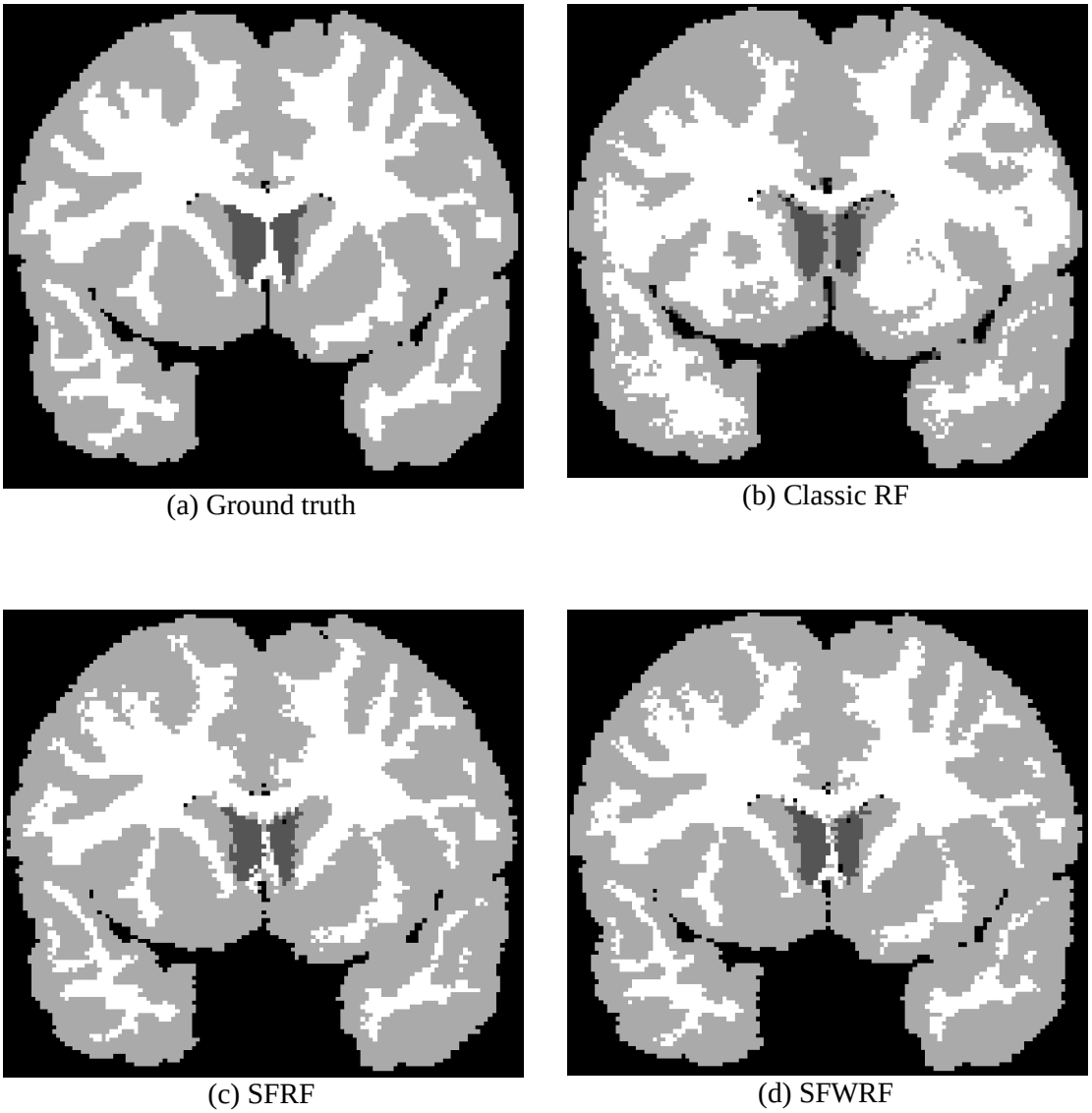
	Offline feature selection	Training	Testing
RF	0	16 hours	1.3 second
Entropy weighted RF	0	16 hours	1.8 second
WRF	0	16 hours	2.2 second
SFRF	11 hours	95 minutes	1.5 second
SRF	11 hours	16 hours	1.3 second
SFWRF	11 hours	95 minutes	2.3 second
SWRF	11 hours	16 hours	2.3 second

3.4.2 Brain Structures Segmentation in MR Images

Visual comparison between RF, SFRF and SFWRF to segment brain MRI is shown in Figure 3.21. Note how the classic Random Forests failed to segment a few regions accurately while both the smart fast Random Forests and the smart fast-weighted Random Forests outperformed the classic one. Slight enhancement from the smart fast Random Forests to the smart fast-weighted Random Forests can be visually seen. On the other hand, Table 3.5 shows the Jaccard indices from several segmentation algorithms. Note that our implementation of the classic Random Forests versus the one implemented in (Yi et al., 2009) give slightly lower Jaccard index although the same dataset was used. The reason for this is that we could not reproduce Yi et al. (2009) exactly as key parameters were not defined in the manuscript. On the other hand, the last two rows in the table show better Jaccard indices over the classic Random Forests. Finally, Table 3.4 shows the offline feature selection, training and testing times. Testing time is reported for segmenting one volume. Although SFWRF requires the most testing time, it is comparable to the classic Random Forests and the fast Random Forests with equal voting.

Table 3.4. Times to compute Feature scores, training and testing for the classic Random Forests, smart fast Random Forests and smart fast-weighted Random Forests on the brain MRI dataset.

	Offline feature selection	Training	Testing
RF	0	78 hours	11 seconds
SFRF	26 hours	6 hours	15 seconds
SFWRF	26 hours	6 hours	23 seconds



(a) Ground truth (b) Classic RF

(c) SFRF (d) SFWRF

Figure 3.21. Typical segmentation results on a brain MRI. (a) shows a 2D manually segmented slice from the 3D volume. (b) shows the segmentation using the classic Random Forests. (c) shows the segmentation using the Smart Fast Random Forests with equal voting. (d) shows the segmentation using the Smart Fast weighted Random Forests. (a-d) *Black*, *dark gray*, *gray* & *white* represent background, CSF, GM, and WM, respectively.

Table 3.5. Comparison between the proposed and several state of the art techniques. This table is based on (Yi et al., 2009). Mean Jaccard indices are shown for segmenting CSF, GM and WM.

Method	CSF	GM	WM
Adaptive MAP	0.069	0.564	0.567
Biased MAP	0.071	0.558	0.562
Fuzzy c-means	0.048	0.473	0.567
Maximum-a-posteriori (MAP)	0.071	0.550	0.554
Maximum-likelihood	0.062	0.535	0.551
Tree-Structure k-means	0.049	0.477	0.571
MPM-MAP	0.227	0.662	0.683
BSE/BFC/PVC	—	0.595	0.664
Constrained GMM	—	0.680	0.660
Spatial-varying GMM	—	0.768	0.734
Coupled surface	—	0.701	—
FSL (Zhang et al., 2001)	—	0.756	—
SPM (Ashburner and Friston, 1997)	—	0.790	—
MAP with histograms (Yi et al., 2009)	0.549	0.814	0.710
Classic Random Forests (Yi et al., 2009)	0.614	0.838	0.731
Classic Random Forests	0.604±0.08	0.811±0.03	0.714± 0.04
Smart Fast Random Forests	0.635±0.07	0.840±0.03	0.762±0.05
Smart Fast-weighted Random Forests	0.648±0.08	0.851±0.03	0.784±0.05

3.5 Discussion

We have shown that smart fast/non-fast-weighted Random Forests outperforms the classic Random Forests as well as many state of the art techniques, see Table 3.2 and Table 3.5. Three main contributions enhanced the classification accuracy over the classic Random Forests. First, offline feature selection helped choose a smaller feature pool which contains mainly "good" features. Second, relying on the pre-computed feature scores helped significantly speed up the training time while keep the testing accuracy almost similar. Third, trees' weighted voting produced a more accurate probabilistic decision since each tree contributes unevenly to the final decision depending to its strength to classify an unseen example .

The proposed Random Forests has proven its applicability to segment different objects of interest in medical images. Moreover, we validated its performance on two image modalities. Although classic Random Forests can be implemented in an efficient and parallel way, training is still time consuming. Therefore, we not only showed better classification accuracy in the smart fast-weighted Random Forests but also much faster training. Slight improvement in classification accuracy can be achieved in the SFWRF if the classic training happens at each tree node (SWRF) but training time will be significantly large. This is critical in some cases where training has to be done several times while tuning the main parameters. In addition, in cross validation evaluation methods, training has to be repeated. On the other hand, the proposed technique requires a pre-training step in which all features gather a global score. This step is time-consuming and an extra overhead compared to the classic Random Forests. Nevertheless, finding a global score for each feature is totally independent which results of a near-real time feature selection if computational power exists. Furthermore, because machine learning techniques rely on a gold standard training set, the resultant classifier is tightly related to a quality of the training examples. In our study, manual segmentation from only one expert was used to train our classifiers although the use of several manual segmentation is preferable.

Finally, an extension to this work could be by enriching the feature pool with more features that can characterise different medical data accurately, e.g., HOG (Dalal and Triggs, 2005), local phase (Rajpoot et al., 2009a), etc. In addition, more interest is drawn recently in hybrid discriminative and generative models, e.g., (Tu et al., 2008), which makes our proposed technique a good candidate to combine with a generative model.

Chapter 4

Evaluating Image Fusion for Improved Fetal Femoral Boundary Definition and Quantification

In Chapter 2, we discussed the problems concerning ultrasound images with the focus of imaging bony structures like the fetal femur. Although the quality of 3D and 4D ultrasound imaging continues to improve, it does not compare with CT or MRI in terms of anatomical definition. Ultrasound imaging of fetal bony structures is limited by significant acoustic shadowing and signal drop out. As a result, three-dimensional imaging and volumetric measurement of skeletal structures has unique challenges. For automatic volumetric quantification and diagnosis there is a clear need for a novel methodology which maximises the anatomical definition obtained from one or more ultrasound scans. In this chapter, we propose an automatic wavelet-based 3D image fusion technique to combine multiple ultrasound images taken from different angles of the fetal femur. The material properties of the femoral tissues result in high attenuation

of parts of the femur in a single scan. The method enhances the femoral boundary definition and provides a better anatomical image of the femur.

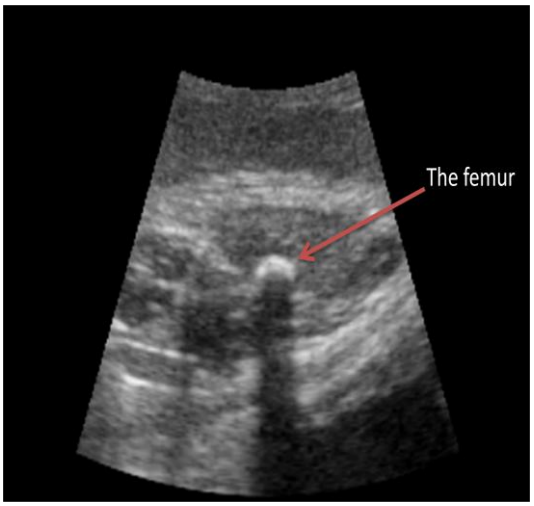
A literature review of current medical image fusion techniques with the focus on how ultrasound image fusion compensates for missing structures can be found in Section 2.6. This chapter is organised as follows. We first motivate looking at image fusion in section 4.1. We then illustrate in section 4.2 the method of image alignment and fusion. In section 4.3, we describe the experimental setup to evaluate the approach. Section 4.4 shows the qualitative and quantitative results. Finally, a brief discussion is provided in section 4.5. Early versions of this work have been published as (Yaqub et al., 2010a) and (Yaqub et al., 2011a).

4.1 Introduction

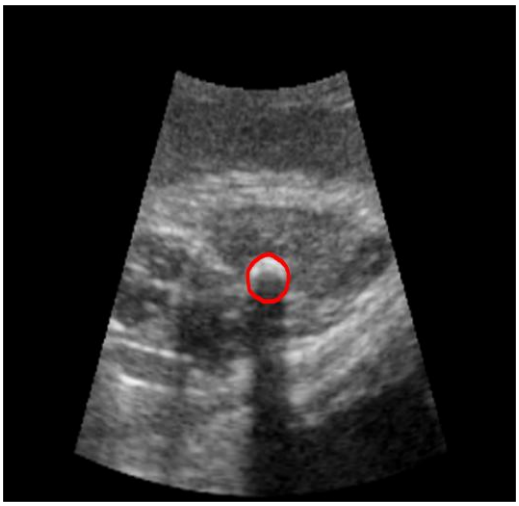
In Chapter 2, we have showed how ultrasound imaging fails to completely image bony structures because of the high reflection of an ultrasound beam when it faces bone. As a fetus grows, the bone becomes more calcified which increases signal attenuation and hence lower parts of the bone with respect to the beam direction are partially invisible in ultrasound images. Although both femoral epiphysis are less calcified than the mid-shaft bone, they tend to be partially invisible. This is mainly because they are relatively further away from the scan signal than the mid-shaft during image acquisition. Researchers (Chang et al., 2007, Mahon, 2007, Mahon et al., 2009, Ioannou et al., 2010, Yaqub et al., 2010a, Yaqub et al., 2010b, Yaqub et al., 2011a) have tried to manually segment the fetal femur in 3D ultrasound images but an assumption of what is considered part of the bone has to be made clear beforehand. For instance, (Chang et al., 2007, Mahon, 2007, Mahon et al., 2009) have manually segmented the femur on 2D

axial slices. They have segmented the lower part of the femur by making an assumption that the femur has a cylindrical shape. They assume that a reflection of the lower part of the femur should appear on a few slices in the axial view therefore a semi-circular shape was assumed to aid manual segmentation of each 2D axial slice. This assumption significantly increases the segmented area on a 2D slice especially at the distal and proximal epiphyses. An example of using this assumption is depicted in Figure 4.1 (b). This assumption is subjective and (Mahon, 2007, Mahon et al., 2009) showed low intra-observer reproducibility when measuring the femur volume. The main reason for this subjectivity is the absence of edge information at the bottom border of the femur especially at the femur epiphyses. On the other hand, Yaqub et al. (2010b) made a different assumption which is to segment what is visible by eye. This assumption is simple and more likely to be more reproducible than the first assumption although reproducibility issues were not investigated in that work. Figure 4.1 (c) shows a pictorial example of this type of manual segmentation. Another way of segmenting the femur was proposed by Ioannou et al., (2010) in which they draw a straight line directly after they segment the upper part, Figure 4.1 (d). This method showed good inter and intra-observer reproducibility. We use this protocol to create manual segmentations to validate our work in this chapter.

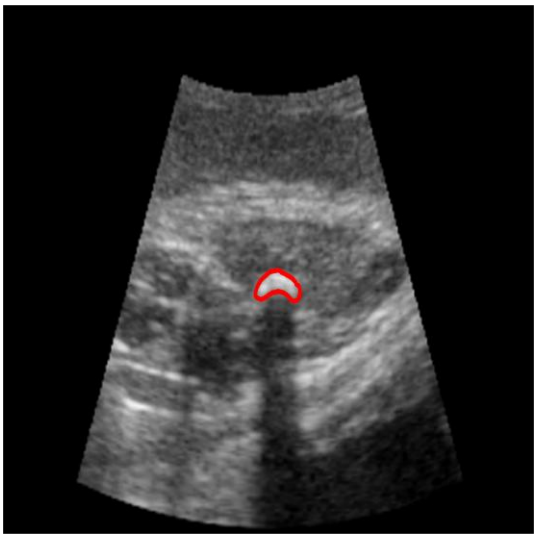
The discussion above highlights that there is no clear way of measuring the fetal femoral volume. The main reason for this is the partially invisible part of the femur which is caused by the reflection of ultrasound beams when they hit the upper part of the femur. As a result, developing a method that accurately compensates for the missing part of the femur to enable more realistic measurements is important.



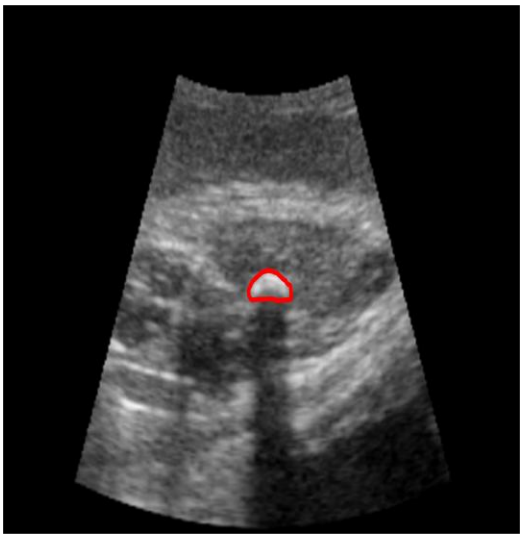
(a) 2D axial slice



(b) Manual segmentation as in (Mahon, 2007, Mahon et al., 2009)



(c) Manual segmentation as in (Yaqub et al., 2010b)



(d) Manual segmentation as in (Ioannou et al., 2010, Yaqub et al., 2010a, Yaqub et al., 2011a)

Figure 4.1. Different manual segmentation assumptions. (a) Original 2D axial slice. The result of manually segmenting the image in (a) according to (b) draw a semi-circular shape (Chang et al., 2007, Mahon, 2007, Mahon et al., 2009), (c) segment what you see (Yaqub et al., 2010b), and (d) make a straight line for the invisible region after segmenting the upper part (Ioannou et al., 2010, Yaqub et al., 2010a, Yaqub et al., 2011a).

In this chapter, we hypothesise that image enhancement of 3D ultrasound is important to improve quantification, measurement and thus diagnosis. Therefore, we propose an image fusion technique to combine multiple ultrasound volumes of the fetal femur acquired from several angles. The technique aims to successfully align the

volumes and merge them to provide better anatomical definition in the reconstructed volume. We call each volume used in the fusion a single view or scan and the resultant volume a fused view.

The technique has two stages. First, image registration is used to align the multiple ultrasound views of the femur. Second, image fusion is used to merge the aligned images. In this chapter, we validate the technique on in vivo and in vitro femur datasets and show how fused images provide better boundary definition of the object of interest. We also show that image fusion improves intra and inter-observer reproducibility of volume measurement compared with manual femur segmentation. We compare two other fusion techniques (Max and Mean intensity fusion) to demonstrate how wavelet-based fusion preserves important information and eliminates irrelevant detail.

4.2 Method

4.2.1 Image Registration

As the femur is a rigid body it is valid to use a 3D rigid image registration algorithm to align fetal femur scans for the same fetus. While aligning several images to a one coordinate space, one of them has to be selected to represent the reference image to which the remaining images (floating images) will be aligned to. In this work we follow a similar approach to (Grau et al., 2007, Rajpoot, 2009). The optimisation problem can be formulated as follows

$$\arg\max_{T} S(T) = S\left(I_r, T(I_f^i)\right), \tag{4.1}$$

where S is the similarity measure, I_r is the reference image, I_f^i is the ith floating image and T is the transformation function that is used to map I_f^i into I_r coordinate space.

The optimisation was performed between I_r (usually the first ultrasound volume) and I_f^i (the remaining ultrasound volumes) for every case. This process is time consuming but can be performed simultaneously. We used the Normalised Cross Correlation (NCC) as a similarity measure as in (Grau et al., 2007, Rajpoot, 2009). Although other similarity measures can be used, NCC has experimentally succeeded to align most cases and it has worked well on our datasets. To find the global maxima of the similarity measure, a multi-resolution approach with multiple initialisations was used. This method has been demonstrated to avoid local maxima (Grau et al., 2007, Rajpoot, 2009). The Powell optimiser was used to maximise the similarity criteria. The main values of parameters used during optimisation were 1) number of iterations=10, 2) step length=0.05 and 3) tolerance value=0.0001. In addition, we used tri-linear interpolation to re-sample the transformed voxels.

4.2.2 Wavelet-based Image Fusion

Having aligned images, fusion combines intensities from multiple scans. Ultrasound images contain a high amount of speckle and can have a weak boundary definition. We have modified the wavelet-based fusion proposed in (Rajpoot, 2009) for fusing the fetal femur. The novelty lies in the way that the frequency components are manipulated after wavelet decomposition. In addition, we have thoroughly validated the method.

The steps in wavelet-based fusion are illustrated in Figure 4.2. The 3D Discrete Wavelet Transform (DWT) is applied to each 3D ultrasound volume to get 8 frequency components from each volume. Figure 4.3 shows an example of wavelet decomposition of the eight components. The low frequency component is an approximation of the original image which is a down-sampled intensity component. The remaining seven are

high frequency components in different orientations. The first high frequency component captures the horizontal intensity variations in the image. Therefore, retaining the first high frequency component is important because it contains useful information since the femur is mainly straight (and horizontal) in all aligned images. In particular, this suggest to maximise the low and the first high frequency (horizontal) components; i.e., MAX(low_r, $low^1{}_f$, …, $low^N{}_f$) and MAX($high(1)_r$, $high(1)^1{}_f$, …, $high(1)^N{}_f$) where r is the reference image and $^i{}_f$ is the i_{th} floating image. On the other hand, the other six high frequency components mainly contain speckle and/or other non-horizontal detail and tissue artefacts. Therefore, we average the remaining 6 non-horizontal high frequencies. After finding MAX_{low}, MAX_{high1} and $AVG_{high2\text{-}7}$, we apply the inverse 3D-DWT to get the fused 3D image reconstruction.

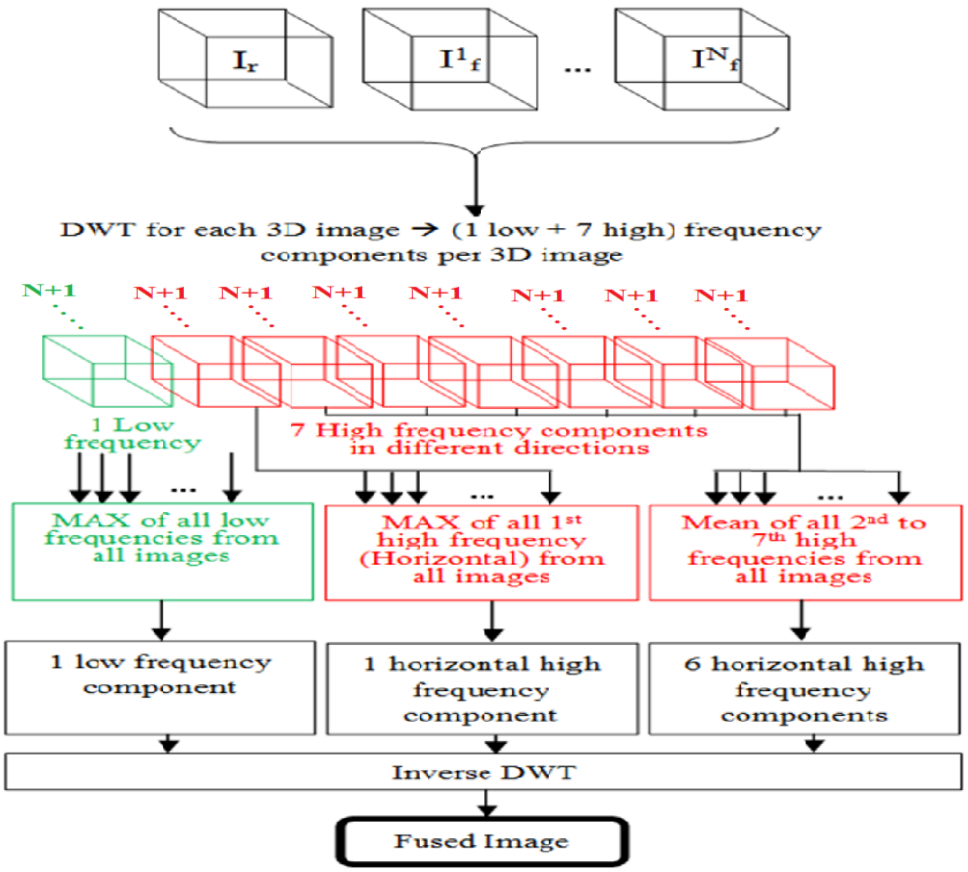


Figure 4.2. Framework of the wavelet-based image fusion.

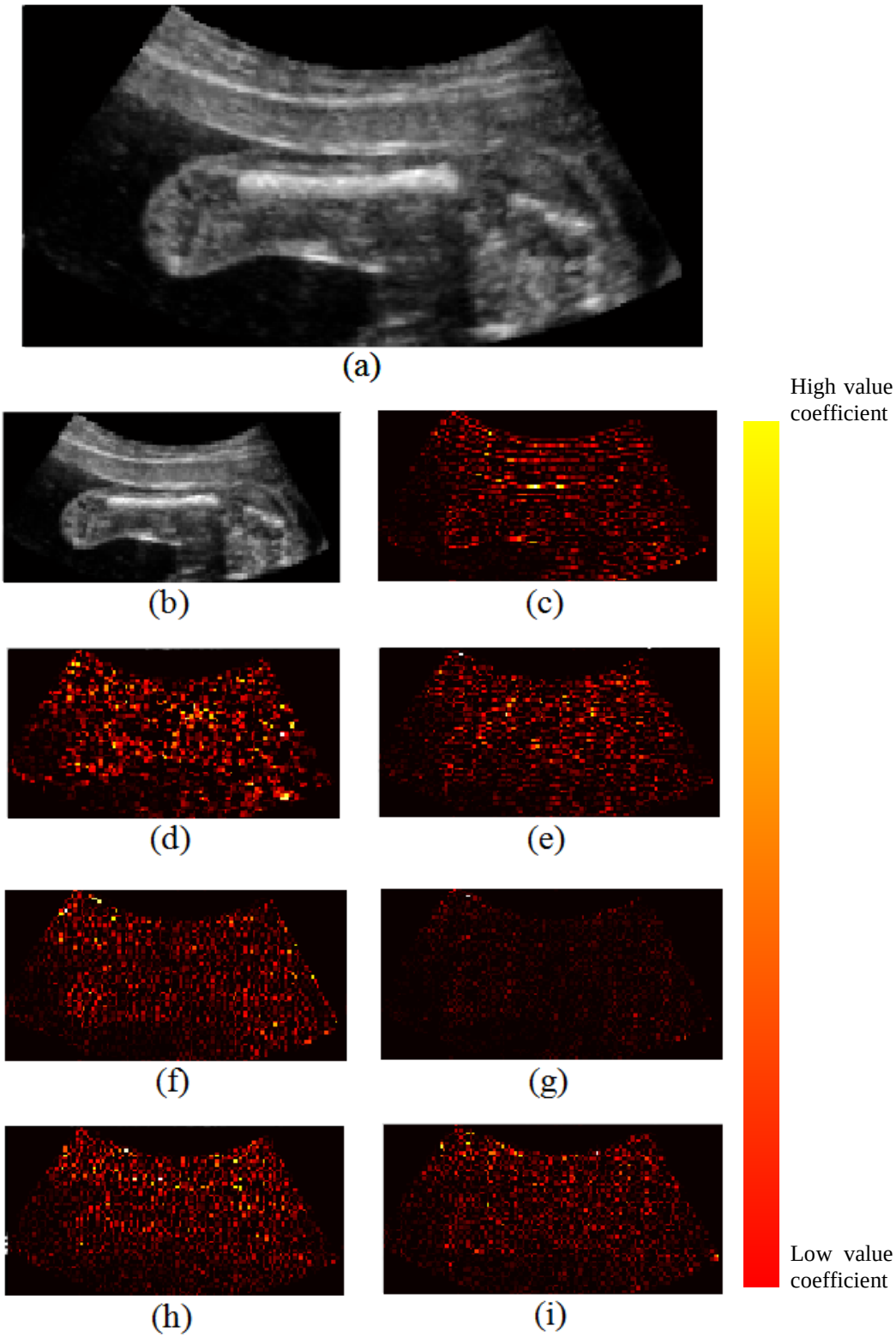


Figure 4.3. Example of a 3D wavelet decomposition of the fetal femur. Only 2D slice (a) is shown with its corresponding one level 8 wavelet decomposition components. (b) is the low frequency (approximation). (c) is the first high frequency component (horizontal details). (d - i) are the remaining six high frequency components. We colour the high frequency components for visualisation purpose. Colours represents the value of wavelet coefficients (Yellow means high, red

There are other ways to combine aligned images including Mean intensity, Max intensity, Median and Mean-Shift intensity (Rajpoot, 2009, Rajpoot et al., 2009b, Gooding et al., 2010). More details about such techniques can be found in Section 2.6.

4.2.3 Implementation Details

The rigid registration was coded in C++ with the use of the Insight ToolKit package (www.itk.org, Ibanez et al., 2005). The fusion code was developed using Matlab (version R2008a Natick, Massachusetts: The MathWorks Inc., 2008). The code runs on a 2.8 GHz quad core PC with 8GB of RAM. The registration between two 3D ultrasound images on average took three to four minutes. On the other hand, fusion required approximately five seconds to merge four aligned images. The registration time is high but this is because it is a multi-resolution, multi-initialisation algorithm. In general, registration between different views is independent and hence we have done this simultaneously.

4.3 Experiments

The main idea so far is to acquire, align and fuse multiple scans for an object that is partially attenuated. The hypothesis is to recover the correct volume of the object of interest. On the other hand, validation of this technique is not straightforward because of the absence of a gold standard. For instance, fusing multiple ultrasound scans for a fetal femur will generate a reconstructed volume of the femur. Comparing the reconstructed femur volume to the exact actual volume of the fetal femur is not possible unless a CT scan is performed which is generally prohibited during pregnancy. MRI could be another option but it is mainly used to image soft tissues and it requires the fetal thigh to be fixed to acquire a 3D volume of it which is a hard constraint to satisfy because of

fetal movement. Moreover, MRI is mainly used to image 20 weeks or elder fetuses where less movement is happening. In this work, we have validated our technique first using a commercially available fetal phantom however we found that it hardly mimics the real fetal femur. Second we attempted to build animal phantoms but with limited success. Third, in vivo ultrasound volumes were used to confirm our hypothesis. Finally, a post-mortem case where femoral ultrasound volumes and a CT scan of the leg were acquired and used to validate the soundness of the alignment and fusion.

4.3.1 Datasets

In vitro and in vivo data were used to validate the proposed technique. In the in vitro case, the femur from a fetal phantom, a chicken and rats were scanned (see details below). In the in vivo case, scans from healthy fetuses and one post-mortem case were used for validation. In the first in vivo study, fetal scans were obtained from 25 healthy pregnant women following informed consent and institutional ethics committee approval (see Appendix A for ethics approval). In the second in vivo study, 6 fetal femoral ultrasound scans were acquired on the femur of a stillborn fetus. A CT scan was also taken of the fetal thigh after termination. All ultrasound scans were acquired using a Philips HD9 scanner with a mechanical V7-3 transducer (Philips Healthcare, Bothell, Washington, USA). The CT scans were acquired on a LightSpeed™ VCT scanner (GE Healthcare).

4.3.1.1 Human Fetal Phantom

A fetal phantom (CIRS Model 068 Fetal Ultrasound Biometrics Phantom, CIRS, Norfolk, Virginia, USA) was used to study the effect of fusion on the femur. The phantom was scanned from 10 different angles. Although the entire femur is visible in

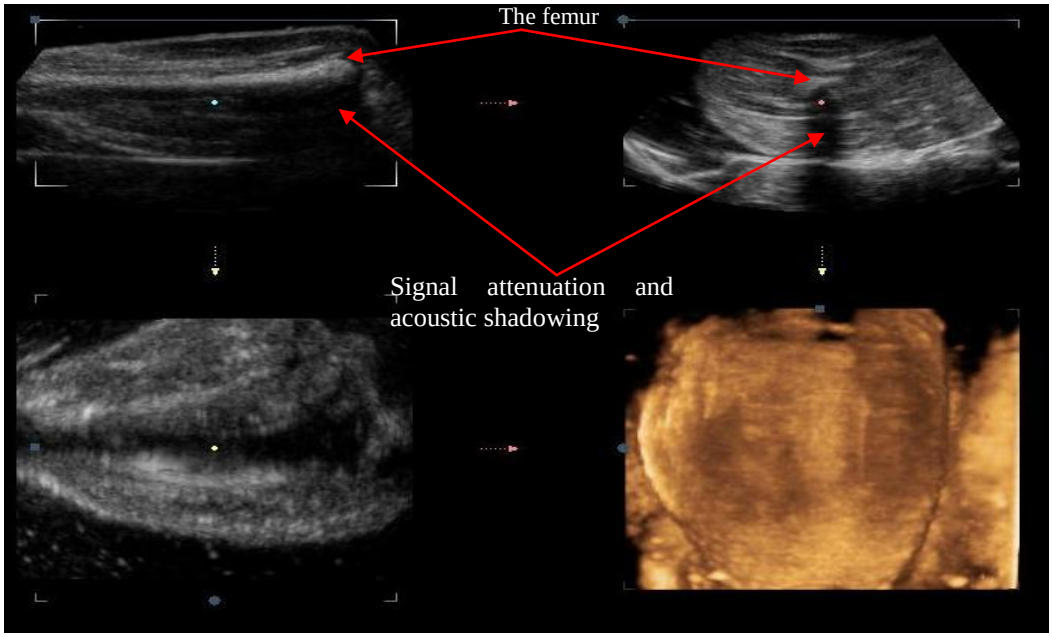
one scan, see Figure 4.9 (a), the border of the femur is not clear. Moreover, contrast and SNR are low in the single scan compared to the fused image, see Figure 4.9 (b).

4.3.1.2 Animal Phantoms

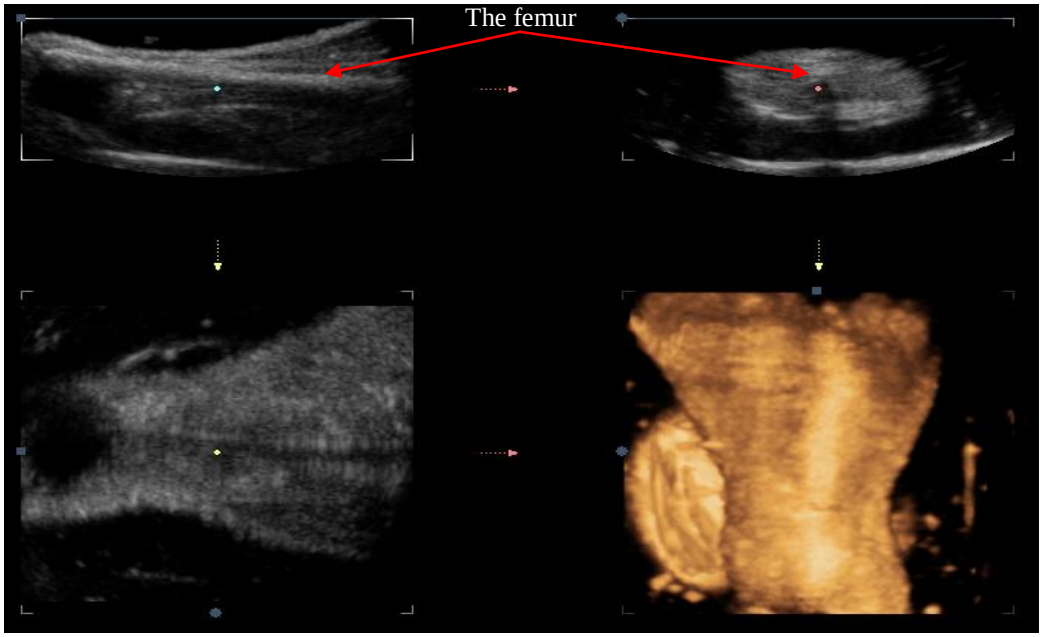
The main goal of building an animal phantom is to validate the fusion technique by comparing the segmented volumes from the single views and the fused view with a femoral physical volume. The real volume of the femur can be measured using water displacement or manually segmenting the femur from a CT scan of the phantom. The hypothesis is that measuring femoral volume on a fused view is closer to the real femoral volume than measuring it using a single view scan. Therefore, three animal femur phantoms were made. The femurs were taken from a chicken and two rats. All phantoms were built following a standard procedure for creating tissue-mimicking phantoms (Hall et al., 1997).

The first phantom was a chicken femur phantom where feathers and skin were removed from a chicken femur, which was then put inside a cooked gelatine. Figure 4.4 shows two 3D ultrasound scans (relatively, good and bad scans) taken from different angles of the chicken femur phantom. In this figure, three 2D views of the chicken femur and a 3D visualisation for each scan are shown for visual comparisons. One can notice in Figure 4.4 (a) in the top right image, for instance, the attenuation of the signal underneath the femur. Unfortunately, the femur was not easily visible in all scans. This is mainly because the chicken femur is highly calcified, hence signal attenuation was high and acoustic shadowing quite prominent. This is by contrast to the fetal femur which is partially calcified and less attenuation and shadowing occur. Moreover, the femur was too long and could not be captured by one sweep. In conclusion, a chicken

femur phantom cannot physically mimic the fetal femur so we could not proceed further with this experiment.



(a) A relatively good image



(b) A relatively bad image

Figure 4.4. 3D ultrasound for the chicken leg phantom. (a) and (b) show two ultrasound scans from different angles. (a) has relatively good quality image compared to (b). In both sub-figures (a) and (b), top left: longitudinal plane, top right: transverse plane, bottom left: coronal plane, and bottom right: 3D visualisation.

The second phantom was a rat femur phantom derived from a one year old male rat which had been sacrificed for another study. The soft tissue was removed from the rat thigh which was then scanned, see Figure 4.5. Note that the top part of the femur is nicely visible but the lower part was severely attenuated. Notice, also from Figure 4.5, that it is hard to assess the lower femoral border because of acoustic shadowing and signal attenuation. The high bone calcification which induced severe attenuation prevented us from correctly quantifying the femur and this suggested that results could be erroneous. We therefore stopped this experiment and proceeded to do the next one.

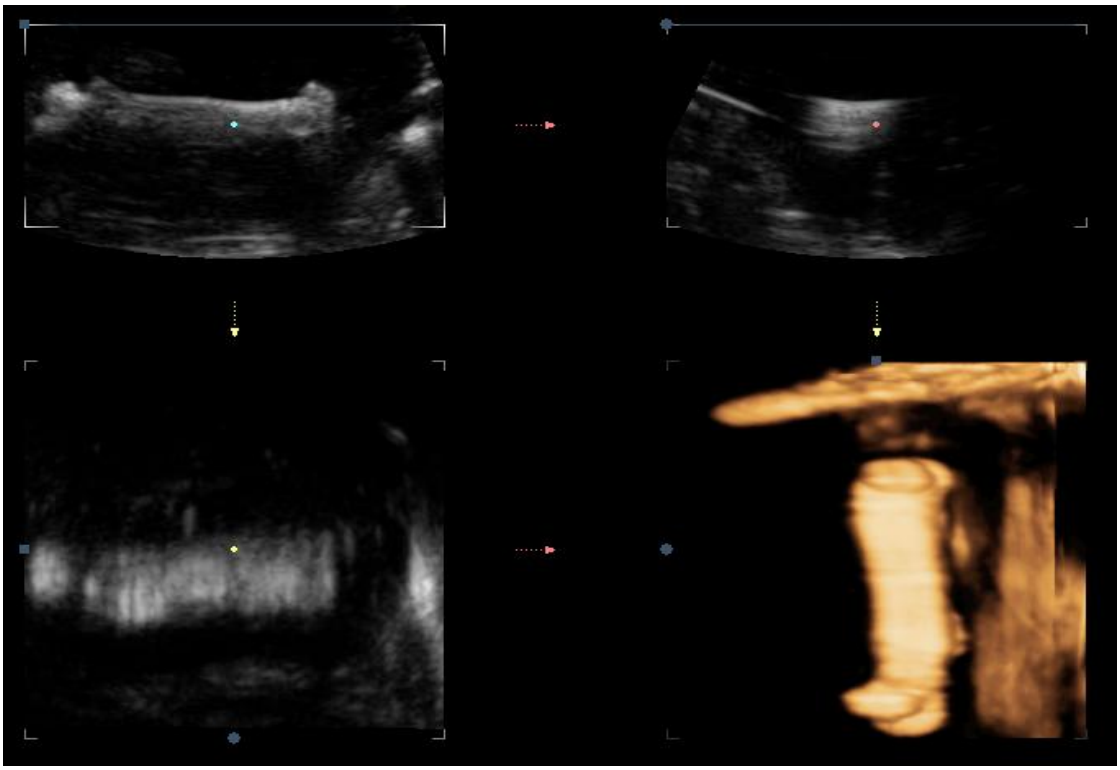


Figure 4.5. Rat femur phantom. Top left: longitudinal plane, top right: transverse plane, bottom left: coronal plane, and bottom right: 3D visualisation. Notice in the top left plane (longitudinal view) how unclear the bottom edge of the femur.

Finally, we hypothesised that using a rat, a few months younger and with soft tissue on the leg would reduce the attenuation. Younger animal have less calcified bone which would provide less attenuation. Therefore we constructed a rat leg phantom that contained the whole leg of a younger rat. The rat in this experiment was a seven months old female. Unfortunately, the images were still found to be of poor image quality, see Figure 4.6. Since none of the phantoms provided good enough quality images, we abandoned validation of fusion using animal phantoms and moved to analysis on real clinical data.

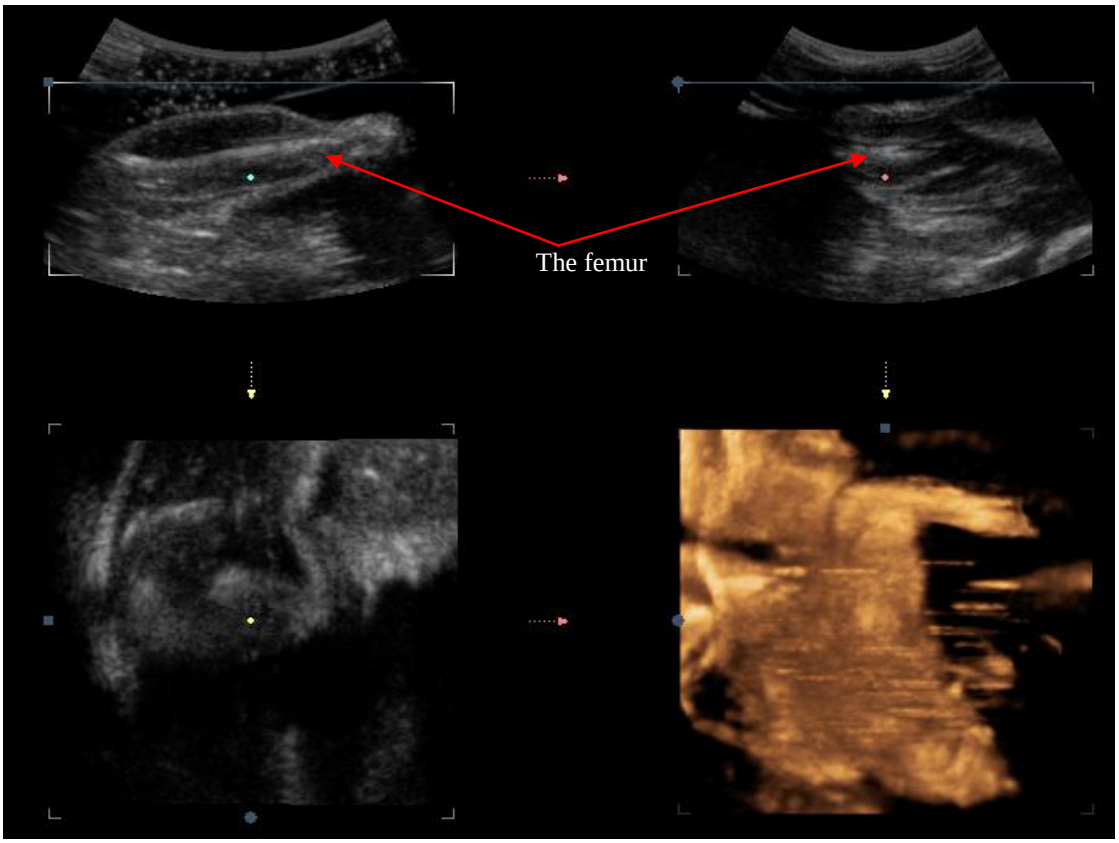


Figure 4.6. Rat leg phantom. Top left: longitudinal plane, top right: transverse plane, bottom left : coronal plane, and bottom right: 3D visualisation.

4.3.1.3 Normal Cases

Scans from 25 healthy pregnancies were used (mean gestational age 24 completed weeks, range 18 – 30 weeks). For each fetus, 6 3D ultrasound scans of the longitudinal view femur were taken from different angles. They were scanned and the effect of fusion on these scans was investigated. The 6 scans of the same femur were taken with a consistent protocol (i.e. a total of 150 ultrasound volumes were used). All 6 scans included the entire femur acquired in the longitudinal view and the reference scan needs to be in a near horizontal position; the central part of the ultrasound beam was over the mid shaft (2 scans), the distal epiphysis (2 scans) or the proximal epiphysis (2 scans), see Figure 4.7. The first scan was the reference scan to which the remaining scans were aligned. Scan dimensions were roughly 120 × 230 × 120 voxels and voxels dimensions were ($0.33 \times 0.33 \times 0.33$ mm^3).

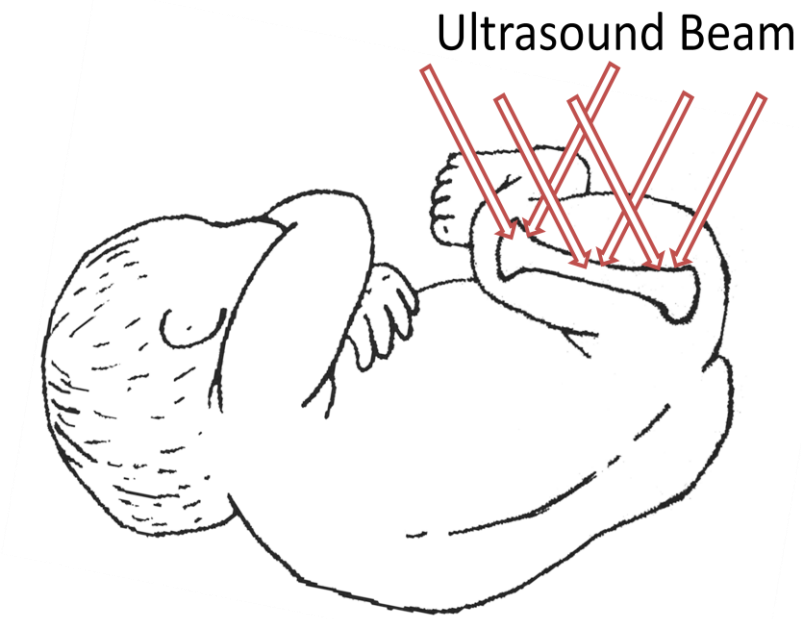


Figure 4.7. Scanning protocol that shows ultrasound beam direction of the fetal femur.

4.3.1.4 Post-mortem Case

Although the normal cases study showed promising results (see Section 4.4), it does not validate the registration accuracy neither prove the technical correctness of the proposed technique. This is because no gold standard that tells us the exact femur volume exists. Therefore, we validated the technique with one post-mortem case. The case was taken from a study where women who are expected to deliver a stillborn fetus within 72 hours are offered the opportunity to take part in this study, following a sensitive discussion, detailed explanation of methods and objectives and an informed consent. Women who agree to take part are offered an ultrasound examination before delivery in order to obtain 3D fetal ultrasound volumes for several fetal structures including the femur. Following delivery, the stillborn baby undergoes a whole body CT and a specific CT for the baby's legs. Ethics approval can be found in Appendix A. Figure 4.8 shows an example of axial (transverse), sagittal (longitudinal) and coronal views for one 3D ultrasound and the CT.

The fetus was 18 weeks +5 days. 6 ultrasound volumes were acquired following the protocol described in the normal cases study. In an initial attempt, the six volumes were registered to each other while varying the reference image. Only one or two volumes were successfully aligned to each reference image. The registration mainly failed because of the poor quality of the ultrasound volumes compared to the normal babies. To handle this, we manually cropped the volumes around the femur and then aligned the cropped volumes. This resulted in ultrasound volumes of approximately 44 × 140 × 44 voxels of ($0.33 \times 0.33 \times 0.33$ mm^3) voxel dimensions. The CT volume was 73 × 512 × 512 voxels with ($0.6 \times 0.292969 \times 0.292969$ mm^3) voxel size.

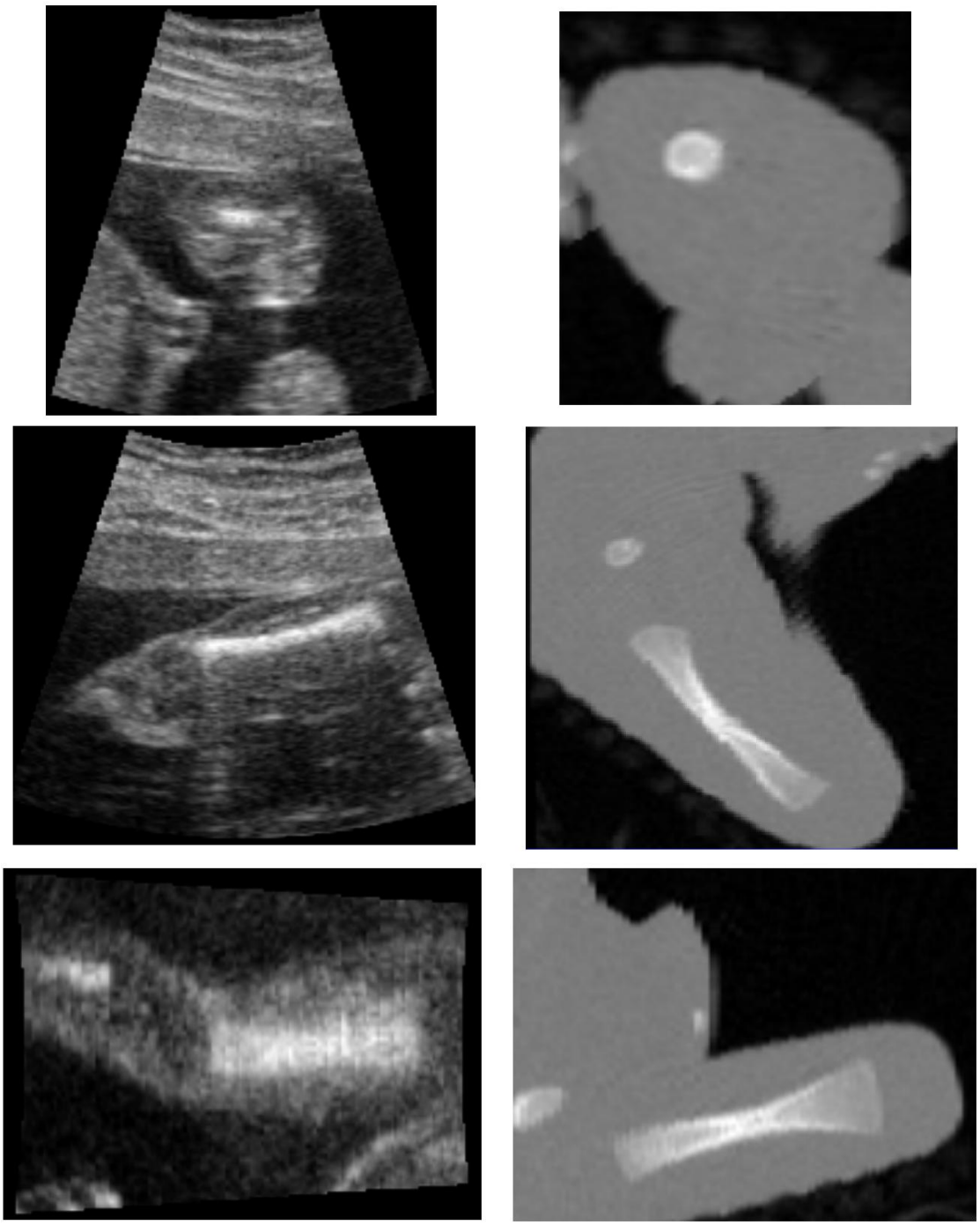


Figure 4.8. Ultrasound vs. CT of the femur. Axial (Top row), longitudinal (middle row), and coronal (bottom row) of a 3D ultrasound scan of the fetal femur (left column) and its CT scan (right column)

4.3.2 Qualitative Analysis

The quality of the best single view, Wavelet-based fused, Max-based fused and Mean-based fused were assessed by an experienced clinician (A. P.). The best single view

among all single views for each patient case was visually chosen. A score from 1 to 10 was given to each scan such that 10 means good femoral definition and 1 means poor femoral definition. The ranking took into account 1) the femoral border especially from the distal and proximal epiphyses, 2) the amount of signal attenuation and image quality just underneath the femur, 3) the overall contrast of the femur and 4) the variation between the femur and the surrounding tissues. Patient information was anonymised and images were randomly presented during the ranking process. Since the fused images can be distinguished by human eye by looking to the border of the field of view, the images were manually cropped around the femur before the ranking to hide this effect. The cropping included as much tissue as possible and at least the whole thigh visible. 24 volumes were analysed with each approach giving a total of (24×4=96) volumes.

4.3.3 Quantitative Analysis

For quantitative assessment, three intensity-based enhancement measures were estimated (Rajpoot, 2009). The measures are percentage change of contrast, percentage change of Contrast to Noise Ratio (CNR) and percentage change of Signal to Noise Ratio (SNR), and they are defined as:

$$\Delta Contrast = \left(\frac{\mu_{fused}^{femur} - \mu_{fused}^{background}}{\frac{1}{M}\sum_{i=1}^{M} \mu_i^{femur} - \mu_i^{background}} - 1 \right) * 100 \tag{4.2}$$

$$\Delta CNR = \left(\frac{CNR_{fused}}{\frac{1}{M}\sum_{i=1}^{M} CNR_i} - 1 \right) * 100 \quad \text{where} \tag{4.3}$$

$$CNR = \frac{\mu^{femur} - \mu^{background}}{\sqrt{(\sigma^{femur})^2 + (\sigma^{background})^2}}$$

$$\Delta SNR = \frac{\Delta SNR^{femur} + \Delta SNR^{background}}{2} \quad \text{where}$$

$$\Delta SNR^{object} = \left(\frac{20 * \log\left(\frac{\mu_{fused}^{object}}{\sigma_{fused}^{femur}} \right)}{\frac{1}{M} \sum_{i=1}^{M} 20 * \log\left(\frac{\mu_i^{object}}{\sigma_i^{object}} \right)} - 1 \right) * 100 \tag{4.4}$$

where μ and σ are the mean and the standard deviation within a region R, respectively; M is the number of the single view images used in the fusion process. R is a 3D representative region from the object of interest (the femur and background) of size $(7)^3$ voxels. The background region was selected from the thigh tissue directly above the femur. Since R is a small 3D region and cannot capture the whole 3D structure, 10 different regions, 5 for the background and 5 for the femur were used from every image. ΔContrast, ΔCNR and ΔSNR were calculated for these regions and the median value was reported. To compare wavelet-based fusion with other techniques, we compared percentages for wavelet-based, Mean-based and Max-based fused images.

In order to assess intra and inter-observer reproducibility, manual volume measurements were performed by two experts (M. Y. and C. I.) on the fused and the best single view (visually chosen) for each case. The first expert (M.Y.) is an engineer with an experience in ultrasound image analysis. The second (C.I.) is an obstetrician with an experience in ultrasound scanning and image analysis and he acquired all scans. Both experts followed a consistent protocol during manual segmentation. A standard multiple plane segmentation technique was used as previously described for the femur by Chang et al. (2007) with slight modification. Briefly, the cross-sectional imaging plane was selected and the outline of the femur was traced manually on consecutive slices. Slices that are close to the distal and proximal epiphyses were segmented

differently than the ones in the middle. Slices close to the mid shaft were segmented based on what is visible assuming the entire femur is visible. Slices on both epiphysis ends are treated in slightly different ways since we assume that the lower part of the femur is partially invisible. In these regions, the upper part of the femur was segmented since it is visible then a straight line was made to connect the lower border of the femur. Segmenting such a plane is described in Figure 4.1 (d). The femoral volume was then calculated automatically by integration of the constituent slices. The software used to do the manual tracing was the Medical Imaging Interaction Toolkit MITK, version 0.12.2, (c) 2003-2008 German Cancer Research Center, Division of Medical and Biological Informatics; The first expert segmented the single and fused volumes once, while the second expert segmented them twice. The first expert also segmented three more single views (total of 4 single views) in five cases to compare the segmented fused view to all segmented single views. In total, the first expert segmented 63 volumes while the second expert segmented 96 volumes.

The segmented ultrasound volumes were used in order to 1) assess the intra and inter-observer reproducibility; 2) investigate the effect of fusion on the measured femur volume; and 3) compare the fused image with the union and the intersection of the all single views. For reproducibility analysis Bland Altman plots of measurement differences were used; also the Coefficients of Variation (CV) and the Intraclass Correlation Coefficients (ICC) were calculated.

In the post-mortem case, the registration and fusion stages were applied to the ultrasound volumes to create a fused volume. To investigate the volume increase in the fused volume, we manually segmented the femur in the single ultrasound views, the fused view and the CT volume. The hypothesis is that the measured femoral volume

from the fused view is larger than the femoral volumes measured from the single views individually and it is also closer to the measured femoral volume from the CT.

4.4 Results

4.4.1 Datasets

In the in vitro case, none of the animal phantoms were used in the fusion process because of the previously mentioned problems. On the other hand, the fetal phantom was scanned from 10 different angles and 9 out of 10 views were successfully aligned.

In the in vivo case, volume fusion was successful in 24 out of 25 cases with a median of 4 merged views per fusion (range 2-6). One out of the twenty-five cases was rejected because of the low overlap between images which led to the misalignment of all six scans. In the post-mortem case, 5 out of 6 single view volumes were correctly aligned. One volume was rejected because it had significantly poor image quality.

4.4.2 Qualitative Results

The fused view of the fetal femur phantom, Figure 4.9 (b), has a well defined region, clear boundary and high SNR compared to a single view image of the phantom, Figure 4.9 (a).

Table 4.1 shows the mean ± standard deviation, highest and lowest scores assigned by the clinician for best single view, wavelet-based fused, max-based fused and Mean-based fused. The mean score for the wavelet-based fused view is higher than other scores. Visual comparisons between our method and the Max-based and Mean-based intensity fusion show that wavelet-based and Max-based fusion techniques preserve the contour of the femur compared to the Mean-based fusion technique. On the other hand,

Max-based fused images are more noisy compared to the other two techniques since this approach keeps the maximum voxel intensity among all single views. See Figure 4.10 for a visual comparison between the three fusion techniques. The wavelet-based method preserves more meaningful information than the other two and suppresses noise.

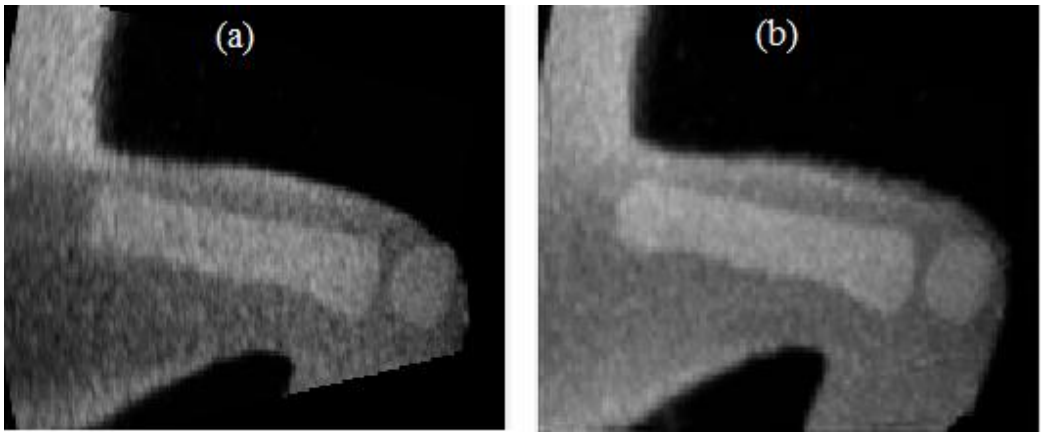


Figure 4.9. The femur of a fetal phantom. (a) 2D slice of one single view with (b) the corresponding 2D wavelet-based fused slice.

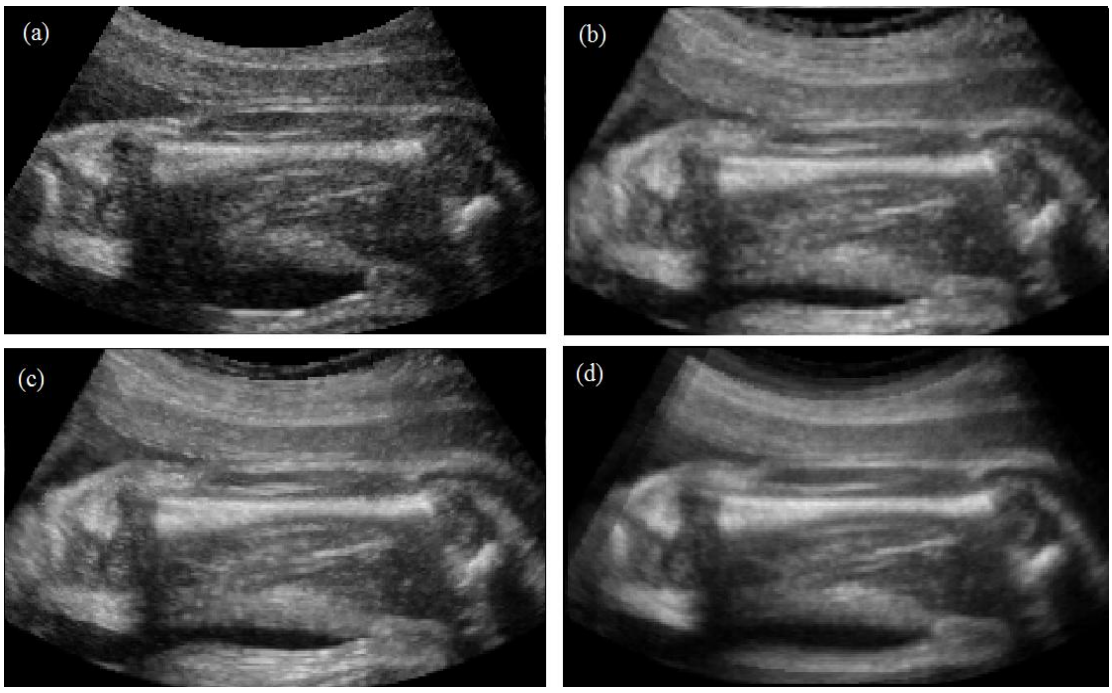


Figure 4.10. Visual comparisons between different fusion methods. Single 2D slice out of a 3D volume is shown for convenience. (a) Original 2D slice, (b) Wavelet-based fusion, (c) Max intensity fusion, and (d) Mean intensity fusion.

Table 4.1.The mean ± standard deviation, highest and lowest scores given to the 24 volumes. The scores are for best single view, wavelet-based fused, max-intensity fused and mean-intensity fused

	Best single	Wavelet-based fused	Max-intensity fused	Mean-intensity fused
μ±σ	6.29 ± 1.04	6.79 ± 1.14	6.67 ± 1.34	6.50 ± 1.32
Highest	8	9	9	9
Lowest	4	5	3	4

To judge the accuracy of a registration algorithm on the 24 normal patient cases, we visually validated the registration results to show that the femur is correctly aligned and also other structures, e.g., thigh skin, knee tissues, etc, are correctly aligned. For instance, see green rectangles and the skin tissue in Figure 4.15 (a-c). In the post-mortem case, we show in the next section that volume increase in the fused volume is genuine. This implies that the registration between single view volumes is accurate.

4.4.3 Quantitative Results

The percentage change of contrast, CNR and SNR are shown in Figure 4.11 for the 24 cases and the fetal femur phantom for the different methods (Wavelet, Mean and Max). Notice that ΔContrast, ΔCNR and ΔSNR are always positive except for one case (patient 9) where only ΔContrast is negative. This implies that the contrast, CNR and SNR of the fused view is higher than the ones of all single views for all cases and the phantom. In addition, Figure 4.11 shows ΔContrast, ΔCNR and ΔSNR for the three fusion techniques (Wavelet-based, Mean and Max). Notice that the wavelet-based fusion provides increased contrast and SNR images relative to the other two fusion techniques while having comparable CNR. The box plot in Figure 4.11 (d) also shows how wavelet-based fusion outperforms the other two techniques. Moreover, Table 4.2 shows the mean and standard deviation of ΔContrast, ΔCNR and ΔSNR on all cases for

the three fusion techniques. It demonstrates how the proposed technique provides better image quality than the other two while maintaining good SNR.

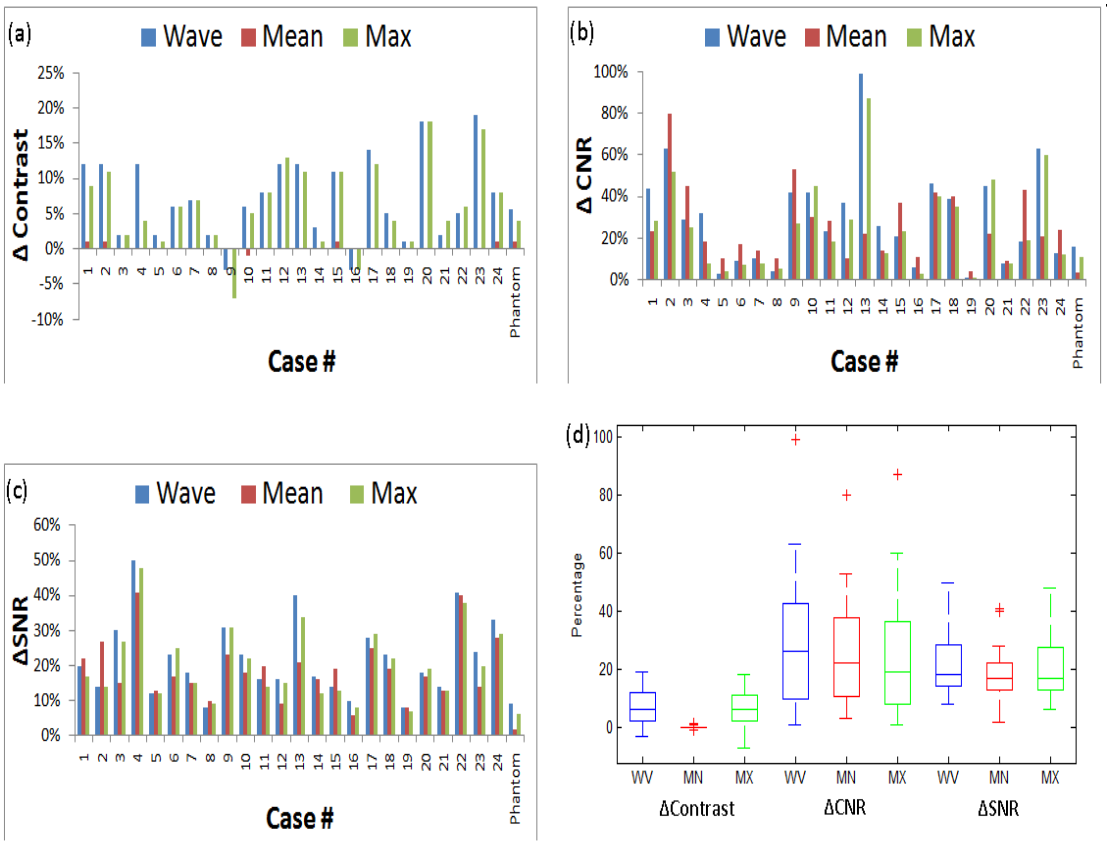


Figure 4.11. (a-c) respectively show the percentage change of contrast, percentage change of CNR and percentage change of SNR. Each figure shows the wavelet-based (blue), Mean (red) and Max (green) fusion techniques where x-axis represents the case ID for the 24 cases and the fetal phantom. (d) represents a Box plot which shows the percentage change of contrast (the first three columns on x-axis), percentage change of CNR (forth, fifth and sixth columns on x-axis) and percentage change of SNR (seventh, eighth and ninth columns on x-axis) for Wavelet-based (WV, *blue*), Mean (MN, *red*) and Max (MX, *green*) fusion techniques.

Table 4.2. The mean ± standard deviation of the ΔContrast, ΔCNR and ΔSNR on the three fusion techniques.

	Wavelet	Mean	Max
ΔContrast (μ±σ)	7 ± 5.8%	0 ± 0.5%	6 ± 5.8%
ΔCNR (μ±σ)	30 ± 23.2%	25 ± 17.9%	25 ± 21.2%
ΔSNR (μ±σ)	22 ± 11.0%	18 ± 9.2%	20 ± 10.5%

Reproducibility of femur volume measurement was improved on the fused compared with single datasets, both within the same observer (intra-observer) and between observers (inter-observer). Intra-observer CV decreased from 5.8% in the single view to 4.6% for the fused view. Inter-observer CV decreased from 10.0% in the single view to 4.7% for the fused view. Intra-observer ICC increased from 0.977 (95% confidence intervals 0.948 – 0.990) for the single view to 0.987 (0.970 – 0.994). Similarly, inter-observer ICC increased from 0.931 (0.849 - 0.970) to 0.985 (0.964 – 0.994). Figure 4.12 shows the Bland Altman plots of intra-observer differences for the single and fused view. The equivalent inter-observer differences are demonstrated on Figure 4.13.

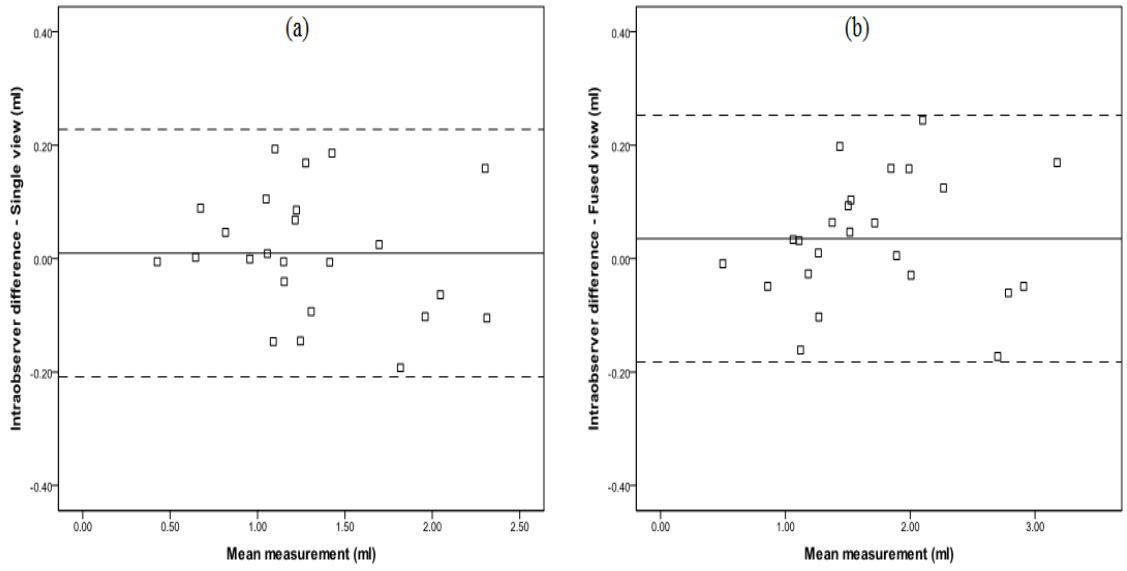


Figure 4.12. Bland-Altman plots for (a) intra-observer measurement differences in the single view (mean difference 0.01 ml; 95% limits of agreement 0.23 to -0.21 ml) and (b) intra-observer measurement differences in the fused view (mean difference 0.04 ml; 95% limits of agreement 0.25 to -0.18 ml).

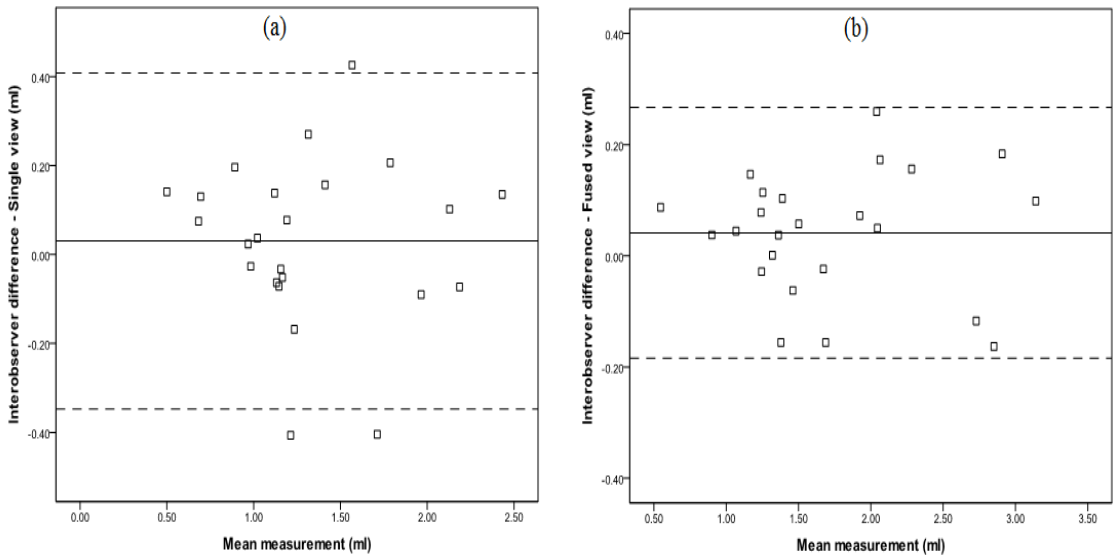


Figure 4.13. Bland-Altman plots for (a) inter-observer measurement differences in the single view (mean difference 0.03 ml; 95% limits of agreement 0.41 to -0.35 ml) and (b) inter-observer measurement differences in the fused view (mean difference 0.04 ml; 95% limits of agreement 0.27 to -0.18 ml).

Single views from 5 out of the 24 femoral cases were manually segmented. Each one has four single views. The percentage intersection of the segmented single views with the segmented fused view was low at 47.7±7%. Although the four single views were aligned, each one highlights a different part of the femur, thus demonstrating the potential benefits of a fused image. For the five patient cases, the fused view had approximately a 16±4% larger volume than the mean volume of the 4 single views. On the other hand, the volume of union between the four single views was larger than the volume of the fused view for both datasets. The union volume was 22±6% larger than the fused volume. This is mainly because of the unclear boundaries of the distal and proximal epiphysis which in turn led to an inaccurate manual segmentation. Figure 4.14 shows volume comparisons for the five cases while Figure 4.15 the result of fusion on an ultrasound case. It also shows manual segmentation results.

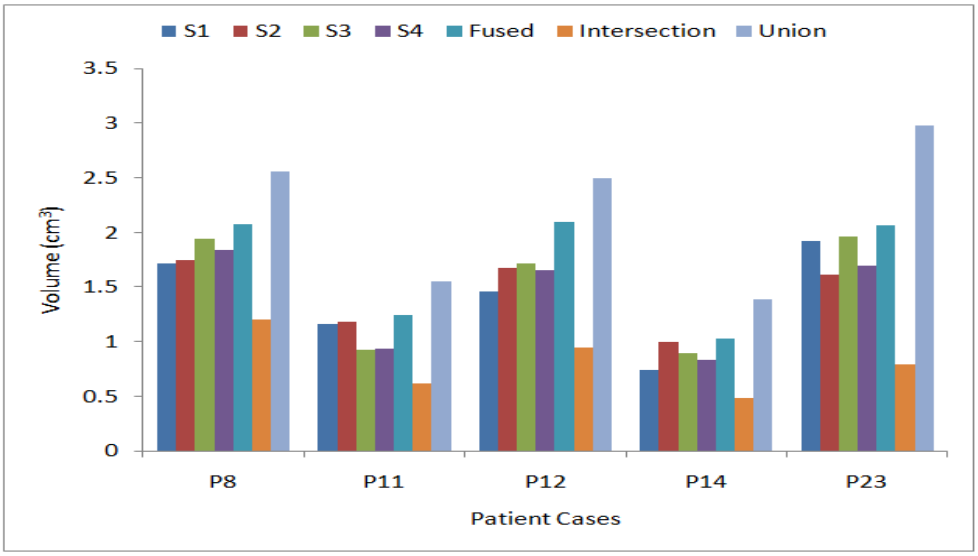


Figure 4.14. Comparing the femur volumes on four single views, fused view, union and intersection of the single views. Si represents the i^{th} segmentation while Pj represents the j^{th} patient.

In the post-mortem study, femoral volumes were calculated from the manually segmented single views, fused view and the CT. Figure 4.16 shows an example on one 2D slice with its manual segmentation in the five single views and the fused one. Figure 4.17 shows one 2D slice from the fused image as well as the manual segmentation from the fused and single views. Notice in this figure how the manual segmentation from the fused image provides extra information that is captured from all single views. Finally, results from the measured femoral volumes on all single views, fused view and CT are reported in Figure 4.18 and Table 4.3. Figure 4.18 depicts the absolute volume difference of the measured femoral volumes of all ultrasound images with respect to the CT while Table 4.3 shows the original measured volumes on all images. Mean volume increase in the fused view and the CT are 13.2% and 12.8% respectively. The femur volume from both the fused view and the CT are very close (difference is $2mm^3$). These volumes support our hypothesis that a fused view provides more realistic volumetric measurements compared to a single view.

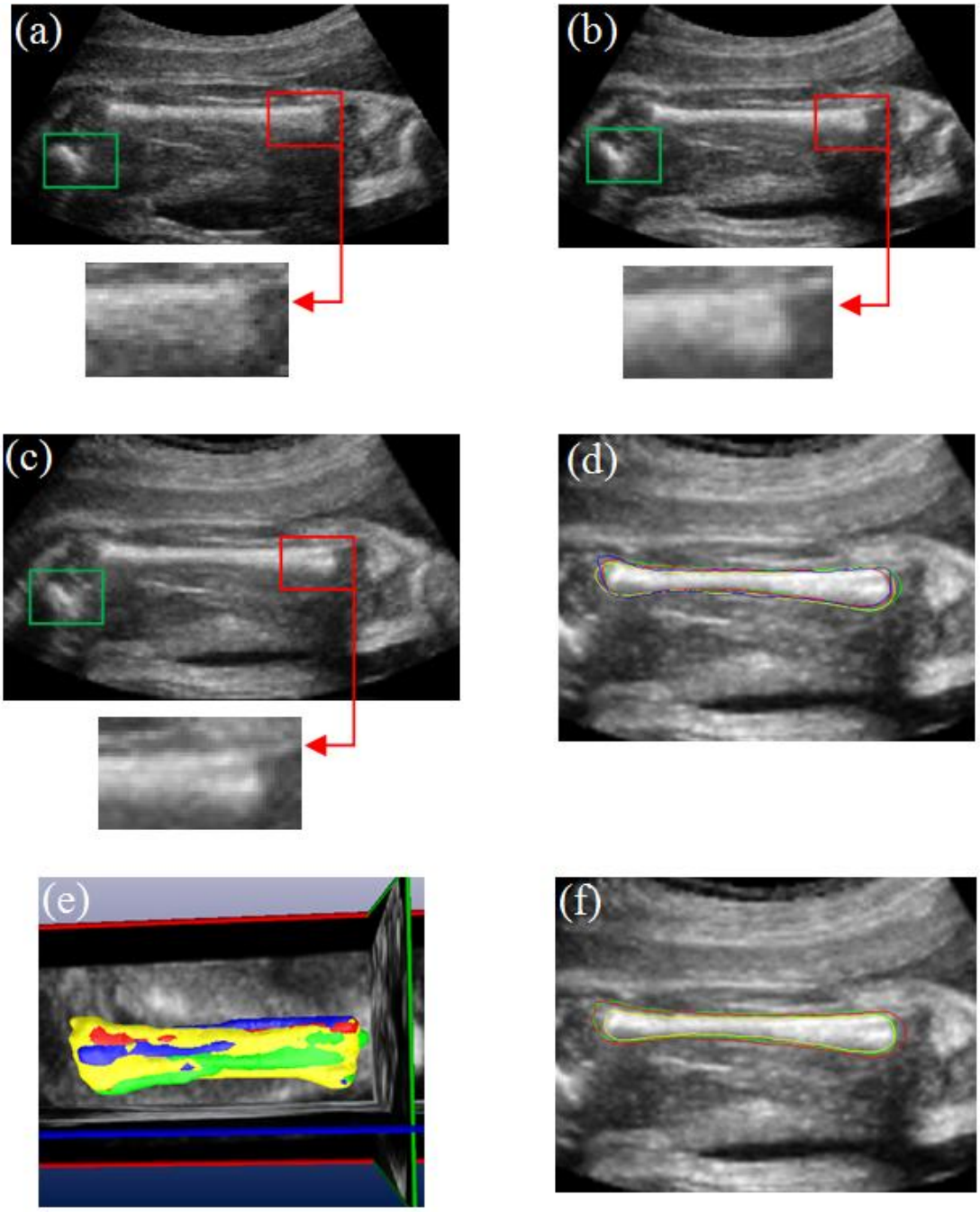


Figure 4.15. Detailed comparisons using the manual segmentation on aligned and fused view. 2D slices out of a 3D volume are shown for convenience. (a) and (b) show two Aligned single views while (c) shows the fused view. The green rectangle shows the correct alignment of another structure (partial tibia bone). (d) shows the segmentation from four single views over the fused image while (e) shows the 3D surface of the segmentation of the 4 single views over the fused image. (f) shows the segmentation on the fused image where fused (green), union (red) and intersection (yellow).

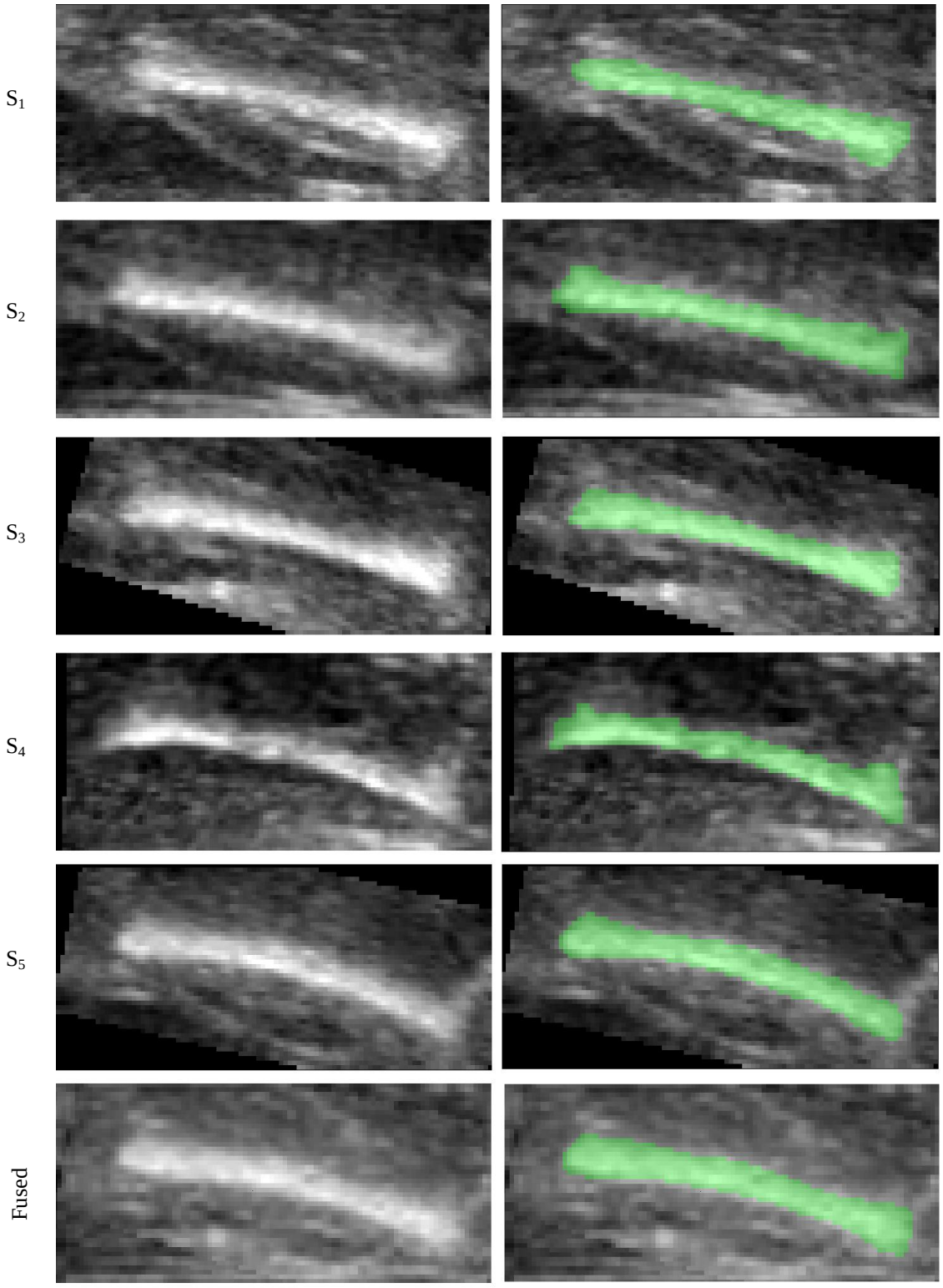


Figure 4.16. 2D slices and their manual segmentation for the five single views (S_1 to S_5) and the fused view (Fused).

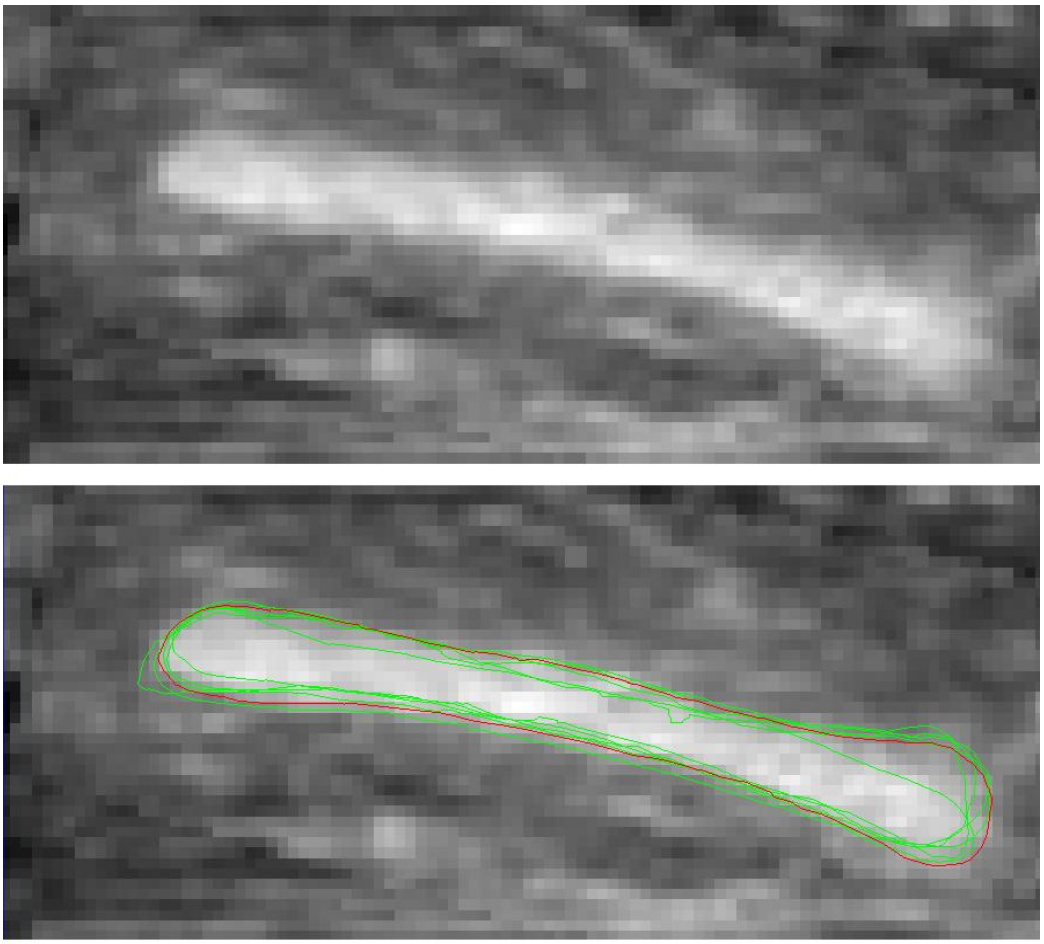


Figure 4.17. Top: original 2D slice from the fused view. Bottom: manual segmentations from the fused (red) with segmentation from all single views (green).

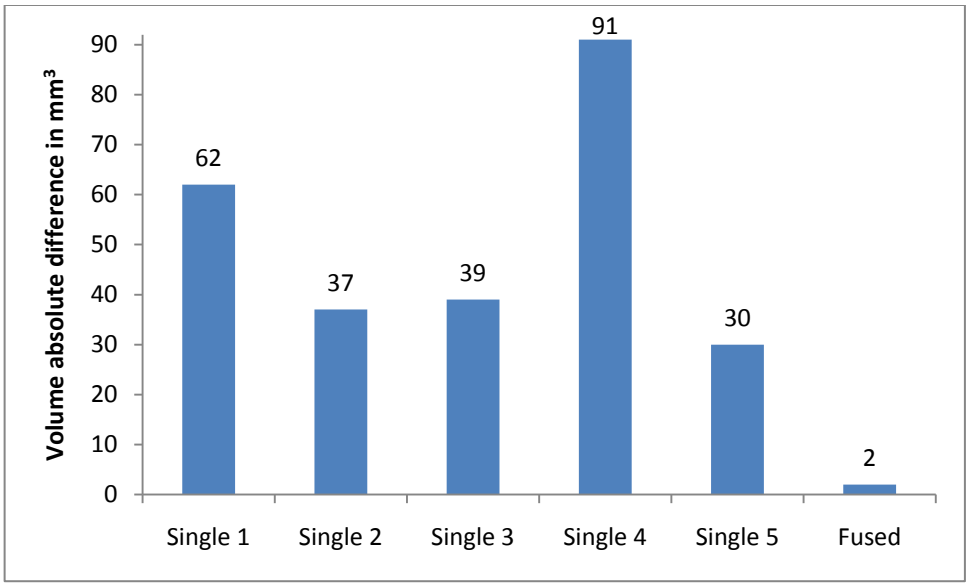


Figure 4.18. Absolute difference between CT femur volume and all single views and the fused one. Volumes in mm^3.

Table 4.3. The manually measured volume of the femur of a stillborn baby on all single views, fused view and the CT. Volumes are in mm^3.

Single 1	Single 2	Single 3	Single 4	Single 5	Fused	CT
343	368	366	314	375	407	405

4.5 Discussion

We have hypothesised that image fusion enhances the quality of ultrasound images of bony structures which leads to a more accurate and reproducible assessment and measurement. We presented a novel processing in the wavelet domain to improve femoral boundary definition.

Qualitative and quantitative validation on in vitro and in vivo 3D ultrasound supported our hypothesis. We showed how wavelet-based fusion outperforms other standard techniques of the fetal and phantom femur. Since we average the first high frequency component, speckle will be reduced which may be important to preserve in some applications. Interestingly, we showed that the intersected femur volume between four segmented aligned single view 3D ultrasound images is roughly 48% of the fused femur volume. In addition, the fused femur volume is approximately 16% more than the mean of 4 single views. Moreover, a CT scan from a stillborn baby helped us validate that the volume increase in fused view is genuine. This implies that 3D quantification from single views may be inaccurate. This also proves that missing information can be compensated if multiple ultrasound volumes for a fetal structure are aligned and fused. Although we applied the technique to the fetal femur, it is applicable to other fetal structures with minimal tuning. We believe that this approach will make it easier and more accurate to measure volumetric fetal structures especially bony ones which will hopefully lead to a better diagnosis.

The work has shown promising applicability. However, extra work is required from the sonographer to image multiple volumes and then apply the method to generate a fused view. We believe that fusion is important and ultrasound manufacturers should consider embedding a fusion capability into the machine.

Finally, notice that the fusion techniques employed in this thesis rely on image registration by geometric transformations. On the other hand, other methods have been developed in the literature that use generative models to do the registration. Since the main target in the medical imaging domain is to derive a more complete view of a 3D object, image registration using Super-resolution techniques (Capel and Zisserman, 2003, Pickup et al., 2009) might also improve image quality. Such techniques fuse information from several low-resolution images into a single clearer high-resolution image. Such techniques have shown promising results in other domains and might be interesting to investigate in the future.

Chapter 5

Maternal Vitamin D Concentration during Pregnancy: A Population-based Cohort Study

This chapter mainly covers the statistical comparisons between the automatic measurements from the proposed segmentation techniques and fetal gestational age and maternal vitamin D. We first introduce fetal and maternal characteristics that were used in a previous study (Mahon, 2007). We then describe automatic measurements that can be derived from the different variations of the segmentation technique. In Section 5.2, we describe data cleaning and normalisation steps. We use linear regression analysis to compare the relationship between the automatic measurements and fetal and maternal characteristics. Finally, we discuss our results and present conclusions in Section 5.3.

5.1 Characteristics and Measurements

5.1.1 Maternal and Fetal Characteristics

In Section 2.3.1, we described the Southampton Women's Survey study. In that study, several manual measurements from the ultrasound volumes as well as several maternal and fetal characteristics were recorded (Mahon, 2007). We are only interested in a few of these for comparisons. The two most important characteristics are fetal gestational age and maternal vitamin D level. Fetal gestational age varies from 18 to 21 weeks. The main measurements we calculate and compare with the SWS measurements are the femur volume, femur length, cross sectional area, splaying indices and splaying angles. Two splaying indices and two splaying angles were calculated.

5.1.2 Automatic Measurements

Several measurements were automatically calculated from the 450 segmented femur volumes. Geometric demonstration of some of these measurements are given in Section 3.3.4 on page 57. Table 5.1 summarises the seventeen measurements used in our comparisons. Notice that CSA, PSI, MSI, Sp1 and Sp2 have three measurements (average, maximum and minimum). This is because there are two epiphyses ends (distal and proximal). Therefore, Avg represents the mean of the two measurements from both ends, Max represents the larger of the two and Min represents the smallest of the two. The main difference between Sp1 and Sp2 are the way that these angles are automatically calculated from the segmented femur volume. Measurements were found five times using five variations of the segmentation technique (namely, axial (Ax), longitudinal (Long), classic RF (RF), Smart Fast RF (SFRF) and Smart Fast-Weighted RF (SFWRF)). Notice that we did not include all variations of the segmentation technique proposed in Chapter 3 but we included the ones that cover all proposed ideas

(1- the use of the Random Forests technique, 2- the Smart selection of features and 3- the weighted voting) The axial and longitudinal methods represent the segmentation using the classic Random Forests on 2D slices separately. Training and testing in the axial method were performed on 2D images from the axial view while 2D longitudinal images were used in the longitudinal method. Description of the 3D Random Forests methods (the classic RF, SFRF and SFWRF) were given in Section 2.5 (page 20) and Chapter 3. This means that 17×5=60 measurement variables were automatically found.

Table 5.1. List of all automatic measurements and their abbreviations.

Measure	Description
FV	The Femur Volume in cm^3
FL	The Femur Length in cm
Avg, Max & Min CSA	Cross Sectional Area in cm^2
Avg, Max & Min PSI	Splaying index = CSA/FL as in equation (2.1)
Avg, Max & Min MSI	Splaying index = CSA/FL^2
Avg, Max & Min Sp 1	Splaying angle 1
Avg, Max & Min Sp 2	Splaying angle 2

5.2 Experimental Results

5.2.1 Date Checking and Cleaning

Data checking and cleaning were performed on all automatic measurements to check and remove outliers. If an outlier was found, the calculated variable was checked to try

to fix any probable problem. For example, we noticed a few outliers in the femur volume using the classic Random Forests technique. Therefore, we checked the technique and found that it usually over-segments several cases. As a result, slightly more strict thresholds were used to restrict the technique to provide less false negatives. The mean and standard deviation of all variables are reported in Table 5.2. Notice that we report comparisons not only for all cases but also for girls and boys to check if there is any significant difference with respect to gender. Kernel density estimates and box plots of the femur volume and femur length are used to check the measurement distribution of raw measurements. Figure 5.1 and Figure 5.2 show these plots for the femur volume and femur length respectively.

Table 5.2. Mean ± standard deviation of all automatic measurements.

Both	Axial	Long	RF	SFRF	SFWRF
FV	0.664±0.237	0.68±0.235	0.743±0.245	0.583±0.169	0.569±0.126
FL	3.223±0.264	3.399±0.299	3.482±0.356	3.227±0.27	3.26±0.241
Max CSA	0.015±0.004	0.015±0.004	0.017±0.006	0.014±0.004	0.014±0.003
Avg CSA	0.014±0.004	0.014±0.004	0.016±0.004	0.012±0.003	0.012±0.002
Min CSA	0.012±0.004	0.012±0.004	0.014±0.004	0.01±0.003	0.01±0.002
Max PSI	0.467±0.129	0.442±0.11	0.493±0.142	0.426±0.115	0.42±0.101
Avg PSI	0.419±0.112	0.398±0.095	0.463±0.119	0.375±0.087	0.37±0.074
Min PSI	0.37±0.114	0.354±0.093	0.394±0.11	0.324±0.078	0.32±0.069
Max MSI	0.145±0.043	0.13±0.031	0.142±0.041	0.132±0.035	0.13±0.033
Avg MSI	0.13±0.034	0.117±0.026	0.134±0.036	0.117±0.027	0.114±0.026
Min MSI	0.114±0.034	0.104±0.026	0.114±0.033	0.101±0.025	0.099±0.024
Max Sp1	9.54±5.056	9.783±4.95	9.397±4.692	9.397±4.692	9.326±4.712
Avg Sp1	6.614±3.588	6.847±3.384	6.685±3.615	6.685±3.615	6.609±3.34
Min Sp1	3.687±3.312	3.911±3.306	3.972±3.574	3.972±3.574	3.893±3.314
Max Sp2	9.504±5.215	9.884±4.844	10.593±5.66	9.391±4.825	9.093±4.633
Avg Sp2	6.702±3.545	7.145±3.486	7.694±4.132	6.87±3.609	6.581±3.291
Min Sp2	3.9±3.152	4.406±3.322	4.795±3.802	4.349±3.563	4.069±3.117
Girls	**Axial**	**Long**	**RF**	**SFRF**	**SFWRF**
FV	0.674±0.241	0.695±0.244	0.746±0.241	0.592±0.176	0.575±0.129
FL	3.223±0.263	3.416±0.305	3.496±0.339	3.237±0.275	3.269±0.233
Max CSA	0.015±0.004	0.015±0.004	0.017±0.005	0.014±0.004	0.014±0.003
Avg CSA	0.014±0.004	0.014±0.004	0.016±0.004	0.012±0.003	0.012±0.002
Min CSA	0.012±0.004	0.012±0.004	0.014±0.004	0.011±0.003	0.01±0.002
Max PSI	0.474±0.127	0.448±0.11	0.493±0.138	0.43±0.112	0.424±0.096
Avg PSI	0.425±0.114	0.404±0.095	0.461±0.112	0.379±0.085	0.373±0.071
Min PSI	0.376±0.118	0.36±0.093	0.397±0.103	0.327±0.077	0.322±0.066
Max MSI	0.148±0.041	0.131±0.03	0.141±0.038	0.133±0.034	0.13±0.031
Avg MSI	0.132±0.034	0.118±0.026	0.132±0.033	0.117±0.026	0.115±0.024
Min MSI	0.116±0.035	0.105±0.025	0.114±0.031	0.102±0.025	0.099±0.023
Max Sp1	9.526±5	9.732±4.855	9.474±4.564	9.474±4.564	9.392±4.618
Avg Sp1	6.578±3.59	6.702±3.295	6.677±3.496	6.677±3.496	6.694±3.339
Min Sp1	3.63±3.355	3.671±3.283	3.88±3.552	3.88±3.552	3.995±3.356
Max Sp2	9.398±5.067	9.769±4.699	10.631±5.388	9.376±4.789	9.142±4.673
Avg Sp2	6.652±3.435	7.038±3.328	7.777±4.003	6.811±3.539	6.702±3.39
Min Sp2	3.906±3.166	4.307±3.297	4.924±3.871	4.247±3.497	4.263±3.329
Boys	**Axial**	**Long**	**RF**	**SFRF**	**SFWRF**
FV	0.653±0.232	0.663±0.223	0.74±0.25	0.574±0.161	0.563±0.123
FL	3.224±0.265	3.381±0.292	3.467±0.375	3.216±0.265	3.251±0.25
Max CSA	0.015±0.004	0.015±0.004	0.017±0.006	0.014±0.004	0.013±0.003
Avg CSA	0.013±0.004	0.013±0.004	0.016±0.005	0.012±0.003	0.012±0.002
Min CSA	0.012±0.004	0.012±0.004	0.014±0.004	0.01±0.003	0.01±0.002
Max PSI	0.459±0.13	0.435±0.109	0.493±0.147	0.421±0.118	0.416±0.106
Avg PSI	0.412±0.11	0.391±0.095	0.464±0.126	0.37±0.088	0.366±0.078
Min PSI	0.364±0.11	0.347±0.094	0.39±0.116	0.32±0.079	0.318±0.073
Max MSI	0.143±0.044	0.129±0.032	0.143±0.043	0.131±0.037	0.129±0.036
Avg MSI	0.128±0.034	0.116±0.027	0.135±0.039	0.116±0.028	0.114±0.027
Min MSI	0.113±0.032	0.103±0.026	0.113±0.035	0.1±0.026	0.099±0.025
Max Sp1	9.556±5.127	9.838±5.061	9.314±4.834	9.314±4.834	9.254±4.82
Avg Sp1	6.652±3.593	7.004±3.478	6.693±3.747	6.693±3.747	6.519±3.346
Min Sp1	3.748±3.27	4.169±3.318	4.071±3.603	4.071±3.603	3.783±3.273
Max Sp2	9.617±5.379	10.007±5.003	10.553±5.951	9.408±4.875	9.042±4.599
Avg Sp2	6.755±3.667	7.26±3.652	7.604±4.273	6.933±3.691	6.451±3.184
Min Sp2	3.893±3.143	4.512±3.354	4.656±3.731	4.458±3.638	3.861±2.865

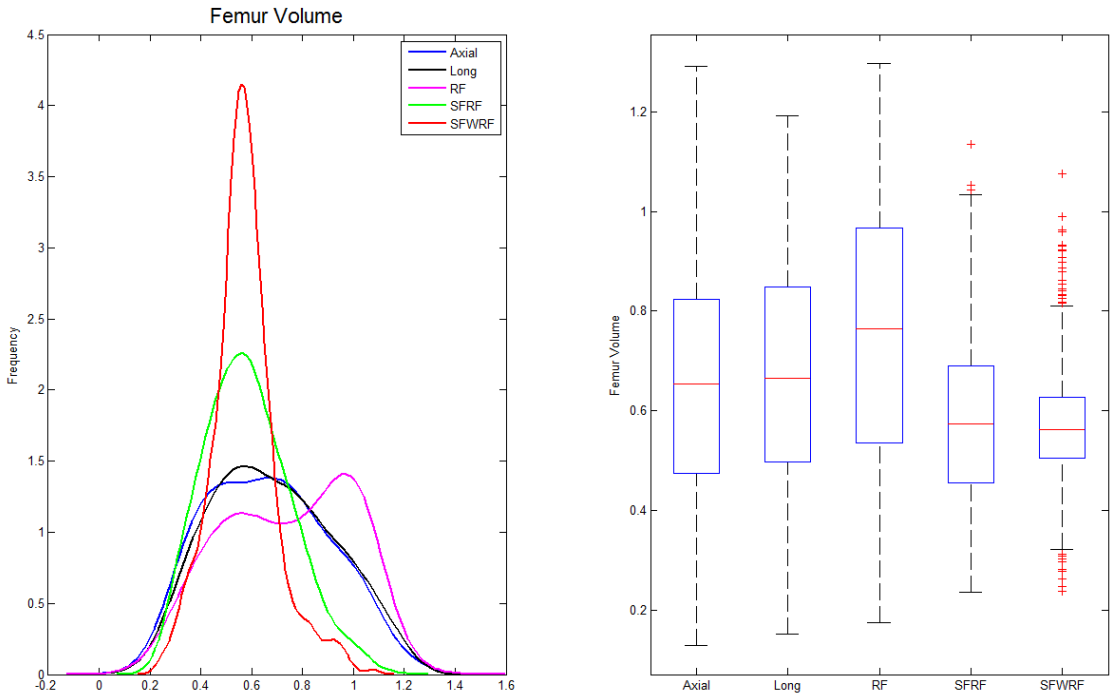


Figure 5.1. Kernel density and box plot for the automatically measured femur volume on 450 fetal femur ultrasound volumes using five variations of the proposed segmentation technique.

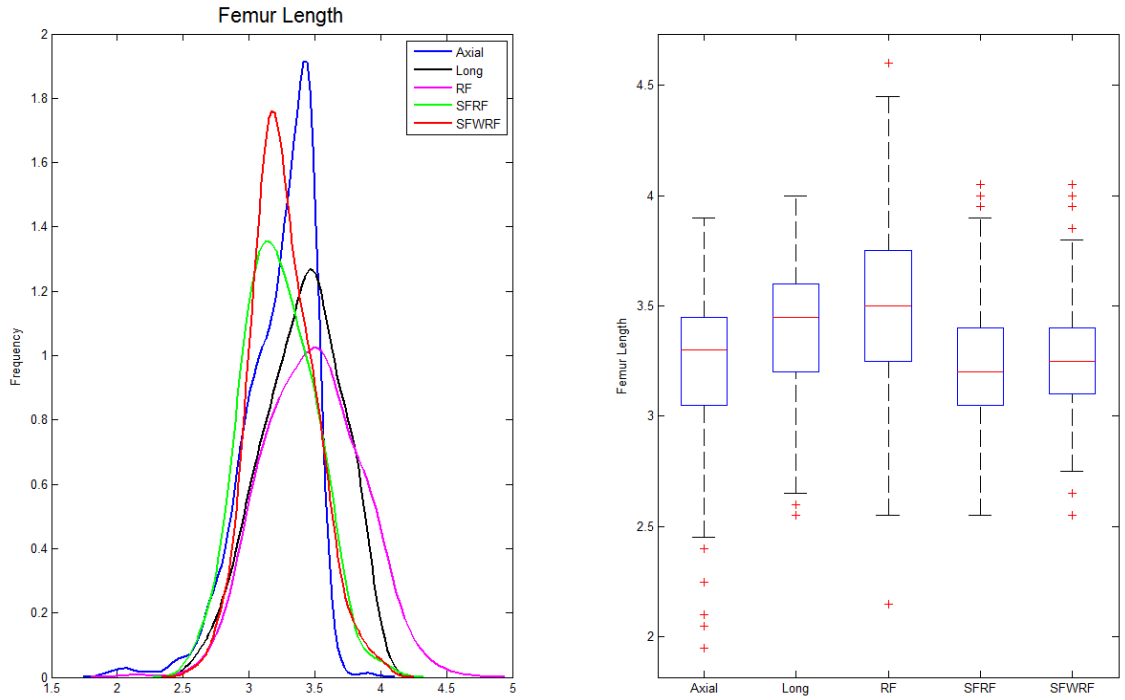


Figure 5.2. Kernel density and box plot for the automatically measured femur length on 450 fetal femur ultrasound volumes using five variations of the proposed segmentation technique.

5.2.2 Data Normalisation

This step is important when comparing several outcomes (automatic measurements, see Table 5.1) which have different units. Therefore, we normalised these measurements to make statistical comparisons like linear regression meaningful. In linear regression, the slope of the regression line (β) is related to the unit of the two variables in regression. Therefore, normalisation allows data on different scales to be compared by bringing all scales to a common one. We used the standard score method to normalise all variables. The standard score is calculated as

$$z = \frac{x - \mu}{\sigma}, \tag{5.1}$$

here x is a vector of values for a variable, μ and σ are scalars that represent the mean and the standard deviation of x, and z is a vector of normalised values of the variable x.

5.2.3 Linear Regression Analysis

Two statistical experiments were performed where linear regression analysis was used to find correlations between automatic and manual measures. These experiments studied 1) the relationship between the normalised automatic measurements and fetal gestational age and 2) the relationship between normalised the automatic measurements and maternal vitamin D.

5.2.3.1 Fetal Gestation Age verses Automatic Measurements

The gestational age range is 18-21 weeks. Figure 5.3 to Figure 5.4 show scatter charts and linear regression analysis between the normalised femur volume (SD) and femur length (SD) from 438 volumes and gestational age of fetuses. Note that the gestational age of 12 fetuses were not recorded so they were dropped in the analysis. More specific linear regression analysis where gender is taken into consideration and for all automatic

measurements is shown in Table 5.3. Note that Table 5.3 spreads on two pages because it is large. From Table 5.3, note how the fetal gestation and the FV, FL and to some extent CSA is positively correlated. In other words, an increase in fetal gestational age implies an increase in these three measures which sounds biologically correct. On the other hand, negative significant correlation between MSI and fetal gestation age was found. This indicates that elder fetuses tend to develop smaller ratio between CSA and FL which could mean that the cross-sectional growth of the femur is relatively smaller than the longitudinal growth. Finally, both boys and girls develop similar correlation with gestational age.

Here is an example that helps understand the interpretation numbers in Table 5.3 and the two successive tables. In first row we can see that FV in both genders has the following statistical correlation 0.399***(0.18, 0.62) with fetal gestation age. This means the slope of the regression line is 0.399. The slope is positive which implies a positive linear relationship. The larger the slope the stronger the correlation. But remember that the slope could be negative in some situations (e.g., MSI and fetal gestation age in Table 5.3) which implies negative correlation and hence the smaller the slope the stronger negative correlation. The three stars mean that the p-value in this correlation is less than 0.001. The confidence interval in this correlation is between 0.18 and 0.62.

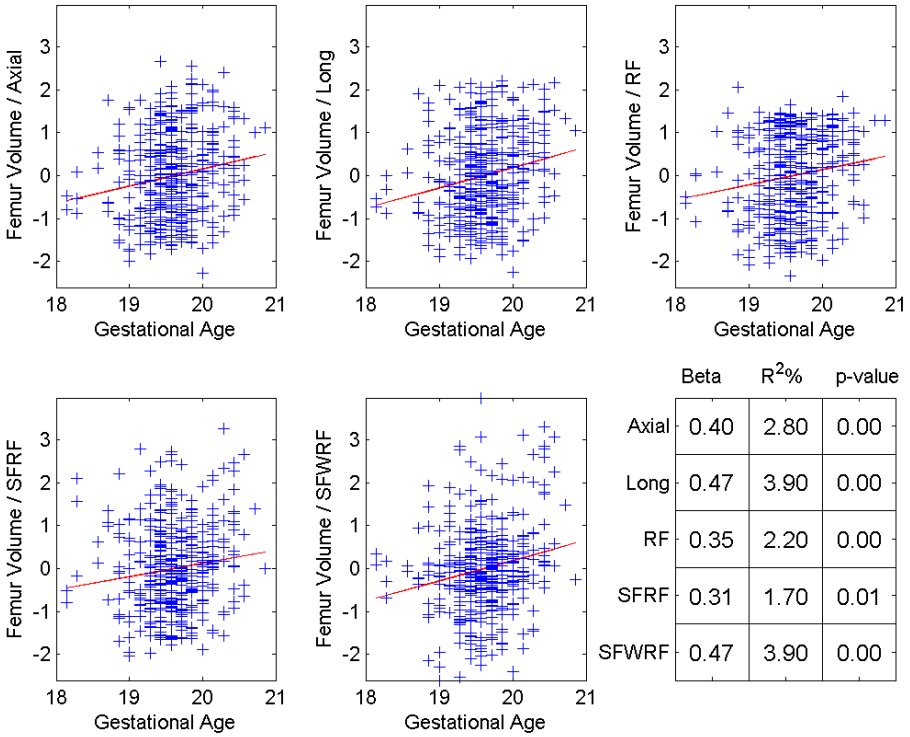


Figure 5.3. Linear regression analysis between normalised femur volume and gestational age of fetuses. Each sub-figure varies depending on the segmentation technique (axial, longitudinal, classic RF, smart fast RF and smart fast-weighted RF).

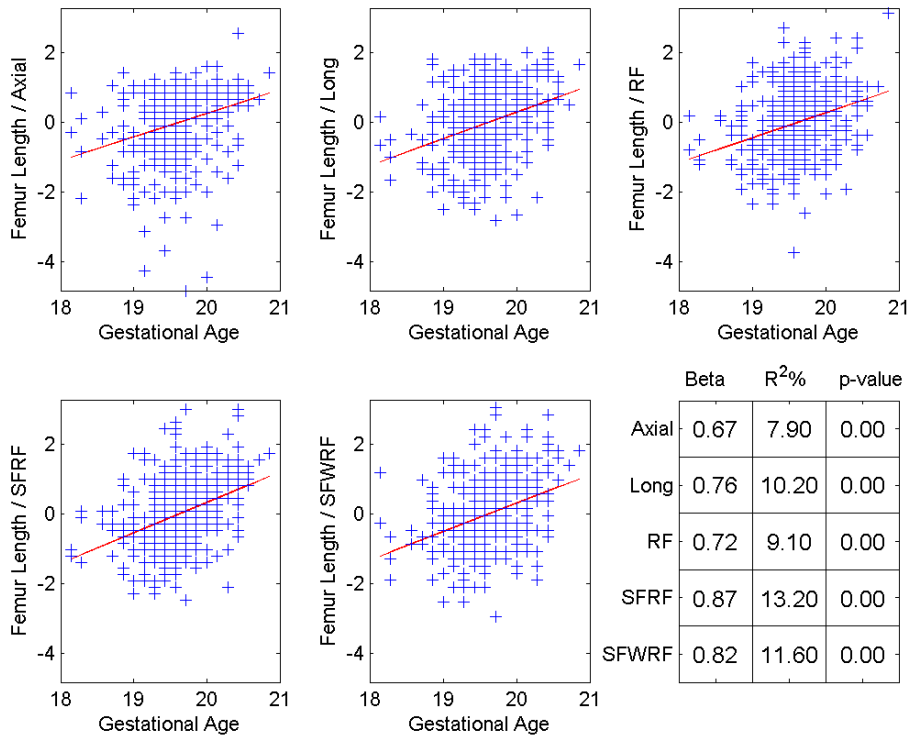


Figure 5.4. Linear regression analysis between normalised femur length and gestational age of fetuses. Each sub-figure varies depending on the segmentation technique (axial, longitudinal, classic RF, smart fast RF and smart fast-weighted RF).

Table 5.3. Relationship between SD femoral volume, femur length, average, maximum and minimum CSA, average, maximum and minimum splaying indices (PSI and MSI) and average, maximum and minimum of two splaying angles and 7 day increment in gestational age during second trimester. The relationship was repeated for the five variations of the technique. n is the number of cases in each group. The numbers here are in the following format: Beta [Stars] [(Confidence Interval)] where Beta represents the slope of the regression line, [Stars] is an optional field that represents the p-value. If the p-value >0.05 then no p-value and confidence interval are shown. If the p-value≤0.05 & >0.01 then one star is shown. If the p-value≤0.01 & >0.001 then two stars are shown. If the p-value≤0.001 then three stars are shown. [(Confidence Interval)] is shown if p-value≤0.05.

	n	436	225	211
		All	Girls	Boys
Axial	**FV (SD)**	0.399***(0.18, 0.62)	0.434**(0.13, 0.74)	0.357*(0.02, 0.69)
	FL (SD)	0.672***(0.46, 0.89)	0.568***(0.28, 0.86)	0.807***(0.48, 1.13)
	AvgCSA (SD)	0.244*(0.02, 0.47)	0.291	0.188
	MaxCSA (SD)	0.254*(0.03, 0.48)	0.234	0.285
	MinCSA (SD)	0.208	0.322*(0.02, 0.63)	0.062
	AvgPSI (SD)	0.072	0.145	-0.017
	MaxPSI (SD)	0.075	0.076	0.08
	MinPSI (SD)	0.058	0.2	-0.123
	AvgMSI (SD)	-0.136	-0.041	-0.253
	MaxMSI (SD)	-0.12	-0.099	-0.142
	MinMSI (SD)	-0.123	0.043	-0.334*(-0.66, -0.01)
	Avg Sp1 (SD)	-0.081	-0.105	-0.043
	Max Sp1 (SD)	0.181	0.311*(0.01, 0.61)	0.013
	Min Sp1 (SD)	0.139	0.236	0.011
	Avg Sp2 (SD)	-0.035	-0.023	-0.044
	Max Sp2 (SD)	0.097	0.153	0.022
	Min Sp2 (SD)	0.198	0.22	0.168
Long	**FV (SD)**	0.473***(0.25, 0.69)	0.521***(0.21, 0.83)	0.416*(0.1, 0.74)
	FL (SD)	0.766***(0.55, 0.98)	0.684***(0.39, 0.98)	0.88***(0.57, 1.19)
	AvgCSA (SD)	0.28*(0.06, 0.5)	0.339*(0.03, 0.65)	0.21
	MaxCSA (SD)	0.307**(0.08, 0.53)	0.331*(0.02, 0.64)	0.284
	MinCSA (SD)	0.218	0.314*(0.01, 0.62)	0.1
	AvgPSI (SD)	0.05	0.135	-0.054
	MaxPSI (SD)	0.085	0.13	0.032
	MinPSI (SD)	0.002	0.121	-0.147
	AvgMSI (SD)	-0.252*(-0.48, -0.03)	-0.146	-0.387*(-0.74, -0.04)
	MaxMSI (SD)	-0.203	-0.137	-0.285
	MinMSI (SD)	-0.269*(-0.49, -0.04)	-0.133	-0.443*(-0.79, -0.1)
	Avg Sp1 (SD)	0.027	-0.066	0.151
	Max Sp1 (SD)	0.069	0.289	-0.219
	Min Sp1 (SD)	-0.032	0.115	-0.23
	Avg Sp2 (SD)	0.012	-0.049	0.097
	Max Sp2 (SD)	0.052	0.158	-0.087
	Min Sp2 (SD)	0.059	0.134	-0.042
RF	**FV (SD)**	0.351**(0.13, 0.57)	0.365*(0.07, 0.66)	0.334
	FL (SD)	0.72***(0.5, 0.94)	0.571***(0.3, 0.85)	0.92***(0.58, 1.26)
	AvgCSA (SD)	0.072	0.062	0.089
	MaxCSA (SD)	0.169	0.16	0.184
	MinCSA (SD)	0.003	0.012	-0.006
	AvgPSI (SD)	-0.19	-0.153	-0.237
	MaxPSI (SD)	-0.053	-0.016	-0.1
	MinPSI (SD)	-0.248*(-0.46, -0.03)	-0.198	-0.309
	AvgMSI (SD)	-0.439***(-0.65, -0.23)	-0.364**(-0.64, -0.09)	-0.537**(-0.87, -0.21)
	MaxMSI (SD)	-0.298**(-0.52, -0.07)	-0.218	-0.404*(-0.76, -0.04)
	MinMSI (SD)	-0.478***(-0.69, -0.26)	-0.399**(-0.68, -0.12)	-0.58***(-0.92, -0.24)
	Avg Sp1 (SD)	-0.096	-0.141	-0.04
	Max Sp1 (SD)	0.018	-0.081	0.153
	Min Sp1 (SD)	0.084	0.051	0.13
	Avg Sp2 (SD)	-0.036	-0.039	-0.031
	Max Sp2 (SD)	0.114	-0.025	0.296
	Min Sp2 (SD)	0.089	-0.026	0.243
SFRF	**FV (SD)**	0.312**(0.09, 0.54)	0.408**(0.1, 0.72)	0.194

Table 5.3. Relationship between SD femoral volume, femur length, average, maximum and minimum CSA, average, maximum and minimum splaying indices (PSI and MSI) and average, maximum and minimum of two splaying angles and 7 day increment in gestational age during second trimester. The relationship was repeated for the five variations of the technique. n is the number of cases in each group. The numbers here are in the following format: Beta [Stars] [(Confidence Interval)] where Beta represents the slope of the regression line, [Stars] is an optional field that represents the p-value. If the p-value >0.05 then no p-value and confidence interval are shown. If the p-value≤0.05 & >0.01 then one star is shown. If the p-value≤0.01 & >0.001 then two stars are shown. If the p-value≤0.001 then three stars are shown. [(Confidence Interval)] is shown if p-value≤0.05.

	n	**436**	**225**	**211**
		All	**Girls**	**Boys**
SFRF	**FL (SD)**	0.877***(0.66, 1.09)	0.76***(0.47, 1.05)	1.036***(0.73, 1.35)
	AvgCSA (SD)	0.003	0.065	-0.072
	MaxCSA (SD)	0.029	0.079	-0.03
	MinCSA (SD)	-0.039	0.028	-0.12
	AvgPSI (SD)	-0.303**(-0.53, -0.08)	-0.214	-0.414*(-0.76, -0.07)
	MaxPSI (SD)	-0.227*(-0.45, 0)	-0.155	-0.315
	MinPSI (SD)	-0.338**(-0.56, -0.12)	-0.246	-0.455**(-0.79, -0.11)
	AvgMSI (SD)	-0.6***(-0.82, -0.38)	-0.495***(-0.78, -0.21)	-0.734***(-1.07, -0.4)
	MaxMSI (SD)	-0.49***(-0.71, -0.27)	-0.404**(-0.69, -0.12)	-0.598***(-0.94, -0.25)
	MinMSI (SD)	-0.611***(-0.83, -0.39)	-0.503***(-0.79, -0.21)	-0.748***(-1.08, -0.42)
	Avg Sp1 (SD)	0.072	0.091	0.05
	Max Sp1 (SD)	0.136	0.249	-0.008
	Min Sp1 (SD)	0.058	0.066	0.047
	Avg Sp2 (SD)	-0.002	0.066	-0.09
	Max Sp2 (SD)	0.132	0.2	0.045
	Min Sp2 (SD)	0.032	0.062	-0.009
SFWRF	**FV (SD)**	0.478***(0.25, 0.7)	0.512***(0.21, 0.82)	0.438**(0.11, 0.77)
	FL (SD)	0.818***(0.6, 1.03)	0.642***(0.36, 0.92)	1.052***(0.72, 1.38)
	AvgCSA (SD)	0.074	0.072	0.083
	MaxCSA (SD)	0.079	0.098	0.06
	MinCSA (SD)	0.02	0.021	0.022
	AvgPSI (SD)	-0.235*(-0.46, -0.01)	-0.189	-0.292
	MaxPSI (SD)	-0.178	-0.119	-0.25
	MinPSI (SD)	-0.269*(-0.49, -0.04)	-0.22	-0.33
	AvgMSI (SD)	-0.478***(-0.7, -0.26)	-0.4**(-0.68, -0.12)	-0.578***(-0.93, -0.22)
	MaxMSI (SD)	-0.401***(-0.62, -0.18)	-0.315*(-0.59, -0.03)	-0.512**(-0.87, -0.16)
	MinMSI (SD)	-0.488***(-0.71, -0.27)	-0.41**(-0.69, -0.13)	-0.59***(-0.94, -0.24)
	Avg Sp1 (SD)	0.152	0.166	0.136
	Max Sp1 (SD)	-0.055	0.002	-0.127
	Min Sp1 (SD)	0.011	-0.062	0.11
	Avg Sp2 (SD)	0.073	0.138	-0.01
	Max Sp2 (SD)	-0.058	-0.072	-0.039
	Min Sp2 (SD)	0.084	0.047	0.139

5.2.3.2 Maternal Vitamin D verses Automatic Measurements

To compare maternal vitamin D to the pre-mentioned automatic measurements (Table 5.1), two linear regression analysis experiments were conducted. These experiments studied:

1. the relationship between automatic measurements and 10 nmol increments in maternal 25OHD during the second trimester. See Table 5.4.
2. the relationship between automatic measurements and maternal vitamin D status of less than or equal 50nmol during the second trimester where maternal vitamin D status of greater than 50nmol was used as a reference group. See Table 5.5.

Note that these tables spread on two pages because they are large.

From Table 5.4, note that generally no significant correlation was found between the automatic measurements and 10 nmol increment in maternal 25OHD during second trimester although some statistical significance can be seen in girls (e.g., FV in Longitudinal segmentation) but not in boys. On the other hand, more significant correlation was found between some of the measurements and maternal vitamin D status (≤50nmol) during second trimester, see Table 5.5. Again, statistical correlation was found in girls but not in boys.

Table 5.4. Relationship between SD femoral volume, femur length, average, maximum and minimum CSA, average, maximum and minimum splaying indices (PSI and MSI) and average, maximum and minimum of two splaying angles and 10 nmol increment in maternal 25OHD during second trimester adjusting for gestational age. The relationship was repeated for the five variations of the technique. n is the number of cases in each group. The numbers here are in the following format: Beta [Stars] [(Confidence Interval)] where Beta represents the slope of the regression line, [Stars] is an optional field that represents the p-value. If the p-value >0.05 then no p-value and confidence interval are shown. If the p-value≤0.05 & >0.01 then one star is shown. If the p-value≤0.01 & >0.001 then two stars are shown. If the p-value≤0.001 then three stars are shown. [(Confidence Interval)] is shown if p-value≤0.05.

	n	363	184	179
		All	Girls	Boys
Axial	**FV (SD)**	-0.016	-0.036	0.005
	FL (SD)	-0.012	-0.025	0.002
	AvgCSA (SD)	-0.013	-0.027	0.002
	MaxCSA (SD)	-0.014	-0.018	-0.009
	MinCSA (SD)	-0.01	-0.033	0.014
	AvgPSI (SD)	-0.012	-0.023	0
	MaxPSI (SD)	-0.015	-0.014	-0.015
	MinPSI (SD)	-0.007	-0.03	0.017
	AvgMSI (SD)	-0.011	-0.017	-0.005
	MaxMSI (SD)	-0.015	-0.008	-0.022
	MinMSI (SD)	-0.003	-0.024	0.019
	Avg Sp1 (SD)	0.009	-0.002	0.022
	Max Sp1 (SD)	0.023	0.027	0.018
	Min Sp1 (SD)	0.02	0.009	0.03
	Avg Sp2 (SD)	0.019	0.013	0.027
	Max Sp2 (SD)	0.027	0.035	0.019
	Min Sp2 (SD)	-0.007	-0.033	0.02
Long	**FV (SD)**	-0.026	-0.054*(-0.1, 0)	0.006
	FL (SD)	-0.02	-0.022	-0.014
	AvgCSA (SD)	-0.026	-0.048	-0.001
	MaxCSA (SD)	-0.025	-0.038	-0.01
	MinCSA (SD)	-0.024	-0.055*(-0.1, -0.01)	0.01
	AvgPSI (SD)	-0.024	-0.048*(-0.1, 0)	0.003
	MaxPSI (SD)	-0.022	-0.035	-0.008
	MinPSI (SD)	-0.022	-0.057*(-0.11, -0.01)	0.015
	AvgMSI (SD)	-0.019	-0.043	0.006
	MaxMSI (SD)	-0.017	-0.028	-0.005
	MinMSI (SD)	-0.019	-0.054*(-0.1, -0.01)	0.019
	Avg Sp1 (SD)	-0.001	0.001	-0.003
	Max Sp1 (SD)	0.014	0.023	0.004
	Min Sp1 (SD)	0.015	0.026	0.001
	Avg Sp2 (SD)	0.007	0.024	-0.012
	Max Sp2 (SD)	0.021	0.051*(0, 0.1)	-0.011
	Min Sp2 (SD)	0.001	-0.015	0.017
RF	**FV (SD)**	0.003	-0.021	0.028
	FL (SD)	0.007	0.009	0.008
	AvgCSA (SD)	0.007	-0.012	0.027
	MaxCSA (SD)	0.015	0.002	0.029
	MinCSA (SD)	-0.01	-0.035	0.018
	AvgPSI (SD)	0.003	-0.018	0.024
	MaxPSI (SD)	0.013	-0.003	0.028
	MinPSI (SD)	-0.015	-0.041	0.013
	AvgMSI (SD)	0	-0.021	0.02
	MaxMSI (SD)	0.01	-0.007	0.027
	MinMSI (SD)	-0.018	-0.043	0.009
	Avg Sp1 (SD)	0.006	0.028	-0.018
	Max Sp1 (SD)	0.005	-0.005	0.017
	Min Sp1 (SD)	-0.019	-0.016	-0.021
	Avg Sp2 (SD)	0.017	0.044*(0.01, 0.08)	-0.013
	Max Sp2 (SD)	-0.004	-0.009	0.003
	Min Sp2 (SD)	-0.003	-0.002	-0.003
SFRF	**FV (SD)**	-0.008	-0.029	0.014
	FL (SD)	-0.012	-0.003	-0.02
	AvgCSA (SD)	-0.007	-0.026	0.014

Table 5.4. Relationship between SD femoral volume, femur length, average, maximum and minimum CSA, average, maximum and minimum splaying indices (PSI and MSI) and average, maximum and minimum of two splaying angles and 10 nmol increment in maternal 25OHD during second trimester adjusting for gestational age. The relationship was repeated for the five variations of the technique. n is the number of cases in each group. The numbers here are in the following format: Beta [Stars] [(Confidence Interval)] where Beta represents the slope of the regression line, [Stars] is an optional field that represents the p-value. If the p-value >0.05 then no p-value and confidence interval are shown. If the p-value≤0.05 & >0.01 then one star is shown. If the p-value≤0.01 & >0.001 then two stars are shown. If the p-value≤0.001 then three stars are shown. [(Confidence Interval)] is shown if p-value≤0.05.

	n	**363**	**184**	**179**
		All	**Girls**	**Boys**
SFRF	**MaxCSA (SD)**	-0.007	-0.013	-0.001
	MinCSA (SD)	-0.004	-0.04	0.034
	AvgPSI (SD)	-0.003	-0.027	0.023
	MaxPSI (SD)	-0.003	-0.011	0.005
	MinPSI (SD)	-0.001	-0.043	0.043
	AvgMSI (SD)	0.002	-0.024	0.028
	MaxMSI (SD)	0.002	-0.008	0.011
	MinMSI (SD)	0.002	-0.041	0.046
	Avg Sp1 (SD)	-0.027*(-0.05, 0)	-0.045*(-0.08, -0.01)	-0.009
	Max Sp1 (SD)	0.024	0.037	0.01
	Min Sp1 (SD)	0.006	0.005	0.008
	Avg Sp2 (SD)	-0.01	-0.018	-0.003
	Max Sp2 (SD)	-0.006	-0.016	0.004
	Min Sp2 (SD)	0.004	-0.002	0.01
SFWRF	**FV (SD)**	-0.02	-0.026	-0.012
	FL (SD)	-0.008	0.007	-0.019
	AvgCSA (SD)	-0.017	-0.021	-0.012
	MaxCSA (SD)	-0.016	-0.003	-0.029
	MinCSA (SD)	-0.012	-0.035	0.015
	AvgPSI (SD)	-0.013	-0.022	-0.004
	MaxPSI (SD)	-0.012	-0.003	-0.021
	MinPSI (SD)	-0.008	-0.036	0.021
	AvgMSI (SD)	-0.008	-0.02	0.005
	MaxMSI (SD)	-0.007	-0.003	-0.012
	MinMSI (SD)	-0.005	-0.033	0.025
	Avg Sp1 (SD)	-0.005	0.004	-0.015
	Max Sp1 (SD)	0.016	0.027	0.004
	Min Sp1 (SD)	-0.019	-0.003	-0.033
	Avg Sp2 (SD)	0.002	0.025	-0.023
	Max Sp2 (SD)	0.003	0.011	-0.005
	Min Sp2 (SD)	-0.008	0.013	-0.027

Table 5.5. Relationship between SD femoral volume, femur length, average, maximum and minimum CSA, average, maximum and minimum splaying indices (PSI and MSI) and average, maximum and minimum of two splaying angles and maternal vitamin D status (≤50nmol) during second trimester where maternal vitamin D status of (>50nmol) was used as a reference group adjusting for gestational age. The relationship was repeated for the five variations of the technique. n is the number of cases in each group. The numbers here are in the following format: Beta [Stars] [(Confidence Interval)] where Beta represents the slope of the regression line, [Stars] is an optional field that represents the p-value. If the p-value >0.05 then no p-value and confidence interval are shown. If the p-value≤0.05 & >0.01 then one star is shown. If the p-value≤0.01 & >0.001 then two stars are shown. If the p-value≤0.001 then three stars are shown. [(Confidence Interval)] is shown if p-value≤0.05.

	n	363	184	179
		All	Girls	Boys
Axial	**FV (SD)**	-0.2	-0.373*(-0.67, -0.07)	-0.015
	FL (SD)	-0.023	-0.19	0.144
	AvgCSA (SD)	0.185	0.331*(0.03, 0.64)	0.027
	MaxCSA (SD)	-0.22*(-0.43, 0)	-0.295	-0.136
	MinCSA (SD)	0.129	0.335*(0.04, 0.64)	-0.096
	AvgPSI (SD)	0.191	0.302	0.066
	MaxPSI (SD)	-0.234*(-0.45, -0.02)	-0.26	-0.199
	MinPSI (SD)	0.113	0.301	-0.095
	AvgMSI (SD)	0.189	0.243	0.123
	MaxMSI (SD)	-0.232*(-0.45, -0.01)	-0.199	-0.258
	MinMSI (SD)	0.089	0.239	-0.08
	Avg Sp1 (SD)	-0.019	-0.024	-0.01
	Max Sp1 (SD)	0.141	0.2	0.086
	Min Sp1 (SD)	0.101	0.067	0.137
	Avg Sp2 (SD)	0.024	0.041	0.013
	Max Sp2 (SD)	0.191	0.341*(0.04, 0.64)	0.036
	Min Sp2 (SD)	-0.097	-0.217	0.025
Long	**FV (SD)**	-0.248*(-0.46, -0.04)	-0.473**(-0.78, -0.17)	-0.005
	FL (SD)	-0.088	-0.198	0.025
	AvgCSA (SD)	0.24*(0.03, 0.45)	0.432**(0.13, 0.73)	0.031
	MaxCSA (SD)	-0.276**(-0.49, -0.07)	-0.402**(-0.71, -0.1)	-0.138
	MinCSA (SD)	0.171	0.422**(0.12, 0.72)	-0.099
	AvgPSI (SD)	0.233*(0.02, 0.44)	0.414**(0.12, 0.71)	0.035
	MaxPSI (SD)	-0.274*(-0.48, -0.06)	-0.371*(-0.67, -0.07)	-0.165
	MinPSI (SD)	0.152	0.406**(0.11, 0.7)	-0.123
	AvgMSI (SD)	0.203	0.344*(0.06, 0.63)	0.048
	MaxMSI (SD)	-0.245*(-0.45, -0.04)	-0.29*(-0.57, -0.01)	-0.191
	MinMSI (SD)	0.119	0.35*(0.05, 0.65)	-0.131
	Avg Sp1 (SD)	0.008	0.057	-0.043
	Max Sp1 (SD)	0.062	0.121	0.013
	Min Sp1 (SD)	-0.005	0.01	-0.023
	Avg Sp2 (SD)	0.044	0.137	-0.05
	Max Sp2 (SD)	0.133	0.297*(0, 0.59)	-0.042
	Min Sp2 (SD)	-0.028	-0.13	0.08
RF	**FV (SD)**	-0.074	-0.28	0.144
	FL (SD)	0.085	0.01	0.161
	AvgCSA (SD)	0.001	0.162	-0.172
	MaxCSA (SD)	0.01	-0.071	0.099
	MinCSA (SD)	0.056	0.324*(0.04, 0.61)	-0.227
	AvgPSI (SD)	0.032	0.183	-0.13
	MaxPSI (SD)	-0.018	-0.088	0.06
	MinPSI (SD)	0.087	0.347*(0.07, 0.63)	-0.19
	AvgMSI (SD)	0.059	0.176	-0.068
	MaxMSI (SD)	-0.045	-0.091	0.008
	MinMSI (SD)	0.108	0.331*(0.06, 0.6)	-0.13
	Avg Sp1 (SD)	0.025	0.193	-0.151
	Max Sp1 (SD)	-0.053	-0.149	0.045
	Min Sp1 (SD)	-0.107	-0.108	-0.108
	Avg Sp2 (SD)	0.081	0.256*(0.03, 0.48)	-0.098
	Max Sp2 (SD)	0.009	-0.057	0.069
	Min Sp2 (SD)	-0.07	-0.053	-0.093
SFRF	**FV (SD)**	-0.135	-0.284	0.033
	FL (SD)	-0.043	-0.062	-0.026

Table 5.5. Relationship between SD femoral volume, femur length, average, maximum and minimum CSA, average, maximum and minimum splaying indices (PSI and MSI) and average, maximum and minimum of two splaying angles and maternal vitamin D status (≤50nmol) during second trimester where maternal vitamin D status of (>50nmol) was used as a reference group adjusting for gestational age. The relationship was repeated for the five variations of the technique. n is the number of cases in each group. The numbers here are in the following format: Beta [Stars] [(Confidence Interval)] where Beta represents the slope of the regression line, [Stars] is an optional field that represents the p-value. If the p-value >0.05 then no p-value and confidence interval are shown. If the p-value≤0.05 & >0.01 then one star is shown. If the p-value≤0.01 & >0.001 then two stars are shown. If the p-value≤0.001 then three stars are shown. [(Confidence Interval)] is shown if p-value≤0.05.

	n	**363**	**184**	**179**
		All	**Girls**	**Boys**
SFRF	**AvgCSA (SD)**	0.125	0.26	-0.025
	MaxCSA (SD)	-0.15	-0.202	-0.087
	MinCSA (SD)	0.06	0.289	-0.191
	AvgPSI (SD)	0.111	0.246	-0.042
	MaxPSI (SD)	-0.139	-0.185	-0.084
	MinPSI (SD)	0.039	0.274	-0.217
	AvgMSI (SD)	0.082	0.199	-0.049
	MaxMSI (SD)	-0.116	-0.146	-0.078
	MinMSI (SD)	0.015	0.226	-0.216
	Avg Sp1 (SD)	-0.104	-0.162	-0.042
	Max Sp1 (SD)	0.047	0.113	-0.017
	Min Sp1 (SD)	0.071	-0.011	0.151
	Avg Sp2 (SD)	-0.048	-0.128	0.04
	Max Sp2 (SD)	0.052	0.009	0.098
	Min Sp2 (SD)	0.026	-0.016	0.068
SFWRF	**FV (SD)**	-0.106	-0.157	-0.049
	FL (SD)	0.026	0.103	-0.063
	AvgCSA (SD)	0.071	0.114	0.024
	MaxCSA (SD)	-0.102	-0.048	-0.157
	MinCSA (SD)	0.011	0.166	-0.154
	AvgPSI (SD)	0.065	0.143	-0.02
	MaxPSI (SD)	-0.098	-0.073	-0.121
	MinPSI (SD)	0.005	0.189	-0.191
	AvgMSI (SD)	0.048	0.143	-0.055
	MaxMSI (SD)	-0.082	-0.082	-0.079
	MinMSI (SD)	-0.004	0.182	-0.205
	Avg Sp1 (SD)	0.031	0.07	-0.009
	Max Sp1 (SD)	0.035	0.121	-0.049
	Min Sp1 (SD)	-0.092	-0.104	-0.084
	Avg Sp2 (SD)	0.016	0.112	-0.081
	Max Sp2 (SD)	0.084	0.144	0.021
	Min Sp2 (SD)	-0.148	-0.052	-0.25

5.3 Discussion

In this chapter, several statistical tests have been performed for the automatic measurements calculated using five variations of the Random Forests classifier. Just by looking at Table 5.1 one can see that these five variations of the segmentation technique provide slightly different measurements. There are a few points to highlight

1. Most variations of the segmentation technique have approximately similar mean measurements. See Table 5.1, Figure 5.1 and Figure 5.2.
2. SFWRF provides less standard deviation than the other four. This could mean that SFWRF is more reproducible and gives a more precise measure. However, this may also mean that SFWRF is less generalisable which can make the technique over-fit on new cases. There is currently no clear answer to why SFWRF has tighter spread around the mean but this could have been happened because of its strict choice of features, See Chapter 3 for more information.
3. The classic RF provides bimodal distribution for some measurements, for instance Figure 5.1, which means that the segmented region using this technique fall in two regions of different distribution. This also proves that the classic RF segment more non-femur tissues therefore it over-segments the femur. We also noted this point in technical terms in Section 3.4.1.
4. The different variations of the segmentation technique led to similar correlation coefficients within a technique for the different measures.
5. When examining the effect of fetal gestational age on femoral measurements, all the variations of the technique tended to perform the same.
6. Positive correlation was found between the fetal gestation and the femur volume, femur length and to some extent cross-sectional area.

7. Negative correlation was found between the fetal gestation age and MSI.
8. When examining the effect of vitamin D on femoral measurements, the 2D segmentation on the longitudinal view tended to perform better than the others.
9. Stronger statistical relationship between the measurements and fetal gestational age and maternal vitamin D level was found in girls than in boys. This is a very interesting note to further investigate since we believe that a biological factor leads to this phenomenon.

Chapter 6

Conclusions and Future Work

In this thesis we have developed techniques to quantify 3D fetal femoral ultrasound images. Important considerations we made when developing these techniques were accuracy, precision (reproducibility), automation, real-time implementation and clinical correlation with fetal and maternal characteristics. We achieved our goals with excellent results. In this chapter, we summarise and conclude our findings. However, we believe that this is still the beginning of a relatively under-developed area of fetal medicine (fetal bone segmentation and quantification) therefore we provide a few directions for future studies.

6.1 Conclusions

Our main contributions spread over three main chapters (Chapter 3, Chapter 4 and Chapter 5). In Chapter 3, we provided a solution to how to accurately, precisely and automatically quantify fetal femur in 3D ultrasound images. In Chapter 4, we tackled the problem of poor image quality of the 3D fetal femoral ultrasound compared to other medical image modalities, e.g., MRI and CT. Finally, clinical correlations and analyses

were performed in Chapter 5 to assess the relationship between the automatically quantified measurements and fetal and maternal characteristics.

6.1.1 Image Segmentation using Smart + Fast + Weighted Random Forests

Although the goal of the thesis was motivated by its clinical use (quantification of fetal femoral characteristics in 3D ultrasound images), we developed our work in Chapter 3 independently toward a generic technique which is able to segment different structures in a variety of medical image modalities. In Chapter 3, we developed a novel technique based on the Random Forests classifier. The new technique produced better segmentation accuracy compared to the classic Random Forests classifier. The added offline feature selection step provided a smaller subset of "good" features which helped build stronger trees. The use of the global score that is calculated during the offline feature selection helped provide much faster training while keeping excellent accuracy of the technique. Moreover, the weighted voting mechanism in the testing stage reduced the bias toward weaker trees. These ideas have led to a more accurate and precise measurement of different fetal femoral characteristics, e.g., the femur volume, length, cross-sectional areas, etc. The variations of the Random Forests technique were validated on segmenting the fetal femur in 3D ultrasound images and adult brain structures in MRI volumes. All in all, Smart Weighted Random Forests and Smart Fast-Weighted Random Forests showed superior accuracy compared to the others (see Table 3.2 on page 71 and Table 3.5 on page 74). Furthermore, Smart Fast-Weighted Random Forests required less training time but slightly more testing time. As a result, we conclude that the SWRF provides the best accuracy but SFWRF provides comparable accuracy to SWRF while being much faster during training. We also believe that SFWRF is more generalisable than SWRF. All these variations have

showed to provide a novel and well-validated framework for different medical image segmentation problems.

6.1.2 Image Fusion

In Chapter 3, we showed that the accuracy of the developed segmentation technique is better than many state-of-the-art techniques. However, the accuracy is still relatively low (in segmenting the fetal femur: mean Jaccard index 69%, see Table 3.2 on page 71) compared to the achievable accuracy in different segmentation problems for instance segmenting the left ventricle from cardiac MR (Yun et al., 2010). This is mainly because of the poor quality of 3D ultrasound images compared to other medical image modalities as well as the huge attenuation of the ultrasound beam when it hits the fetal femur. Therefore, we developed in Chapter 4 a technique that enhances the quality of 3D fetal femoral ultrasound images by combining several volumes of the same femur acquired from slightly different scanning angles. After acquiring these volumes, the process starts with aligning these volumes (image registration) and then merging the aligned volumes into one (image fusion).

In Chapter 4, a wavelet-based fusion technique in which a novel decomposition of wavelet coefficients showed better enhancement of the femur definition than two standard fusion techniques (Mean and Max intensity fusion). We also showed that fetal biometry on the fused volume is more reproducible than on single views especially for volumetric measurements. In addition, we showed that the fused view captures genuine missing information by comparing the fused view and all single views with a CT scan of one fetal femur. Our conclusion is that image fusion in 3D fetal femoral ultrasound leads not only to a more accurate biometry but also to a more reproducible biometry which in turn to a better diagnosis.

6.1.3 Automatic Measurements verses Fetal Gestational Age and Maternal Vitamin D Concentration during Pregnancy

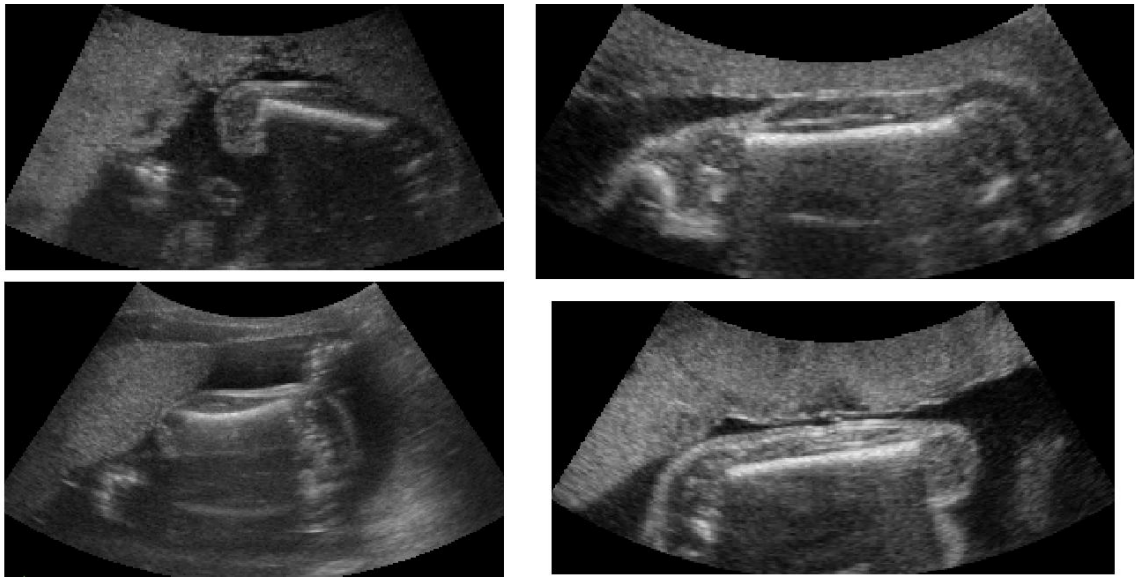
The clinical merit of our thesis was illustrated in Chapter 5. In this chapter, we studied the relationship between several measurements that were automatically found and fetal gestational age and maternal vitamin D level during pregnancy. A statistical significance between fetal gestational age and a few geometric femoral measurements such as the femur volume, length and splaying index (MSI = CSA/FL^2) was found. Finally, the relationship between maternal vitamin D and a few automatic femoral measurements in girls was found to be stronger than in boys. This suggests that low maternal vitamin D level during pregnancy affects female fetuses more than male fetuses.

6.2 Future Work

We believe that there are several possible promising extensions of this thesis in both technical and clinical directions. We split the suggested directions of improvement into 1) dataset, 2) segmentation technique, 3) fusion technique and 4) clinical evaluation.

6.2.1 Dataset

The main ultrasound dataset (SWS) used in this thesis is now approximately nine years old. With the advancement of ultrasound machines, it would be interesting to apply the proposed segmentation technique on ultrasound images from a state-of-the-art ultrasound machine. We believe that a more accurate segmentation can be achieved compared to that achieved on the SWS dataset. For instance, Figure 6.1 (a) shows a 2D slice from the SWS dataset while Figure 6.1 (b) shows an equivalent 2D slice from a recent machine. Note how there is significantly higher contrast, spatial resolution and detail of the femur and surrounding tissues.



(a) From GE Voluson 730 (b) From Philips HD9 System

Figure 6.1. Two 2D slices from different 3D volumes of two ultrasound machines namely (a) GE Voluson 730 and (b) Philips HD9 System.

6.2.2 Segmentation Using The Random Forests Technique

There are several ways to extend the segmentation framework. Here are a few suggestions:

1. *Applications*. The proposed new Random Forests could be applied to different segmentation problems. In fact, we have already helped other colleagues to apply the classic Random Forests classifier to segment different structures in ultrasound images for instance, the fetal skull in 3D ultrasound, the left ventricle in 3D adult echocardiography and the fetal stomach in 2D ultrasound. They all showed promising results.
2. *Feature Sets*. We have used classic features that can be calculated quickly although more advanced ones can be utilised. For instance, local phase might be used to provide a richer description of images since it has been shown to have good discrimination power in 3D Echocardiography images (Rajpoot et al.,

2009a). Furthermore, there are other feature descriptors that showed a good discrimination power in natural image classification like Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005, Schroff et al., 2008), Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), etc. However, these features are computationally expensive compared to the classic ones we used in this thesis.

3. *Object Detection*. Although the goal of the proposed technique was to provide an image segmentation solution, the technique can be adjusted to do object detection. This can be done by creating training examples from regions that contain the object(s) of interest instead of voxels, i.e., features that represent regions. Recently, huge interest in automatic object detection in medical images has been demonstrated and a few good ideas have been developed (Carneiro et al., 2007, Carneiro et al., 2008, Criminisi et al., 2009, Criminisi et al., 2010).
4. *True Parallelism*. Although we implemented our technique such that each tree can be trained or tested on a separate processor, better parallelism could be achieved. In Chapter 3, we mentioned that our implementation can achieve a factor of min(number of trees, number of processors) speed up relative to a sequential implementation. On the other hand, if the number of processors is more than the number of trees, our current implementation cannot utilise the extra processors and each tree will run on one processor. True parallelism could be achieved by utilising all available processors and the recursive training or testing of one tree might run on several processors. This would significantly reduce the training and testing time. In addition, a GPU implementation of the segmentation technique might give a massive reduction in training and testing time (Sharp, 2008).

6.2.3 Image Fusion

We validated how the fused image genuinely compensates for missing information by comparing it with one CT scan. Further validation needs to be done to validate the technique on a reasonable number of cases. We also studied the effect of image fusion on manual segmentation and measurement. However, how image fusion affects the automatic segmentation would be an interesting issue to investigate. Finally, we believe that ultrasound manufacturers should consider embedding a real-time functionality of an ultrasound volume fusion which will hopefully provide better quality volumes and in turn a more accurate diagnosis.

6.2.4 Clinical Evaluation

In spite of the weak correlation we found between maternal vitamin D and some of the automatic measurements, several other maternal predictors can be used to assess the correlation with the automatic measurements like, maternal nutrition, education, age, origin, etc. In addition, in the linear regression analysis we did in Chapter 5, given that vitamin D concentration can be higher in mothers during Summer and lower during Winter because of low sun lights which is the main source of vitamin D in people, the season in which the ultrasound image for each fetus was acquired can be used to group fetuses into categories. This might be an important factor to use in the regression analysis. Finally, a randomised control trail study instead of a cohort study (SWS) on mothers of low level of vitamin D can answer more subtle issues.

Appendix A. Studies Ethics

A. Ultrasound Assessment Of Fetal Musculo-Skeletal Development (SWS)

An approval for the Southampton Women's Survey protocols was granted by the Southampton and South West Hampshire Joint Research Ethics Committee. Ethics approval for this survey and the use of data collected was granted by the Local Research Ethics Committee Nos: 307/97, 089/99,153/99.

B. lntergrowth-21st

This research study is the International Fetal and Newborn Growth Consortium for the 21st Century. The Research Ethics Committee is an advisory committee to South Central Strategic Health Authority. The REC reference number is 08/H0606/139.

C. CT/Ultrasound Post-mortem Case

This case was provided from the 3D Fetal Imaging study where the main objective is to compare fetal bone volumes measured with 3D obstetric ultrasound to those measured with a post-mortem CT. The REC reference number is 09/H0605/124.

Bibliography

[1] AMIT, Y. & GEMAN, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, **9**, 1545-1588.

[2] ANQUEZ, J., ANGELINI, E. D. & BLOCK, I. (2008). Segmentation of Fetal 3d Ultrasound Based on Statistical Prior and Deformable Model. *International Symposium on Biomedical Imaging*, 17 - 20.

[3] ARCHIP, N., ROHLING, R., COOPERBERG, P. & TAHMASEBPOUR, H. (2005). Ultrasound Image Segmentation Using Spectral Clustering. *Ultrasound in medicine & biology*, **31**, 1485 - 1497.

[4] ASBMR (2006). *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism*, Rittenhouse Book Distributors.

[5] ASHBURNER, J. & FRISTON, K. (1997). Multimodal Image Coregistration and Partitioning - a Unified Framework. *NeuroImage*, **6**, 209-217.

[6] BALAFAR, M., RAMLI, A., SARIPAN, M. & MASHOHOR, S. (2010). Review of Brain Mri Image Segmentation Methods. *Artificial Intelligence Review*, **33**, 261-274.

[7] BOSCH, A., ZISSERMAN, A. & MUOZ, X. (2007). Image Classification Using Random Forests and Ferns. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.

[8] BREIMAN, L. (1996). Bagging Predictors. *Machine Learning*, **24**, 123-140.

[9] BREIMAN, L. (2001). Random Forests. *Machine Learning* **45**, 5-32.

[10] BREIMAN, L., FRIEDMAN, J., STONE, C. J. & OLSHEN, R. A. (1984). *Classification and Regression Trees*, Chapman and Hall/CRC.

[11] CAPEL, D. & ZISSERMAN, A. (2003). Computer Vision Applied to Super Resolution. *Signal Processing Magazine, IEEE*, **20**, 75-86.

[12] CARNEIRO, G., GEORGESCU, B., GOOD, S. & COMANICIU, D. (2007). Automatic Fetal Measurements in Ultrasound Using Constrained Probabilistic Boosting Tree. *Medical Image Computing and Computer Assisted Intervention*, 571-579.

[13] CARNEIRO, G., GEORGESCU, B., GOOD, S. & COMANICIU, D. (2008). Detection and Measurement of Fetal Anatomies from Ultrasound Images Using

a Constrained Probabilistic Boosting Tree. *IEEE Transactions on Medical Imaging*, **27**, 1342 - 1355

[14] CARUANA, R. & NICULESCU-MIZIL, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *International Conference on Machine Learning*.

[15] CHALANA, V., WINTER, T. C. I., DR, C., DR, H. & KIM, Y. (1996). Automatic Fetal Head Measurements from Sonographic Images. *Academic Radiology*, **3**, 628 - 635.

[16] CHANG, C.-H., TSAI, P.-Y., YU, C.-H., KO, H.-C. & CHANG, F.-M. (2007). Prenatal Detection of Fetal Growth Restriction by Fetal Femur Volume: Efficacy Assessment Using Three-Dimensional Ultrasound. *Ultrasound in Medicine & Biology*, **33**, 335-341.

[17] CHEN, C., LIAW, A. & BREIMAN, L. (2004). Using Random Forest to Learn Imbalanced Data. Technical Report 666, Statistics Department, University of California at Berkeley.

[18] COOPER, C., JAVAID, M. K., TAYLOR, P., WALKER-BONE, K., DENNISON, E. & ARDEN, N. (2002). Fetal Origins of Osteoporotic Fracture. *Calcified Tissue International*, **70**, 391 - 394.

[19] COOPER, C., WESTLAKE, S., HARVEY, N., JAVAID, K., DENNISON, E. & HANSON, M. (2006). Review: Developmental Origins of Osteoporotic Fracture. *Osteoporosis International*, **17**, 337-347.

[20] CRIMINISI, A., SHOTTON, J. & BUCCIARELLI, S. (2009). Decision Forests with Long-Range Spatial Context for Organ Localization in Ct Volumes. *MICCAI workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA)*.

[21] CRIMINISI, A., SHOTTON, J., ROBINSON, D. & KONUKOGLU, E. (2010). Regression Forests for Efficient Anatomy Detection and Localization in Ct Studies. *MICCAI workshop in Medical Computer Vision 2010: Recognition Techniques and Applications in Medical Imaging*.

[22] DALAL, N. & TRIGGS, B. (2005). Histograms of Oriented Gradients for Human Detection. *IEEE Computer Vision and Pattern Recognition*, 2, 886–893.

[23] DENNISON, E. M., ARDEN, N. K., KEEN, R. W., SYDDALL, H., DAY, I. N. M., SPECTOR, T. D. & COOPER, C. (2001). Birthweight, Vitamin D Receptor Genotype and the Programming of Osteoporosis. *Paediatric and Perinatal Epidemiology*, **15**, 211-219.

[24] FENSTER, A. & DOWNEY, D. B. (1996). 3-D Ultrasound Imaging: A Review. *Engineering in Medicine and Biology Magazine, IEEE*, **15**, 41-51.

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

[26] GALE, C., ROBINSON, S., HARVEY, N., JAVAID, M., JIANG, B., MARTYN, C., GODFREY, K., COOPER, C. & GROUP, P. A. H. S. (2008). Maternal Vitamin D Status During Pregnancy and Child Outcomes. *European Journal of Clinical Nutrition*, **62**, 68-77.

[27] GEREMIA, E., MENZE, B. H., CLATZ, O., KONUKOGLU, E., CRIMINISI, A. & AYACHE, N. (2010). Spatial Decision Forests for Ms Lesion Segmentation in Multi-Channel Mr Images. *Medical Image Computing and Computer Assisted Intervention*.

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

[29] GRAU, V., BECHER, H. & NOBLE, J. A. (2007). Registration of Multiview Real-Time 3-D Echocardiographic Sequence. *IEEE Transactions on Medical Imaging*, **26**, 1154-65.

[30] HALL, T. J., BILGEN, M., INSANA, M. F. & KROUSKOP, T. A. (1997). Phantom Materials for Elastography. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, **44**, 1355-1365.

[31] HANNA, C. & YOUSSEF, A.-B. (1997). Automated Measurements in Obstetric Ultrasound Images. *IEEE International Conference on Image Processing*, 3.

[32] HARALICK, R. M. (1983). Image Segmentation Survey. *Fundamentals in computer vision*, 209-223.

[33] IBANEZ, L., SCHROEDER, W., NG, L. & CATES, J. (2005). *The Itk Software Guide*, Kitware Inc.

[34] Http://Www.Medscinet.Net/Intergrowth/Default.Aspx. INTERGROWTH-21ST PROJECT. Accessed on 18/10/2010.

[35] IOANNOU, C., YAQUB, M., NOBLE, A., JAVAID, K. & PAPAGEORGHIOU, A. (2010). Fetal Femur Volume Measurement Using Fusion of Multiple Three-Dimensional Ultrasound Image Sets: A Pilot Study. *Ultrasound in Obstetrics & Gynecology*, **36**.

[36] JARDIM, S. & FIGUEIREDO, M. (2003). Automatic Contour Estimation in Fetal Ultrasound Images. *International Conference on Image processing*, 3, 1065-1068.

[37] JARDIM, S. & FIGUEIREDO, M. (2005). Segmentation of Fetal Ultrasound Images. *Ultrasound in Medicine & Biology*, **31**, 243 - 250.

[38] JAVAID, K. (2005). *Maternal Nutrition, Maternal Body Composition During Pregnancy and Neonatal Bone Mass*, PhD Thesis. University of Southampton, Southampton.

[39] JAVAID, K. & COOPER, C. (2002). Prenatal and Childhood Influences on Osteoporosis. *Best Practice and Research Clinical Endocrinology and Metabolism*, **16**, 349 - 367.

[40] JAVAID, M. K., CROZIER, S. R., HARVEY, N. C., GALE, C. R., DENNISON, E. M., BOUCHER, B. J., ARDEN, N. K., GODFREY, K. M., COOPER, C. & GROUP, P. A. H. S. (2006). Maternal Vitamin D Status During Pregnancy and Childhood Bone Mass at Age 9 Years: A Longitudinal Study. *Lancet*, **367**, 36-43.

[41] KANIS, J. A., JOHNELL, O., ODEN, A., JONSSON, B., DE LAET, C. & DAWSON, A. (2000). Risk of Hip Fracture According to the World Health Organization Criteria for Osteopenia and Osteoporosis. *Bone*, **27**, 585-590.

[42] KING, A. P., RHODE, K. S., MA, Y., YAO, C., JANSEN, C., RAZAVI, R. & PENNEY, G. P. (2010). Registering Preprocedure Volumetric Images with Intraprocedure 3-D Ultrasound Using an Ultrasound Imaging Model. *IEEE Transactions on Medical Imaging*, **29**, 924-937.

[43] LEMPITSKY, V., VERHOEK, M., NOBLE, J. A. & BLAKE, A. (2009). Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3d Echocardiography. *Proceedings of the 5th International Conference on Functional Imaging and Modeling of the Heart*.

[44] LEPETIT, V. & FUA, P. (2006). Keypoint Recognition Using Randomized Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 1465-1479.

[45] LIU, H., XIE, C., CHEN, Z. & LEI, Y. (2006). Segmentation of Ultrasound Image Based on Morphological Operation and Fuzzy Clustering. *Proceedings of the Third IEEE International Workshop on Electronic Design, Test and Applications*, 397-400.

[46] LOWE, D. G. (1999). Object Recognition from Local Scale-Invariant Features. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on.

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

[48] LUPASCU, C. A., TEGOLO, D. & TRUCCO, E. (2010). Fabc: Retinal Vessel Segmentation Using Adaboost. *Information Technology in Biomedicine, IEEE Transactions on*, **14**, 1267-1274.

[49] MAHON, P. A. (2007). *Ultrasound Assessment of Fetal Musculo-Skeletal Development*, PhD Thesis. University of Southampton, Southampton.

[50] MAHON, P. A., COOPER, C., CROZIER, S. R. & GODFREY, K. M. (2009). The Use of 3d Ultrasound to Investigate Fetal Bone Development. *Norsk Epidemiologi*, **19**.

[51] MALLAPRAGADA, V. G., SARKAR, N. & PODDER, T. K. (2008). Autonomous Coordination of Imaging and Tumor Manipulation for Robot Assisted Breast Biopsy. Biomedical Robotics and Biomechatronics, 2008. BioRob 2008. 2nd IEEE RAS & EMBS International Conference on.

[52] 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**, 1317 - 1324.

[53] NOBLE, J. A. & BOUKERROUI, D. (2006). Ultrasound Image Segmentation: A Survey. *IEEE Transactions on Medical Imaging*, **25**, 987-1010.

[54] OREN, M., PAPAGEORGIOU, C., SINHA, P., OSUNA, E. & POGGIO, T. (1997). Pedestrian Detection Using Wavelet Templates. *Computer Vision and Pattern Recognition*, 193-199.

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

[56] PEI, Y., CRIMINISI, A., WINN, J. & ESSA, M. (2007). Tree-Based Classifiers for Bilayer Video Segmentation. Computer Vision and Pattern Recognition.

[57] PHAM, D. L., XU, C. & PRINCE, J. L. (2000). Current Methods in Medical Image Segmentation. Annual Review of Biomedical Engineering.

[58] PICKUP, L. C., CAPEL, D. P., ROBERTS, S. J. & ZISSERMAN, A. (2009). Bayesian Methods for Image Super-Resolution. *Comput. J.*, **52**, 101-113.

[59] PRAGER, R. W., IJAZ, U. Z., GEE, A. H. & TREECE, G. M. (2010). Three-Dimensional Ultrasound Imaging. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, **224**, 193-223.

[60] PRINCE, R. L. & GLENDENNING, P. (2004). 8: Disorders of Bone and Mineral Other Than Osteoporosis. *The Medical Journal of Australia*, **180**, 354-359.

[61] RAJPOOT, K. (2009). *Multi-View 3d Echocardiographic Image Analysis*, PhD Thesis. University of Oxford, Oxford.

[62] RAJPOOT, K., GRAU, V. & NOBLE, J. A. (2009a). Local-Phase Based 3d Boundary Detection Using Monogenic Signal and Its Application to Real-Time 3-D Echocardiography Images. *IEEE international conference on Symposium on Biomedical Imaging: From Nano to Macro*.

[63] RAJPOOT, K., NOBLE, J. A., GRAU, V., SZMIGIELSKI, C. & BECHER, H. (2009b). Multiview Rt3d Echocardiography Image Fusion. *International Conference on Functional Imaging and Modeling of the Heart*, 134-143.

[64] RAUCH, F. & SCHOENAU, E. (2002). Skeletal Development in Premature Infants: A Review of Bone Physiology Beyond Nutritional Aspects. *Arch Dis Child Fetal Neonatal Ed*, **86**, F82-F85.

[65] SCHROFF, F., CRIMINISI, A. & ZISSERMAN, A. (2008). Object Class Segmentation Using Random Forests. *Proceedings of the British Machine Vision Conference*.

[66] SHAPIRO, L. G. & STOCKMAN, G. C. (2001). *Computer Vision*, Prentice Hall.

[67] SHARP, T. (2008). Implementing Decision Trees and Forests on a Gpu. *European Conference on Computer Vision*, 595-608.

[68] SHIMIZU, A., NARIHIRA, T., FURUKAWA, D., KOBATAKE, H., NAWANO, S. & SHINOZAKI, K. (2008). Ensemble Segmentation Using Adaboost with Application to Liver Lesion Extraction from a Ct Volume. *MICCAI*.

[69] SHOTTON, J., JOHNSON, M. & CIPOLLA, R. (2008). Semantic Texton Forests for Image Categorization and Segmentation. *Computer Vision and Pattern Recognition*.

[70] SMIGIELSKI, C., RAJPOOT, K., GRAU, V., MYERSON, S., HOLLOWAY, C., NOBLE, J. A., KERBER, R. & BECHER, H. (2010). Four-Dimensional Fusion Echocardiography: A New Technique for Fusion of Real-Time Three-Dimensional Datasets Obtained from Different Transducer Positions: First Clinical Experience. *Journal of the American College of Cardiology*.

[71] SOLER, P., GERARD, O., ALLAIN, P., SALOUX, E., ANGELINI, E. & BLOCH, I. (2005). Comparison of Fusion Techniques for 3d+T Echocardiography Acquisitions from Different Acoustic Windows. *Computers in Cardiology*.

[72] STATNIKOV, A. & ALIFERIS, C. F. (2007). Are Random Forests Better Than Support Vector Machines for Microarray-Based Cancer Classification? *AMIA Annual Symposium Proceedings*, 686-90.

[73] SUBRAMANIAN, K., LAWRENCE, D. & MOSTAFAVI, M. T. (1997). Interactive Segmentation and Analysis of Fetal Ultrasound Images. *8 th Eurographics Workshop on Visualization in Scientific Computing*.

[74] THOMAS, J., JOUVE, P. E. & NICOLOYANNIS, N. (2006). Optimisation and Evaluation of Random Forests for Imbalanced Datasets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4203 LNAI**, 622-631.

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

[76] TU, Z. (2005). Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering. *International Conference on Computer Vision*.

[77] TU, Z., NARR, K., DOLLAR, P., DINOV, I., THOMPSON, P. & TOGA, A. (2008). Brain Anatomical Structure Segmentation by Hybrid Discriminative/Generative Models. *IEEE Transactions on Medical Imaging*, 27, 495-508.

[78] TUSTISON, N. J., AVANTS, B. B., COOK, P. A., YUANJIE, Z., EGAN, A., YUSHKEVICH, P. A. & GEE, J. C. (2010). N4itk: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, **29**, 1310-1320.

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

[80] Http://En.Wikipedia.Org/Wiki/Endochondral_Ossification, Last Modified 10, March 2009. WIKIPEDIA. Accessed on 2009

[81] The Insight Segmentation and Registration Toolkit. WWW.ITK.ORG. Accessed on 2009.

[82] XU, R. S., MICHAILOVICH, O. & SALAMA, M. (2010). Information Tracking Approach to Segmentation of Ultrasound Imagery of the Prostate. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, **57**, 1748-1761.

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

[84] YAQUB, M., IOANNOU, C., PAPAGEORGHIOU, A., JAVAID, K., COOPER, C. & NOBLE, A. (2011a). Evaluating Image Fusion for Improved Fetal Femoral Boundary Definition and Quantification. *Ultrasound in Medicine & Biology, in press*.

[85] YAQUB, M., JAVAID, K., COOPER, C. & NOBLE, A. (2011b). Efficient Volumetric Segmentation Using 3d Fast-Weighted Random Forests *Medical Image Computing and Computer Assisted Intervention-MLMI, Submitted*.

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

[87] YI, Z., CRIMINISI, A., SHOTTON, J. & BLAKE, A. (2009). Discriminative, Semantic Segmentation of Brain Tissue in Mr Images. *Medical Image Computing and Computer Assisted Intervention*, 12, 558-65.

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

[89] YU JIN, Z. (2001). A Review of Recent Evaluation Methods for Image Segmentation. Signal Processing and its Applications, Sixth International, Symposium on. 2001.

[90] YUN, Z., PAPADEMETRIS, X., SINUSAS, A. J. & DUNCAN, J. S. (2010). Segmentation of the Left Ventricle from Cardiac Mr Images Using a Subject-Specific Dynamical Model. *IEEE Transactions on Medical Imaging*, **29**, 669-687.

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

[92] ZAYED, N., BADWI, A., ELSAYAD, A., ELSHERIF, M. & YOUSSEF, A.-B. (2001). Wavelet Segmentation for Fetal Ultrasound Images. *Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems, 2001. MWSCAS 2001*, 1, 501-504.

[93] ZHANG, Y., BRADY, M. & SMITH, S. (2001). Segmentation of Brain Mr Images through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm. *IEEE Transactions on Medical Imaging*, **20**, 45-57.