Automatic Characterization of Developmental Dysplasia of the Hip in Infants using Ultrasound Imaging

by

Niamul Quader

B.Sc. in Engineering, Islamic University of Technology, 2010
M.Sc. in Engineering, Islamic University of Technology, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Electrical and Computer Engineering)

The University of British Columbia
(Vancouver)

February 2018

© Niamul Quader, 2018
Abstract

Developmental dysplasia of the hip (DDH) is the most common pediatric hip condition, representing a spectrum of hip abnormalities ranging from mild dysplasia to irreducible hip dislocation. Thirty-three years ago, the introduction of the Graf method revolutionized the use of ultrasound (US) and replaced radiography for DDH diagnoses. However, it has been shown that current US-based assessments suffer from large inter-rater and intra-rater variabilities which can lead to misdiagnosis and inappropriate treatment for DDH.

In this thesis, we propose an automatic dysplasia metric estimator based on US and hypothesize that it significantly reduces the subjective variability inherent in the manual measurement of dysplasia metrics. To this end, we have developed an intensity invariant feature to accurately extract bone boundaries in US images, and have further developed an image processing pipeline to automatically discard US images which are inadequate for measuring dysplasia metrics, as defined by expert radiologists. If found adequate, our method automatically measures clinical dysplasia metrics from the US image. We validated our method on US images of 165 hips acquired through clinical examinations, and found that automatic extraction of dysplasia metrics improved the repeatability of diagnoses by 20%.

We extended our automatic metric extraction method to three-dimensional (3D) US to increase robustness against operator dependent transducer placement and to better capture the 3D morphology of an infant hip. We present a new random forests-based method for segmenting the femoral head from a 3D US volume, and a method for automatically estimating a 3D femoral head coverage measurement from the segmented head. We propose an additional 3D hip morphology-derived dysplasia metric for identifying an unstable acetabulum. On 40 clinical hip ex-
aminations, we found our methods significantly improved the reproducibility of diagnosing femoral head coverage by 65% and acetabular abnormalities by 75% when compared to current standard methods.
Lay Summary

Many babies are born with unstable hips which can cause severe mobility issues as they grow older. This condition is known as developmental dysplasia of the hip (DDH), and although it is the most common hip disorder in infants, its diagnosis has been shown to be prone to large variability. The goal of this thesis is to develop methods to automatically identify DDH in infants, using ultrasound (US) imaging. We demonstrate two-dimensional (2D) and three-dimensional (3D) ultrasound systems that automatically extract clinical measurements which can be used by clinicians to diagnose DDH. We show that our automatic 2D measurements improve reliability in repeated diagnoses by 20% on a dataset of 165 clinical hip examinations. We demonstrate a much more considerable improvement of 70% in repeated diagnoses with 3D US-based measurements, evaluated on 40 infant hip examinations.
Preface

The research presented herein was approved by the UBC Clinical Research Ethics Board (CREB), certificate numbers: H14–01448 and H17–01904. This thesis is primarily based on the following articles, resulting from collaboration of multiple researchers.

Studies described in Chapter 3 have been published in:


Studies described in Chapter 4 have been published in:


Studies described in Chapter 5 have been published in:


All published articles were revised and edited by all co-authors. I was the primary author of all of these publications.

[P1, P2]: I contributed to the article’s idea, method implementation and validation scheme under the supervision of Drs. Rafeef Abgharbieh and Antony J. Hodgson.

[P3]: I contributed to the article’s idea, method implementation of method and validation scheme under the supervision of Drs. Rafeef Abgharbieh and Dr. Antony J. Hodgson. Dr. Thomas Savage contributed in retrieving ultrasound data and in providing annotations for the retrieved ultrasound images. Dr. Kishore Mulpuri provided clinical feedback for the study.

[P5, P7]: I contributed to the article’s idea, design of the data collection protocol, implementation of method and validation scheme under the supervision of Drs. Rafeef Abgharbieh and Antony J. Hodgson. Drs. Kishore Mulpuri and Emily Schaeffer contributed in retrieving 2D ultrasound data from infants at British Columbia Children’s hospital and in providing clinical feedback.

[P6]: I contributed to the article’s idea, design of data collection protocol, implementation of method and validation scheme under the supervision of Drs. Rafeef Abgharbieh and Antony J. Hodgson. Dr. Kishore Mulpuri contributed in retrieving 2D ultrasound data from infants at British Columbia Children’s hospital and in providing clinical feedback.

[P4, P8]: I contributed to the article’s idea, design of data collection protocol, implementation of method and validation scheme under the supervision of Drs. Rafeef Abgharbieh and Antony J. Hodgson. Drs. Anthony Cooper and Kishore Mulpuri contributed in acquiring 2D and 3D ultrasound data from infants at British Columbia Children’s hospital and in providing clinical feedback including annota-
tions for the scanned ultrasound images.
# Table of Contents

Abstract .................................................................................. ii
Lay Summary ........................................................................ iv
Preface .................................................................................... v
Table of Contents .................................................................... ix
List of Tables ........................................................................... xiii
List of Figures ........................................................................... xv
Glossary .................................................................................... xxx
Acknowledgments ..................................................................... xxxii
Dedication ................................................................................ xxxiii

1 Introduction ................................................................. 1
  1.1 Thesis Motivation ...................................................... 1
  1.2 Thesis Objectives ..................................................... 3
  1.3 A Clinical Evaluation of Developmental Dysplasia of the Hip 4
    1.3.1 US-based Diagnosis ........................................... 4
  1.4 Towards Minimal User-Interaction in Extracting Dysplasia Metrics 10
    1.4.1 US Bone Imaging .............................................. 10
    1.4.2 US Adequate Plane Detection ............................ 15
1.4.3 Dysplasia Metric Extraction .................................................. 13
1.5 Research Questions Addressed .................................................. 15
1.6 Thesis Overview ................................................................. 17

2 A Systematic Review and Meta-analysis of Variability in Dysplasia
Metrics ................................................................. 19
2.1 Introduction ................................................................. 19
2.2 Study Identification .......................................................... 21
  2.2.1 Search Strategy ......................................................... 22
  2.2.2 Results ................................................................. 22
  2.2.3 Discussion .............................................................. 24
2.3 Methodologic Quality Appraisal ............................................. 25
  2.3.1 Results ................................................................. 26
  2.3.2 Discussion .............................................................. 26
2.4 Meta-analysis of variability in dysplasia metrics ....................... 27
  2.4.1 Extracting Variability and Agreement Statistics .................... 27
  2.4.2 Summarizing Variability Statistics .................................. 27
  2.4.3 Summarizing Agreement Statistics .................................. 28
  2.4.4 Comparing Dysplasia Metrics ....................................... 28
  2.4.5 Estimating Trend of Variability over Time ......................... 28
  2.4.6 Results ................................................................. 29
  2.4.7 Discussion .............................................................. 38
2.5 Summary ................................................................. 45

3 Symmetry and Attenuation-based Ultrasound Bone Imaging ........ 46
3.1 Introduction ................................................................. 46
3.2 Confidence-Weighted Local Phase Features ............................. 49
  3.2.1 Local Phase Symmetry Feature ..................................... 50
  3.2.2 Attenuation-related Features ...................................... 51
  3.2.3 Combined Feature for Bone Surface Localization ................ 53
  3.2.4 Experiment ........................................................... 54
  3.2.5 Discussion ............................................................ 56
3.3 Structured Phase Symmetry ................................................. 58
3.3.1 2D Filters-based Structured Phase Symmetry . . . . . . . 58
3.3.2 3D Filters-based Structured Phase Symmetry . . . . . . . 60
3.3.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . 66
3.4 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66

4 Characterizing Hip Joint using 2D Ultrasound . . . . . . . . . . 68
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
4.2 Localizing Bone and Cartilage Boundaries . . . . . . . . . . . 70
4.2.1 Ilium, Labrum and Acetabulum . . . . . . . . . . . . . . 71
4.2.2 Femoral Head . . . . . . . . . . . . . . . . . . . . . . . . 72
4.3 Scan Adequacy Classification . . . . . . . . . . . . . . . . . . . 75
4.3.1 Experiment . . . . . . . . . . . . . . . . . . . . . . . . . 77
4.3.2 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . 80
4.4 Extraction of 2D Dysplasia Metrics . . . . . . . . . . . . . . . 82
4.4.1 Experiment . . . . . . . . . . . . . . . . . . . . . . . . . 83
4.4.2 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.5 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99

5 3D Ultrasound-based Hip Dysplasia Assessment . . . . . . . . . . 101
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
5.2 Defining 3D Hip Morphology-derived Dysplasia Metrics . . . . 102
5.3 Extracting $\alpha_{3D}$ . . . . . . . . . . . . . . . . . . . . . . 102
5.3.1 Localizing $I$ and $A$ Across the Coronal Plane . . . . . . . 104
5.3.2 Extracting 3D Bone Boundary . . . . . . . . . . . . . . . 104
5.3.3 Approximating Planes of $I$ and $A$ . . . . . . . . . . . . . 105
5.3.4 Calculating $\alpha_{3D}$ . . . . . . . . . . . . . . . . . . . . . 107
5.4 Extracting $FHC_{3D}$ . . . . . . . . . . . . . . . . . . . . . . . 107
5.4.1 Femoral Head Segmentation . . . . . . . . . . . . . . . . . 108
5.4.2 Enhancing Femoral Head Segmentation . . . . . . . . . . . 108
5.4.3 Estimating $FHC_{3D}$ . . . . . . . . . . . . . . . . . . . . . 109
5.5 Experiment, Results and Discussion . . . . . . . . . . . . . . . . 109
5.5.1 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . 121
5.6 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125
6 Conclusions ................................................................. 127

6.1 Thesis Contributions ............................................... 127

6.1.1 A Systematic Review and Meta-analysis of Variability in Dysplasia Metrics .................................................. 128

6.1.2 Symmetry and Attenuation-based US Bone Imaging ... 128

6.1.3 Characterizing Hip Joint using 2D US ....................... 129

6.1.4 A 3D US-based Hip Dysplasia Assessment ............... 130

6.1.5 Summary of Technical Contributions ....................... 131

6.2 Future Work .......................................................... 133

Bibliography ............................................................... 136

A Examples of Automatic Dysplasia Metric Extraction from 2D US Images ................................................................. 153

B Examples of Automatic Dysplasia Metric Extraction from 3D US Images ................................................................. 156
List of Tables

Table 2.1  Initial logic grid aligned with the population-intervention-comparison-outcome (PICO) elements of the review question: For infants at risk of DDH, are US imaging-based diagnoses reproducible? Note that although the comparison intervention is left empty in here, we will be comparing all the US imaging-based diagnoses with each other. ........................................... 22

Table 2.2  Logic Grid with identified keywords added in each of the PICO columns. .............................................. 23

Table 2.3  Logic Grid with keywords and index terms or MeSH headings in each of the PICO columns [mh = mesh headings, kw = keyword, exp=expode]. .............................................. 23

Table 2.4  Final search strategy using keywords and MeSH [mh = mesh headings, kw = keyword, exp=expode]. .......................... 24

Table 2.5  Criteria for evaluating normal and dysplastic hips along with ranges between normal and hip dysplasia for each of the dysplasia metrics. .............................................. 29

Table 2.6  Meta-analysis result for the total variability (σT) in α angle (in °), β angle (in °) and FHC (in %). ................................. 33

Table 2.7  Meta-analysis result for the total variability (σT) in ACA angle (in °) and H angle (in °) and FHC (in %). ........................ 38

Table 4.1  Features used in adequacy classifier. ................................. 78
Table 4.2  Graf-DDH classification of the neonatal hips. Here, type-I ($\alpha > 60^\circ$, mature hip), type-II ($60^\circ < \alpha < 43^\circ$, physiological immature hip) and type-III ($\alpha < 43^\circ$, eccentric hip). The agreement between automatic and manual classification is fair (Cohen’s kappa coefficient=0.61 (CI:0.48 and 0.75)).

Table 5.1  Success rates in acquiring adequate 3D US images and 2D US imageset.
List of Figures

Figure 1.1 A flow-chart outlining the thesis objectives. The red box identifies the 3D US adequacy step, not being addressed in this thesis, necessary for an automatic processing of a 3D US image.

Figure 1.2 The propagation of an ultrasound pulse (white) along one particular scanline (dotted red line). A strong echo is reflected from the ilium bone boundary resulting in a high brightness measure on the image at the bone boundary point.

Figure 1.3 (a) and (b) display the anatomy surrounding a hip joint, with (a) illustrating the femur and (b) the hip bone. Note that the femoral head is not ossified. (c) A B-mode US image of an infant hip acquired from angle of the coronal plane. This is also an example of an adequate US image since this image includes the labrum, ischium, femoral head, acetabulum and flat horizontal ilium. (d), (e) and (f) illustrate example measurements of $\alpha$, $\beta$ and $FHC$.

Figure 1.4 A flow-chart outlining the research questions and core blocks of this thesis. P1 to P10 are publications associated with this thesis (note that P1 has been submitted and P10 is currently being prepared for submission). The vertical arrow bars to the right provide a visual direction to the components of this thesis described in the various chapters.
Figure 2.1  Flow diagram demonstrating study identification process. The reasons for excluding a study were based on the inclusion and exclusion criteria.

Figure 2.2  Variability in $\alpha$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.3  Variability in $\beta$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.4  Variability in $FHC$. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.5  Variability in $ACA$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.6  Variability in $PFD$. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.7  Variability in $M$ measure. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.8  Variability in $H$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.9  Variability in dysplasia metrics based on our meta analysis. Here, the variability of each dysplasia metric was standardized by dividing $\sigma_T$ by the range of that dysplasia metric between normal and dysplastic hips. The inter-exam, inter-rater variability is the most clinically relevant variability.

Figure 2.10  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $\alpha$.

Figure 2.11  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $\beta$.

Figure 2.12  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $FHC$.

Figure 2.13  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $AROC$.

Figure 2.14  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $AL$. 

xvi
Figure 2.15  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for \( PFD \).

Figure 2.16  Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for \( M \).

Figure 2.17  Agreement in dysplasia metrics, expressed as Kappa coefficient between repeated measurements.

Figure 2.18  Forest plot showing correlation coefficients between variability of \( \alpha \) and year of study. Top rows show mean and 95% confidence intervals of the correlation between intra-exam, intra-image, intra-user variability and year of study for \( \alpha \). Last row show the total random effects correlation coefficient estimating the effective change in variability with year of study for \( \alpha \). Here, the summary measure is the center of diamond, and the associated confidence intervals of the summary measure are the lateral tips of the diamond.

Figure 2.19  Forest plot showing correlation coefficients between variability of \( \beta \) and year of study. Top rows show mean and 95% confidence intervals of the correlation between intra-exam, intra-image, intra-user variability and year of study for \( \beta \). Last row show the total random effects correlation coefficient estimating the effective change in variability with year of study for \( \beta \).

Figure 2.20  Forest plot showing correlation coefficients between variability of \( FH\) and year of study. Top rows show mean and 95% confidence intervals of the correlation between intra-exam, intra-image, intra-user variability and year of study for \( FH\). Last row show the total random effects correlation coefficient estimating the effective change in variability with year of study for \( FH\).
Figure 3.1 Flowchart for local phase analysis for a 3D volume collected around the iliac crest area of an adult trauma patient (blue box on segmented CT representing approximate location of US transducer). Here, 3D log-Gabor filtering on the 3D US image results in real and imaginary components, which are used with Equation 3.4 for computing PS response. CPS feature is computed using Equation 3.7. White arrows point to false positive bone surface points.

Figure 3.2 Qualitative results: segmented bone surfaces around bovine femur (first column) and around in-vivo pelvis in human trauma patients. First row shows segmented CT with box representing approximate location of US transducer. Second row shows corresponding US volume. Third row shows PS based on empirical parameters [57]. Last row shows the CPS response.

Figure 3.3 Quantitative results: (a) Bovine phantom. Note that our proposed CPS based segmentation resulted in a 0.302 mm reduction in error compared to PS. (b) In-vivo pelvic data across all subjects (C-1 to C-18). Note that our proposed CPS resulted in a reduction in error which is significant at ($p < 0.01$) based on Wilcoxon signed rank test compared to PS.

Figure 3.4 Example qualitative results. (a) B-mode US image of the femoral head, (b) monogenic signal based PS, (c) directional filter based PS, (d) SPS. Red circles point to false negatives, and green circles point to falsely straightened structures, on the labrum, ilium, acetabulum, and triradiate cartilage. Blue circles point to falsely connected structures. Along the bone and cartilages contours of labrum, ilium, acetabulum, and triradiate cartilage, SPS has less false positive and false negative.
Figure 3.5 Example qualitative results. (a), (h) B-mode US volumes. (b), (i) US volume segments along the posterior-anterior axis that contains the ilium, acetabulum and labrum. (c), (j) OPS responses of (b) and (i), respectively. (d), (k) CSPS responses of (b) and (i), respectively. (e), (l) US slices near the middle of the femoral head chosen by the orthopaedic surgeon for the US volumes (b) and (i), respectively. Yellow, green and red contours are manually labeled ilium, acetabulum and labrum, respectively. (f) and (m) show OPS-based contours of ilium, acetabulum and labrum. (g) and (n) show CSPS-based contours of ilium, acetabulum and labrum. White lines point to false positive bone boundaries.

Figure 3.6 Cumulative distribution function of absolute discrepancy between manually-labeled ilium and OPS- and CSPS-based ilium boundaries.

Figure 3.7 Cumulative distribution function of absolute discrepancy between manually-labeled labrum and OPS- and CSPS-based labrum boundaries.

Figure 3.8 Cumulative distribution function of absolute discrepancy between manually-labeled acetabulum and OPS- and CSPS-based acetabulum boundaries.

Figure 4.1 Example coronal US images of the neonatal hip. (a) Adequate US image showing key hip joint structures. (b) Inadequate US image (missing ilium and labrum). (c), (d) The $\alpha$, $\beta$ angles and FHC measurements extracted manually from the adequate US image in (a).

Figure 4.2 A flow-chart outlining the steps involved in our computer-assisted 2D US-based DDH diagnosis.
Figure 4.3 Overview of bone and cartilage extraction: (a) Example B-mode US image showing the acetabular roof, ilium, labrum, ischium and femoral head. (b) SPS capturing ridges of the B-mode US image. (c) Relative signal strength map or confidence map of the B-mode US image. (d) Segmented bone boundary from SPS, with white arrow pointing to inferior edge of ilium, i.

Figure 4.4 Block diagram showing extraction of I, L and A.

Figure 4.5 Features used in localizing femoral head. Note that only the first HOG feature is shown in here.

Figure 4.6 (a) B-mode US image. (b) Output of femoral head random forests classifier, p. (c) Sobel filtering on p generates edges of the femoral head. (d) Circular Hough transform applied on (c). Peak of circular Hough transform (yellow arrow) provides an estimate of the center coordinates and radius of the femoral head.

Figure 4.7 Out-of-bag classification error vs. number of grown trees in random forest classifiers for adequacy classification. Out-of-bag classification error does not seem to decay considerably after number of trees = 70.

Figure 4.8 Example scan adequacy classification results. (a) Typical 2D B-mode images that were judged adequate by a radiologist (i.e., identified manually). Arrows in the images point to the key structures that were identified manually. In an adequate image, all of the key five structures need to be present. The first four images were classified as adequate by our method, while the last image was classified as inadequate (probability metric = 0.46). (b) Images that were judged to be inadequate by the radiologist. The first four images were classified as inadequate by our method, while the last image was classified as adequate (probability metric = 0.51). Based on the probability metrics produced by our classifier, both of these disagreements in classification could be considered to be borderline cases.
Figure 4.9  ROC curves of our 229 features-based random forests classifier. (a) The five different plots correspond to the five cross-validation experiments we conducted on the validation dataset. (b) The five different plots correspond to the five classifiers each from a different cross-validation dataset and evaluated on the test dataset. 

Figure 4.10  Example qualitative results of manual (labelled as 'M' at right top corners) and automatic measurements (labelled as 'A' at right top corners). (a), (c) and (e) show examples where manual and automatic measurements produced similar values for $\alpha$, $\beta$ and $FHC$, whereas (b), (d) and (f) show examples where the manual and automatic measurements differed noticeably. The white dotted lines represent the manual measurements, while the blue dotted lines represent the automatic measurements and the green in (e) and (f) represent femoral head random forests’ probability output.

Figure 4.11  Scatter plot of automatic versus manual measurements $\alpha$ angle. The red line is the equality line. Blue data points correspond to lower $\alpha_A$ and yellow data points correspond to higher $\alpha_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal.

Figure 4.12  Scatter plot of automatic versus manual measurements $\beta$ angle. The red line is the equality line. Blue data points correspond to lower $\beta_A$ and yellow data points correspond to higher $\beta_A$, so blue correspond to hips that would be judged as normal whereas yellow correspond to hips that would be judged as hip dysplasia.
Figure 4.13  Scatter plot of automatic versus manual measurements $FHC$ angle. The red line is the equality line. Blue data points correspond to lower $FHC_A$ and yellow data points correspond to higher $FHC_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal.  

Figure 4.14  Box-plot of the within-hip standard deviations among the manual and automatic measurements of $\alpha$ angles within all of the 165 hip examinations (i.e., 165 values of $\text{std}(\alpha_M)$ and 165 values of $\text{std}(\alpha_A)$). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The '+' points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in the automatic $\alpha$ angle measurements ($p < 0.01$).  

Figure 4.15  Box-plot of the within-hip standard deviations among the manual and automatic measurements of $\beta$ angles within all of the 165 hip examinations (i.e., 165 values of $\text{std}(\beta_M)$ and 165 values of $\text{std}(\beta_A)$). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The '+' points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in the automatic $\beta$ angle measurements ($p < 0.01$).
Figure 4.16  Box-plot of the within-hip standard deviations among the manual and automatic measurements of $FHC$ angles within all of the 165 hip examinations (i.e., 165 values of std($FHC_M$) and 165 values of std($FHC_A$)). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The ‘+’ points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in the automatic $FHC$ angle measurements ($p < 0.01$).

Figure 4.17  Scatter plot of variability of $\alpha_M$ vs. mean $\alpha_M$. Blue data points correspond to lower mean values of $\alpha_M$ and yellow data points correspond to higher mean values of $\alpha_M$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\alpha_M$ with the severity of hip dysplasia.

Figure 4.18  Scatter plot of variability of $\alpha_A$ vs. mean $\alpha_A$. Blue data points correspond to lower mean values of $\alpha_A$ and yellow data points correspond to higher mean values of $\alpha_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\alpha_A$ with the severity of hip dysplasia.
Figure 4.19  Scatter plot of variability of $\beta_M$ vs. mean $\beta_M$. Blue data points correspond to lower mean values of $\beta_M$ and yellow data points correspond to higher mean values of $\beta_M$, so blue correspond to hips that would be judged as normal whereas yellow correspond to hips that would be judged as hip dysplasia. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\beta_M$ with the severity of hip dysplasia.

Figure 4.20  Scatter plot of variability of $\beta_A$ vs. mean $\beta_A$. Blue data points correspond to lower mean values of $\beta_A$ and yellow data points correspond to higher mean values of $\beta_A$, so blue correspond to hips that would be judged as normal whereas yellow correspond to hips that would be judged as hip dysplasia. The red line is the best fit line. Correlation and p values suggest that there is no significant association in variability of $\beta_A$ with the severity of hip dysplasia.

Figure 4.21  Scatter plot of variability of $FHC_M$ vs. mean $FHC_M$. Blue data points correspond to lower mean values of $FHC_M$ and yellow data points correspond to higher mean values of $FHC_M$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is significant association in the variability of $FHC_M$ with the severity of hip dysplasia.

Figure 4.22  Scatter plot of variability of $FHC_A$ vs. mean $FHC_A$. Blue data points correspond to lower mean values of $FHC_A$ and yellow data points correspond to higher mean values of $FHC_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is significant association in the variability of $FHC_A$ with the severity of hip dysplasia.
Figure 5.1  (a) A rendering of the anatomy of a hip joint showing $A$ (red), $I$ (blue) and the femoral head. (b) A schematic illustration of $\alpha_{3D}$ - the angle between the normals to the fitted planar surfaces approximating $A$ and $I$. (c) A schematic illustration of $FHC_{3D}$, which is defined as the ratio of the volume of the femoral head portion that is medial to $I$ to the total volume of the femoral head.

Figure 5.2  Block diagram showing the extraction of $I$, $A$, femoral head, $\alpha_{3D}$ and $FHC_{3D}$.

Figure 5.3  Block diagram showing the steps for identifying the coronal plane slices that contain $I$ and $A$ (orange colored volume) from an US volume (grey colored volume).

Figure 5.4  (a), (b) and (c) show example US slices within an US volume that contain the $I$ and $A$ structures, whereas (d), (e) and (f) show example US slices in the same US volume that does not contain at least one of $I$ and $A$ structures. (g) Plot showing the slice numbers in the US volume that contain $I$ and $A$.

Figure 5.5  Block diagram showing the extraction of bone and cartilage boundaries from 3D US. Arrows in the last image are pointing to $A$, $I$ and labrum.

Figure 5.6  Overview of our slice-based probability map extraction. (a) Overlay of example US volume and a manually segmented femoral head. (b) $N$ number of $C$ were evaluated using classifier $R_2$ to determine their likelihood of intersecting the femoral head. (c) Back-projected likelihood scores, $L$, for each of the cross-sections $C$. (d) Back-projected responses were summed and normalized to construct voxel-wise probability map, with an overlay of the manually segmented femoral head (yellow).

Figure 5.7  Features used in the classifier $R_3$.

Figure 5.8  (a) Overlay of an example US volume and a manually segmented femoral head. (b) Overlay of the example US volume and its automatically extracted femoral head voxel-wise probability map $P$.  

xxv
Figure 5.9 Out-of-bag classification error for various numbers of grown trees used by the $R_1$ classifier. We selected 70 as being our number of trees for $R_1$. .......................................................... 112

Figure 5.10 Out-of-bag classification error for various numbers of grown trees used by the $R_2$ classifier. We selected 70 as being our number of trees for $R_2$. .......................................................... 113

Figure 5.11 Out-of-bag classification error for various numbers of grown trees used by the $R_3$ classifier. We selected 25 as being our number of trees for $R_3$. .......................................................... 113

Figure 5.12 Cumulative distribution function showing the relative frequency of different $s_d$ values in our data. Maximum value of $s_d$ was 3.5mm. .......................................................... 114

Figure 5.13 Box-plot showing the distribution in acquisition time for both 2D and 3D US for all the six participating surgeons. ....... 116

Figure 5.14 Box-plot showing distribution of acquisition time of both 2D and 3D US. Mean acquisition time of 3D US is slightly less than that of 2D US, however the change was not statistically significant. .......................................................... 116

Figure 5.15 Scatter plot of $FHC_{2D}$ vs. $FHC_{3D}$. The red line is the equality line. Blue data points correspond to lower $FHC_{3D}$ and yellow data points correspond to higher $FHC_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal. .............. 118

Figure 5.16 Scatter plot of $\alpha_{2D}$ vs. $\alpha_{3D}$. Blue data points correspond to lower $\alpha_{3D}$ and yellow data points correspond to higher $\alpha_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal. .......................................................... 119
Figure 5.17 Qualitative results. (a), (b), (d) and (e) show example variability of $\alpha_{2D}$, $\alpha_{3D}$, $FHC_{2D}$ and $FHC_{3D}$ from two 2D and two 3D US images from a hip examination ($\alpha_{2D} = 47^\circ$ and $56^\circ$, $\alpha_{3D} = 45.1^\circ$ and $45.9^\circ$, $FHC_{2D} = 51\%$ and $71\%$, $FHC_{3D} = 46.6\%$ and $47.8\%$). The higher variability in the input 2D US images (and $\alpha_{2D}$ and $FHC_{2D}$ values) can be seen in the manually aligned 2D US images in (c) compared to the lower variability in the manually aligned 3D US images (and $\alpha_{3D}$ and $FHC_{3D}$ values) in (f).

Figure 5.18 Box-plot of the within-hip standard deviations among $FHC_{2D}$ vs. $FHC_{3D}$ values within all of the hip examinations. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The ‘+’ points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to $99.3\%$ coverage if the data are normally distributed. The within-hip standard deviations are significantly less in $FHC_{3D}$ compared to $FHC_{2D}$ ($p < 0.01$).

Figure 5.19 Box-plot of the within-hip standard deviations among $\alpha_{2D}$ vs. $\alpha_{3D}$ values within all of the hip examinations. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The ‘+’ points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to $99.3\%$ coverage if the data are normally distributed. The within-hip standard deviations are significantly less in $\alpha_{3D}$ compared to $\alpha_{2D}$ ($p < 0.01$).
Figure 5.20 Scatter plot of variability of $FHC_{3D}$ vs. mean $FHC_{3D}$. Blue data points correspond to lower mean values of $FHC_{3D}$ and yellow data points correspond to higher mean values of $FHC_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $FHC_{3D}$ with the severity of hip dysplasia.

Figure 5.21 Scatter plot of variability of $\alpha_{3D}$ vs. mean $\alpha_{3D}$. Blue data points correspond to lower mean values of $\alpha_{3D}$ and yellow data points correspond to higher mean values of $\alpha_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\alpha_{3D}$ with the severity of hip dysplasia.

Figure 6.1 A flow-chart outlining the research questions, core blocks and key contributions of this thesis. P1 to P10 are publications associated with this thesis (note that P1 has been submitted and P10 is currently being prepared for submission). The vertical arrow bars to the right provide a visual direction to the components of this thesis described in the various chapters.

Figure A.1 Visualization for the extracted dysplasia metrics from a 2D US image in a healthy hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 70^\circ$, $\beta = 29^\circ$ and $FHC = 73.2\%$.

Figure A.2 Visualization for the extracted dysplasia metrics from a 2D US image in a borderline hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 47^\circ$, $\beta = 40^\circ$ and $FHC = 59.8\%$. 
Figure A.3 Visualization for the extracted dysplasia metrics from a 2D US image in a dysplastic hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 38^\circ$, $\beta = 56^\circ$ and $FHC = 41\%$.  

Figure B.1 Visualization for the extracted dysplasia metrics from a 3D US image in a healthy hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 71.8^\circ$ and $FHC = 72.4\%$. The visualization in here was done automatically using 2017a (MATLAB 2017a, the Mathworks Inc., Natick, MA, USA). The 3D plot generated using MATLAB was later manually rotated in 3D to make the planes of the ilium and acetabulum more apparent. In Chapter 5, the visualization was done manually using AMIRA software (TGS, San Diego, USA).

Figure B.2 Visualization for the extracted dysplasia metrics from a 3D US image in a borderline hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 53.1^\circ$ and $FHC = 47.9\%$. The visualization in here was done automatically using 2017a (MATLAB 2017a, the Mathworks Inc., Natick, MA, USA). The 3D plot generated using MATLAB was later manually rotated in 3D to make the planes of the ilium and acetabulum more apparent.

Figure B.3 Visualization for the extracted dysplasia metrics from a 3D US image in a dysplastic hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 35.6^\circ$ and $FHC = 22.3\%$. 

xxix
Glossary

2D  Two-dimensional
3D  Three-dimensional
ACA  Acetabular contact angle
AL  Arc length
AROC  Acetabular radius of curvature
AVN  Avascular necrosis
BRC  Bony rim coverage
CPS  Confidence-weighted phase symmetry
CSPS  Confidence-weighted structured phase symmetry
CT  Computed tomography
CNN  Convolutional neural networks
DDH  Developmental dysplasia of the hip
FHC  Femoral head coverage
GRRAS  Guidelines for reporting reliability and agreement Studies
HOG  Histogram of oriented gradients
ICC  Intraclass correlation coefficient

xxx
LBP  Local binary patterns
MESH  Medical subject headings
MRI  Magnetic resonance imaging
MSAC  M-estimator sample consensus
OA  Hip osteoarthritis
OPS  Optimized phase symmetry
PFD  Pubo-femoral distance
PS  Phase symmetry
PICO  Population-intervention-comparison-outcome
ROC  Receiver operating characteristics
SFE  Surface fitting error
SPS  Structured phase symmetry
US  Ultrasound
Acknowledgments

I would like to thank my advisors, Drs. Rafeef Abugharbieh and Antony J. Hodgson, for supervising me throughout my doctoral studies, and for spending countless hours towards helping me both within and beyond studies. I would also like to thank my clinical advisor, Dr. Kishore Mulpuri, for supervising my clinical study and for supporting me in my research.

Furthermore, I would like to thank my friends both inside and outside of BiSICL, Surgical Technology Lab, Orthopaedic Research at Children’s Hospital for all of their support. I would also like to thank Drs. Shayesteh Jahanfar and Boris Kuzeljevic for helping me with statistical analysis, Ursula Ellis and Charlotte Beck for helping me with searching articles for a systematic review, and Drs. Anthony Cooper, Thomas Savage and Emily Schaeffer for providing me with valuable clinical feedback. Also, I would like to thank Teo Tammie and Janice Andrade for patiently supporting and assisting me with my clinical study.

Last but not least, I would like to thank my parents, my wife, my sisters and all my family for always being there for me.
Dedication

To my parents, my wife, my sisters and all my family.
Chapter 1

Introduction

Developmental dysplasia of the hip (DDH) refers to a spectrum of hip abnormalities ranging from mild dysplasia with a stable hip, through subluxation, to total hip dislocation [36, 55, 79, 116, 145, 148]. DDH is the most common hip disorder in infants, affecting 0.16% to 2.85% of all newborns [11, 34, 55, 145, 148]. DDH can also cause considerable long-term debilitation if left untreated – e.g. early arthritis is often associated with DDH [145], while failure to diagnose and treat DDH in infancy can lead to costly corrective surgical procedures [119] later. It is therefore important to maximize the accuracy of diagnosing DDH. This thesis focuses on assessing and improving DDH’s diagnostic accuracy.

1.1 Thesis Motivation

Early diagnosis and treatment of DDH is widely recommended [28, 63, 116, 159, 163]. Early treatments include use of the Pavlic harness which is a common nonsurgical treatment that encourages increased hip joint abduction and flexion [132] and has high success rates (60% to 95%) [28, 63, 116, 159, 163]. A delay in early treatment can result in needing more complex treatments at a later age [119]. When the head of the femur is left in an abnormal position, the surrounding anatomy develops abnormally and this, if left untreated, can lead to the need for surgical correction to provide the joint with adequate stability and symmetry [55]. Joint reorientation procedures (i.e. pelvic osteotomies) may be performed initially, and many
patients then progress to requiring joint replacements early in life [55]. Furthermore, artificial hip joints have limited lifespans, so patients may require multiple revision procedures as they grow and develop [55, 163].

While the cost for diagnosing and treating DDH in infants is moderate (e.g. approximately 690 £ per infant receiving diagnosis and a Pavlik harness–based treatment per year in the United Kingdom [163]), failure to detect and treat DDH during early infancy can lead to considerable socioeconomic costs later in life. In particular, DDH in infancy is a major risk factor in the development of adult hip osteoarthritis (OA) [67, 68, 72, 100]. In 2013, Hoaglund performed a meta-analysis study which resulted in the conservative estimate that 10% of all OA patients also had DDH [72]; however, Nakamura [100] estimated that 88% in a study of 2000 consecutive OA patients in Japan had DDH, so the rates may vary considerably by country. Even when considering Hoaglund’s conservative estimate, Price [119] calculated that DDH might be responsible for approximately 25,000 hip replacement procedures annually in the United States. At approximately $50,000 per procedure [142], the direct financial impact of these hip replacements is in the order of $1.25 billion per year in the United States alone, not including the cost of multiple, more expensive revision cases later in life, or other associated economic costs for the patients and their families.

While early treatment of DDH can reduce complications in later life, misdiagnosis and over-treatment of DDH can lead to adverse consequences [112, 147]. The worst of these complications, avascular necrosis (AVN) (the death of bone tissues due to a lack of blood supply), is likewise the most frequent complication arising in both surgical and non-surgical treatments for DDH [112, 147]. A meta-analysis revealed that 1.35% to 10.9% of all infants undergoing treatment might have AVN [92]. The authors did not provide a comparison between the rates of AVN following surgical as opposed to non-surgical treatments. Based on the included studies in Shipman’s review paper [147], surgical treatments are associated with markedly higher rates of AVN compared to nonsurgical treatments (0% to 4% in nonsurgical treatments [21, 39, 117, 165] versus 6% to 46% in surgical treatments [18, 30, 118, 160]). To reduce the rates of DDH treatments leading to AVN, it is important to reduce the rate of treatments performed redundantly, and more importantly, to reduce the number of excessive surgical interventions. Thus, it is
essential that DDH be diagnosed early and reliably.

At an early age (from birth to 6 months of age), the femoral head in the hip joint is primarily cartilaginous, limiting the utility of plain radiographs in visualizing key structures until ossification begins at around the 4 month mark [4]. Also, misdiagnosis with radiographs can occur due to variability in pelvic rotation during image acquisition [155]. Furthermore, radiographs are associated with ionizing radiation [4, 22, 50, 99]. Consequently, US has become the primary imaging modality for detecting acetabular dysplasia and/or hip dislocation in infants at this early stage [4, 104, 137, 150]. However, despite considerable research on diagnosing DDH using US imaging [51, 52, 61, 62, 98], misdiagnosis still occurs frequently - for example, Jaremko demonstrated that through using a standard US-imaging technique, a DDH patient might be interpreted as having healthy hip joints, while a healthy infant could conversely be interpreted as having hip dysplasia [79]. This variability in DDH diagnosis can lead to non-trivial rates (e.g. 29% under-treatment rates [78] and 38% over-treatment rates [147]). Another concern is that, US is less frequently used in many of developing countries because of a lack of hip sonography specialists, particularly in rural and remote areas [7]. Even when diagnosis is done by specialists, manually extracting dysplasia metrics is prone to high variability and error [53, 79, 106, 137]. In this thesis, we hope to improve the repeatability of US-based DDH assessments, which may in turn lead to reduced misdiagnosis rates. Furthermore, we hope to increase the automation of the diagnosis process, which may enable less specialized clinicians to perform effective diagnostic tests.

1.2 Thesis Objectives

The overarching goal of this thesis is to assess and improve the repeatability of US imaging-based diagnosis. More formally, our first objective is to assess the repeatability of currently available dysplasia metrics (objective 1 in Figure 1.1). The rest of the thesis objectives are based on improving repeatability of DDH diagnosis by minimizing user-interaction in extracting dysplasia metrics (objectives 2 and 3 in Figure 1.1).

Towards minimizing user-interaction in extracting dysplasia metrics, our second objective is to automatically process 2D US images of the neonatal hip for
diagnosing DDH - this includes extracting bone and cartilage boundaries (objective 2A in Figure 1.1), assessing adequacy of 2D US images (i.e., whether an US image is adequate for making dysplasia metric measurements), extracting dysplasia metrics, and comparing reproducibility of dysplasia metric measurements using our method against reproducibility of dysplasia metric measurements using other available methods (objective 2B in Figure 1.1). Anticipating that 3D hip morphology-based DDH diagnosis will be more reliable, our final and third objective is to automatically process 3D US images of the neonatal hip for diagnosing DDH (Figure 1.1). In this 3D US-related objective, we focus on extracting bone and cartilage boundaries (objective 3A in Figure 1.1), extracting 3D morphology-derived dysplasia metrics, and comparing reproducibility of the 3D morphology-derived dysplasia metrics extracted using our method against reproducibility of dysplasia metric measurements using 2D US imaging-based methods (objective 3B in Figure 1.1). This thesis does not address the problem of identifying the scan adequacy of 3D US images – a graduate student is currently working on this issue.

1.3 A Clinical Evaluation of Developmental Dysplasia of the Hip

The current clinical practice for the management of DDH is to screen infants using a clinical and/or US-based diagnosis method [147]. Clinical examination tests the stability of the hip within its socket using the Barlow and Ortolani tests [55, 147]. Originally popularized by Graf [51], the US-based methods improve the sensitivity of clinical examination-based diagnoses [40, 54, 136, 163]. Other imaging modalities employed in diagnosing DDH are the X-ray, magnetic resonance imaging (MRI) and computed tomography (CT). X-rays are useful after the femoral epiphysis ossifies and are used as a primary tool after six months of age [4]. MRIs and CTs are only used in the pre- and postoperative evaluation of surgical hip treatments [4].

1.3.1 US-based Diagnosis

Currently, US is the primary imaging modality for detecting acetabular dysplasia and/or hip dislocations in infants at this early stage [104, 137, 150]. US is preferred
over X-rays for diagnosing DDH at early age since the latter involves ionization of matter directly along their path of travel – this can damage living tissue and the DNA within cells [70]. While US is considered safer than X-rays, there are some risks associated with US.

In this subsection, we first provide fundamentals of US imaging, with a focus on US bone imaging since diagnosing DDH requires imaging the bone and cartilage boundaries in an infant hip. We then present a short report on the potential thermal and non-thermal risks associated with performing an infant hip examination using US, and provide evidence suggesting that the risks associated with US imaging in DDH diagnosis is insignificant. Next, we report the currently recom-
mended US–based DDH diagnosis procedures and provide details on each of those procedures.

**Fundamentals of US Imaging of Bone Boundaries:** US images are obtained by using a pulse-echo approach. Here, an US transducer, which is formed by an array of piezoelectric crystals, transmits a spatially localized pulse of US and directs it into a patient along a US scanline (Figure 1.2). For convenience, in the rest of this thesis, we assume that US transducer is at the top of the US image regardless of whichever orientation the US transducer was placed during image acquisition (similar to Figure 1.2). Thus, a mention of an upward direction refers to direction towards the US transducer and a mention of downward direction refers to direction away from the US transducer. As the US pulse travels through a point in the patient, part of the US signal is reflected (US echo), part of the US signal gets absorbed and the remainder of the pulse continues deeper into the patient. These reflections at a particular point depends on the energy of the US pulse and also on the local ratio of acoustic impedance along the line of US pulse travel. When an US pulse enters a bone, the local ratio of acoustic impedance at the bone boundary is large since bone has a considerably larger acoustic impedance in comparison to neighboring tissues. The echo signals are collected at the transducer, and are then converted to brightness or luminance measures (Figure 1.2). The brightness measures corresponding to the bone boundary locations are larger than the brightness measures at points immediately before the bone boundary. Also, since bone strongly attenuates US signal, echoes from beyond the bone boundary are weak and thus brightness measures from points beyond the bone boundary are small. This strong response at the bone boundary along with weak response around the bone boundary generates a local symmetry (or a ridge-like appearance) of brightness measures around the bone boundary in an US image (Figure 1.2).

**Thermal Risks Associated with US Diagnosis:** When US is directed into a patient, there can be risks from heating of tissues. Draper [37] found an increase in temperature by 5°C in the gastrocnemius muscle at a depth of 3 cm when applying a continuous-echo US for 10 minutes with an US signal power of 1.5 W/cm². In
Figure 1.2: The propagation of an ultrasound pulse (white) along one particular scanline (dotted red line). A strong echo is reflected from the ilium bone boundary resulting in a high brightness measure on the image at the bone boundary point.

In practice, US signals are emitted at a power of around 0.7 W/cm² [146]. More importantly, the acquisition time of US is considerably shorter than 10 minutes (e.g. around 2 seconds for acquiring a 3D US volume). Also, we use a pulse-echo US approach which has a considerably smaller heating effect than that of a continuous–echo approach [6]. For all these reasons, we estimate that the increase in heating from a 3D US–based diagnosis would be around 1/1000th of 5°C, and the increase in heating from a 2D US–based diagnosis would be even smaller. Thus, it is unlikely that thermal risks are present in an US–based DDH assessment.

Non–thermal Risks Associated with US–based DDH Diagnosis: Baker’s [6] review article identified two non-thermal risks in association with US – cavitation (or the formation of tiny gas bubbles in the tissues as the result of US vibration) and acoustic streaming (or a localized liquid flow in the fluid around a vibrating bubble as the result of US vibration). Regarding cavitation risks, Baker recommended
caution in continuous long exposure (around 10 minutes) of US near air-filled cavities such as the lungs and intestines. Since our target anatomy is the hip joint and also since the acquisition time of US in DDH diagnosis is short (e.g. around 2 seconds for acquiring a 3D US volume), it is unlikely that cavitation risks are present in a 2D US–based DDH assessment. Regarding acoustic streaming, it is associated with and is secondary to cavitation in terms of risks [6], so it is unlikely that acoustic streaming risks are present in an US–based DDH assessment.

**Recommended US–based DDH Diagnosis Procedure:** The American College of Radiology [104] has standardized US-imaging-based methods for assessing DDH, which includes estimating the acetabular morphology using Graf’s method [51], estimating the femoral head coverage (FHC) using Morin’s method (optional) [98] and assessing the reducibility of the dislocated hip using Harcke’s method [62]. Outlines of these methods are as follows:

**Acetabular morphology assessments using Graf’s method:** Graf’s method [51] involves measuring two angle measurements, $\alpha$ and $\beta$. The $\alpha$ angle is the angle between the acetabular roof and the vertical cortex of the ilium, whereas the $\beta$ angle is that between the vertical cortex of the ilium and the labrum (Figure 1.3 (d) and (e)). A higher $\alpha$ corresponds to healthier hips. The threshold for a normal hip varies between studies. Graf reported an $\alpha$ of $>60^\circ$ to represent a healthy hip [52], whereas Jones reported a healthy hip as being represented by an $\alpha$ of $>55^\circ$ [82]. In contrast to $\alpha$, a lower $\beta$ of $<55^\circ$ corresponds to healthier hips [52]. To improve reproducibility, these measurements are only performed in the B-mode US images collected in the coronal plane, fulfilling the following adequacy criteria: the presence of the labrum, ischium, femoral head, acetabulum and flat horizontal ilium (Figure 1.3 (c)) [52, 79].

**FHC Assessments using Morin’s method:** FHC is defined as the ratio of the acetabular width to the maximal femoral head diameter (Figure 1.3 (f)). A higher FHC corresponds to healthier hips, with a hip with $FHC > 55\%$ being considered healthy. Also, to improve reproducibility, FHC measurements are made only in
the B-mode US images collected in the coronal plane which fulfill the following adequacy criteria: the presence of the labrum, ischium, femoral head, acetabulum and flat horizontal ilium (Figure 1.3(c)) [79].

The Reducibility Assessment using Harcke’s method: In this method, US imaging is used to observe the relative location of the femoral head with respect to the ilium from both the coronal and transverse planes, while a motion and stress maneuver similar to the Barlow and Ortolani examination being performed [61]. The diagnosis outcome is categorical, namely: normal, lax, dislocatable, reducible and not reducible.

Several other dysplasia metrics are available but not commonly used; these include the acetabular contact angle (ACA a measure of the relative shape of the 3-D acetabulum to the vertical cortex of the ilium) [64, 94], combined H angle (H the angle between the vertical cortex of the ilium and the line that connects the lower medial iliac edge to the most distal part of the labrum) [74], rounding index (M: a measure of acetabular convexity) [23], acetabular radius of curvature (AROC a measure of curvature at the perichondrial point) [23], arc length (AL another measure of curvature at the perichondrial point) [23], pubo-femoral distance (PFD a measure of distance between the pubic bone and the femoral head) [157], bony-rim coverage (BRC a measure of ratio of the acetabular width to the maximal femoral head diameter) [157] and L value (L the slope from the vertical cortex of the ilium to the acetabulum) [131]. A detailed analysis of the reproducibility of each in comparison to Graf’s [51], Morin’s [98] and Harcke’s [62] methods, are covered in Chapter 2.

While several methods exist to diagnose DDH using US imaging, all of these methods share two major difficulties. First, it is challenging to identify a standard or adequate 2D US plane that intersects all the necessary bone/cartilage structures (the ilium, acetabulum, ischium, labrum and femoral head) while remaining reproducible under repeated acquisitions. In one study, a group of 250 medical doctors (orthopaedic surgeons, paediatricians and radiologists) performing hip sonography examinations were required to classify four sonogram images [53]. 72% of the medical doctors made mistakes. 64% of these medical doctors were unable to locate the bone/cartilage boundaries and adequate planes [53]. The remaining 36%
were able to identify adequate US images, but miscalculated the dysplasia metric measurements. Although the authors did not report the degree of error in making dysplasia metric measurements, other studies have reported high variability in repeated dysplasia metric measurements from the same hip \cite{79, 107} (e.g. standard deviation approximately \( \approx 7^\circ \) in repeated \( \alpha \) measurements \cite{79}).

### 1.4 Towards Minimal User-Interaction in Extracting Dysplasia Metrics

Analogous to the US-based methods outlined in the previous section, a computer aided automatic DDH assessment should include methods for segmenting the bone and cartilage boundaries from a US image, one for classifying the adequacy of the acquired image for dysplasia metric extraction, and one for extracting the dysplasia metrics from the adequate images. In this section, we outline the relevant background literature for each of these three components necessary for an automatic DDH assessment pipeline.

#### 1.4.1 US Bone Imaging

Automatic assessment of bone boundary in US images can be difficult since US images are typically characterized by high levels of speckle noise, reverberation, anisotropy, and signal dropout, thereby making it demanding to interpret the image and reliably detect its relevant features \cite{103}. Over the last decade, several studies have proposed novel methods for extracting the bone boundaries from US images.

Some of these studies have suggested using phase symmetry responses in US images to localize bone boundaries that are mostly planar \cite{2, 17, 57, 59}. The phase symmetry response is an intensity invariant local symmetry feature. The intensity invariance property of phase symmetry makes it a robust measure against US signal dropout. To address US signal dropout, an intensity invariant feature like phase symmetry is more advantageous than using a time–gain compensation (a method that increases the received signal intensity with depth) since the latter introduces more noise in the US image.

To improve segmentation on non-flat, multi-oriented or curved boundaries Hacihaliloglu \cite{53} proposed the use of isotropic phase symmetric measurements. While
Figure 1.3: (a) and (b) display the anatomy surrounding a hip joint, with (a) illustrating the femur and (b) the hip bone. Note that the femoral head is not ossified. (c) A B-mode US image of an infant hip acquired from angle of the coronal plane. This is also an example of an adequate US image since this image includes the labrum, ischium, femoral head, acetabulum and flat horizontal ilium [79]. (d), (e) and (f) illustrate example measurements of $\alpha$, $\beta$ and $FHC$. 

$FHC = \frac{d}{D}$
we did not find any direct comparison between Hacihaliloglu’s automatically-selected phase symmetry method [59] and Hacihaliloglu’s isotropic phase symmetry-based method [58], qualitative results in these works suggest that the isotropic phase symmetry-based method has noticeably more false positive bone surfaces as compared to the non isotropic phase symmetry-based method.

Other bone-boundary segmentation methods exist that exploit the bone shadowing effect and local image intensity. For example, Fanti [42] used eigen-analysis information from a multi-scale 3D Hessian matrix to enhance sheet-like surfaces for the purpose of generating 3D segmentations of large bones [42]. However, the results from these techniques remain heavily dependent on the quality of the US images used, as well as on the depth and complexity of the imaged bones due to the effects of the shadowing and attenuation of the local intensities [59]. Recently, Hussain [77] proposed combining elastography strain imaging and the envelope power of radio-frequency values to identify bone boundaries. The authors demonstrated a markedly reduced false positive rates in these techniques as compared to those of phase-symmetry-based methods. However, the acquisition of good quality strain imaging requires artefact-free cine loops of decompression-compression cycles, in turn requiring specialization even in the acquisition of 2D US images [161].

In 2014, we introduced a US bone imaging method that combines symmetry and attenuation features to markedly reduce false positive outliers that are present in a phase-symmetry–based bone segmentation [120]. Following our contribution, similar methods of combining symmetry and attenuation features for segmenting bone boundaries were reported [2, 80, 109]. The dynamic programming based bone segmentation methods proposed by Ozdemir [109] and Jia [80] both suffer from long computation times (around 2 minutes per US slice [109] and around 5 seconds per US slice [80]). Anas’s method is relatively faster (around 0.3 second per US slice [2], which is similar to our implementation’s runtime – around 0.25 second per slice [120]), but the bone segmentation filters used are limited to a finite number of orientations, suggesting that the technique may not be well-suited to segmenting round (i.e., not sparse-oriented) bone/cartilage boundaries of the neonatal hip. We need an approach that can extract multi-oriented boundaries and yet remain robust to speckle, signal dropout and soft-tissue outliers. Note that the
computation times reported in the different studies reported here are comparable since the implementations are all on MATLAB and the computers used are comparable (e.g. an intel i7 930 @ 2.80 GHz and 8 GB RAM for Ozdemir’s slow 2 minutes/slice [109] and a Xeon(R) 3.40 GHz CPU computer with 8 GB RAM for our considerably faster 0.25 second/slice method [120]).

1.4.2 US Adequate Plane Detection

Before extracting dysplasia metrics from the bone and cartilage boundaries in an US scan, it is essential to first check whether an US image is adequate for extracting dysplasia metrics. Studies reveal that the process of classifying adequate images in DDH diagnoses is difficult, even for experts [53, 158].

Similar difficulties of identifying a standard plane or an adequate plane are shared by many other US-based diagnoses [44, 91, 95, 96, 102, 130, 166]. A number of machine learning-based solutions have been proposed to identify adequate or standard US images in diverse settings. Applications include: the fetal face [44], abdomen [130], heart [96], and head [95, 102]; the gestational sac [166] and custom made phantoms [91]. These methods involve supervised machine learning approaches (e.g. random forest [102], AdaBoost [130, 166], support vector [95], probabilistic boosting tree [44], deep learning [164]) and are driven by structural features, which are specifically Haar-like features [44, 102, 130, 166] and dynamic texture models [91, 95] or automatically learned features for deep learning. The main limitations to these methods are that they are generally learned and validated on mostly normal patient data. For our application, we need to identify adequate images in both healthy and hip dysplasia patients; to the best of our knowledge, no methods have been yet proposed to automatically identify adequate US images of the neonatal hip.

1.4.3 Dysplasia Metric Extraction

Once an US image is identified, the next step in a DDH diagnosis procedure is to measure dysplasia metrics in the US image. Although US-based DDH diagnosis has been in use for more than three decades [51, 79], very little work has been performed on automatically extracting dysplasia metrics. Hareendranathan [65]
proposed a semi-automatic method for extracting a contour $\alpha$ angle, based on the relative geometry of the ilium and acetabulum boundaries in a 2D US image. They discovered that this contour $\alpha$ has slightly lower intra-examination variability compared with the standard $\alpha$ angle ($\Delta \sigma = 0.2^\circ$ for one rater and $0.4^\circ$ for another rater). Golan [48] proposed using convolutional neural networks (CNN) to automatically segment the ilium and acetabulum boundary in US images of the neonatal hip, and used the segmented bone boundary to estimate $\alpha$ angle values. The authors did not report the variability of their automatically extracted $\alpha$ values.

The lack of a large reduction in variability is perhaps explained by Jaremko’s study [79] in which Jaremko noted that 2D US-based dysplasia metrics were sensitive to variability in positioning the 2D probe over the 3D hip structure. To address this probe-orientation-dependent variability, recent studies proposed using the 3D US to extract 3D hip morphology-derived dysplasia metrics. For example, Hareendranathan [64] proposed the following 3D dysplasia metric: the ACA (measured from 3D US data using manual landmark selections on the ilium and acetabulum, with subsequent angle measurements). This method involves a slice-by-slice analysis process that requires manually selecting at least three seed points in three of the 2D US slices in a 3D US volume and manually separating the acetabulum from the ilium. Using such an interactive method would require valuable clinician time; further, these manual operations introduce a within-image measurement variability of approximately $1^\circ$ [64] and an inter-scan variability of approximately $4^\circ$ [94]. In another study, Mabee [93] proposed manually selecting an optimal 2D US plane from the 3D US so as to reduce the variability which is currently persistent in an only 2D US-based technique. However, the intra-rater variability in the proposed method did not improve considerably compared to the variability in the 2D US-based diagnosis. In terms of semi-automatic analyses, de Luis-Garcia [51] minimized a local structure tensor feature and an interactive user-defined central location based energy term to segment the 3D femoral head of the neonatal hip. However, this semi-automatic work was validated qualitatively only on one example, without performing any quantitative evaluation of the segmentation performance. Furthermore, no method was proposed for using the segmented femoral head for DDH diagnoses.
1.5 Research Questions Addressed

At least twelve dysplasia metrics have been reported for diagnosing DDH using US imaging (see Chapter 2), but no meta-analysis is available to summarize statistics regarding the variability of any of these dysplasia metrics. This makes it difficult to assess the relative utility of each of these metrics. To provide clear estimates on the variability of these measuring systems, we have formulated our first research question as follows:

• Research Question 1: Using a systematic review and meta-analysis, can we summarize estimates of the reproducibility of different dysplasia metrics?

Among these metrics, Graf’s method is the most commonly used [51, 52, 79, 104] and is recommended for use by the Americal College of Radiology [104]. However, manually extracting dysplasia metrics, including Graf’s, is prone to high variability and error [53, 79, 106, 137]. To address this variability and standardize the US-based DDH diagnosis, Germany has established special commissions which have enforced a checklist for hip sonographers to follow [53]. In 2011, one German commission revoked the licenses of up to 43.7% of the hip sonographers in eight German states due to deficiencies in their protocols [158].

To aid clinicians and sonographers in this challenging task of proposingly acquiring the US images and correctly extracting the dysplasia metrics, we have formulated our next three research questions as follows:

• Research Question 2: Can we develop in 2D US images a method that automatically localizes hip bone boundaries? Can we also extend the 2D US bone imaging method to 3D, ensuring that the mean discrepancy between manually-labelled bone boundaries and automatically-extracted bone boundaries is < 2 mm? This 2 mm is based on empirical observations of the thickness of bone boundaries (ranging between 1.5 mm and 4 mm) in US images.

• Research Question 3: Can we develop a method that automatically classifies 2D US images that are adequate for dysplasia metric measurements, with both true positive and negative rates being higher than 90% when the automatic classifications of the images are compared against the expert-labeled classifications?
Research Question 4: Can we develop a method that automatically extracts dysplasia metrics from adequate US images, with at least a moderate agreement (criteria for moderate correlation coefficient: \( r > 0.36 \)) between the automatically and manually extracted dysplasia metrics performed by experts?

In 2013, Graf [53] suggested that the inconsistent hip sonography techniques within and across the health centres are one of the major causes of variability. Thus, we have anticipated that an automatic dysplasia metric extraction method will remove that inconsistency, and provide a more reproducible diagnosis. Our fifth research question was:

Research Question 5: Does the automatically extracted 2D US-based dysplasia metrics improve the reproducibility significantly of the DDH diagnosis?

Later in 2014, Jaremko [79] discovered that a greater source of variability in dysplasia metric extraction is due to changes in the relative locations of the US transducers and the infant subjects’ hips [86]. We anticipate being able to solve this crucial probe-orientation-dependent variability problem using an intrinsic 3D morphology metric derived directly from 3D US scans. However, interpreting 3D US images adds another layer of complexity to manually extracting dysplasia metrics in comparison to interpreting 2D US images. To provide clinicians with an easy-to-use tool that can potentially provide a more reliable diagnoses compared to the current standard 2D US–based diagnosis, we formulate our next two research questions as:

Research Question 6: Can we develop a method that automatically extracts dysplasia metrics able to characterize 3D hip morphology, with at least a moderate agreement (criteria for moderate correlation coefficient: \( r > 0.36 \)) between 3D hip morphology-derived dysplasia metrics and manually extracted 2D dysplasia metrics performed by experts?

Research Question 7: Do automatically extracted 3D US-based dysplasia metrics significantly improve repeatability in DDH diagnosis?
In this thesis, we have not addressed the problem of identifying adequate 3D US scans. For 3D US, this thesis focuses on investigating the potential reduction in variability that is possible with 3D morphology-derived dysplasia metrics.

1.6 Thesis Overview

In addition to this introductory chapter, the thesis includes five chapters. The final chapter discusses the conclusions and directions for future work.

An overview of Chapters 2 to 5 is shown in the flowchart in Figure 1.4. Figure 1.4 also shows relevant publications associated with each of the research questions.

Chapter 2 addresses the first research question: “Using a systematic review and meta-analysis, can we summarize the estimates of reproducibility of different dysplasia metrics?” Here, we conduct a systematic review and meta-analysis to summarize the variability of each available dysplasia metrics, compare the degree of usefulness of each dysplasia metrics in terms of its variability, and, finally, recommend a dysplasia metric for use in DDH assessments.

In Chapter 3, we address the second research question: “Can we develop in 2D US images a method that automatically localizes hip bone boundaries? Can we also extend the 2D US bone imaging method to 3D, ensuring that the mean discrepancy between manually-labelled bone boundaries and automatically-extracted bone boundaries is $< 2\text{mm}$?” Here, we present our bone imaging methods based on the symmetry and attenuation features in US images. Evaluating on US images of 15 infant hips (age 0 to 4 months), we report a mean discrepancy of $< 1.5\text{mm}$ between manually-labelled bone boundaries and automatically-extracted bone boundaries.

Later, in Chapter 4, we address our third, fourth and fifth research questions concerning reliably identifying adequate 2D US images and calculating dysplasia metrics. Here, we develop a machine learning-based method for identifying whether a 2D US image is adequate for dysplasia metric measurements. We also present methods for automatically extracting the three commonly used dysplasia metrics $[\alpha, \beta, FHC]$. We demonstrate that automatic extractions of these dysplasia metrics reduce variability by around 18% to 21% as compared to that of
Figure 1.4: A flow-chart outlining the research questions and core blocks of this thesis. P1 to P10 are publications associated with this thesis (note that P1 has been submitted and P10 is currently being prepared for submission). The vertical arrow bars to the right provide a visual direction to the components of this thesis described in the various chapters.

Chapter 5 addresses our last two research questions concerning reliably diagnosing DDH using US imaging. Here, we propose new dysplasia metrics that characterize 3D hip morphology in infants. We also present our automatic method for extracting these dysplasia metrics. We demonstrate that these 3D hip morphology-derived dysplasia metrics provide a significant reduction in variability in repeated diagnoses as compared to that of their manual 2D counterparts: an approximately 75% reduction for $\alpha_{3D}$ and an approximately 65% reduction for $FHC_{3D}$.)
Chapter 2

A Systematic Review and Meta-analysis of Variability in Dysplasia Metrics

2.1 Introduction

Despite the availability of at least twelve dysplasia metrics (e.g. Graf’s $\alpha$ angle \cite{51}, femoral head coverage \cite{98}, etc.), classification and diagnostic terminology discrepancies exist in DDH diagnosis largely due to the lack of reliability of dysplasia metrics \cite{79,137}. For example, Jaremko \cite{79} showed that poor repeatability in measuring $\alpha$ angle can result in falsely classifying a dysplastic hip as normal, as well as falsely classifying a normal hip as dysplastic. Our goal in this chapter is to establish the reproducibility of the currently available dysplasia metrics.

Investigating reproducibility is the first step towards identifying the usefulness of any dysplasia metric \cite{137}. Subsequent steps entail quantifying of accuracy (reliability and validity), impact on clinical decisions, risk-benefit analysis and impact on short- and long-term clinical outcomes \cite{137}. Diagnosis and treatment decisions for acetabular dysplasia and hip dislocations depend upon both clinical factors and imaging assessment, but diagnostic accuracy via ultrasound has been hampered by the lack of a definitive standard for what constitutes true DDH. Further variabil-
ity arises between-surgeon and across-centre for treatment thresholds, regardless of radiologic measurements. Consequently, many studies attempting to assess the diagnostic accuracy of ultrasound have used outcomes of later examinations, such as those by plain radiograph at two years of age as a measure of actual outcomes. However, since dysplasia may spontaneously resolve during early infancy development in approximately 90% of hips [147], the quality of sensitivity and specificity analysis is poor. Note that while 90% hip dysplasia resolves spontaneously, failing to diagnose and treat early can lead to significant adverse consequences - e.g. Price [119] calculated that DDH might be responsible for approximately 25,000 hip replacement procedures annually in the United States.

A 2006 systematic review conducted on the quality of diagnostic accuracy in US-based DDH diagnosis concluded that the accuracy in such diagnostic tests is poorly established [137]. Further, there have been studies that have examined the impact of ultrasound diagnosis on clinical decisions and ultimately on clinical outcomes without a complete understanding of the reproducibility or accuracy of the test [12, 98, 154, 155, 167]. Consequently, this systematic review focuses on determining the reproducibility of US-based DDH metrics in order to establish the first step in diagnostic process-flow. There can be substantial variability in the quantitative assessment of dysplasia metrics by ultrasound, arising from two main sources: discrepancies between scans and discrepancies between observers [79]. An additional limitation on ultrasound diagnosis is the lack of an available gold standard on the accuracy of diagnosis (sensitivity and specificity) [148]. In the absence of a radiological gold-standard for diagnosis, this review will limit its study of US reproducibility to those diagnoses that have been clinically validated. In the context of DDH, the validity refers to whether the diagnosis can clinically differentiate between a morphologically healthy hip and a severely dysplastic hip. This distinction alone has been fraught with challenges to accurately ascertain, because clear clinical indications for morphological dysplasia in the absence of hip instability or a dislocation may not be evident until the infant has reached walking age or beyond [148]. Thus, in this chapter, we focus on determining the reproducibility of US-based diagnosis in clinically-validated cases of DDH.

The population of interest in this review is infants between the ages of 0 and 6 months at risk of DDH. The intervention/treatment/test for the population is the di-
agnosis of DDH using ultrasound imaging. Our goal is to summarize the reliability in the available dysplasia metrics. For each of the dysplasia metrics, we aimed to summarize variability and/or agreement statistics for six different levels representing the possible range of clinically relevant diagnostic ratings within and across observers. The first level is the intra-exam, intra-image, intra-observer variability. This situation represents how consistent a single observer is in extracting dysplasia metrics when measuring on a single US image taken from a single US examination. The second level is the intra-exam, intra-image, inter-observer variability, representing the variability in agreement between two readers extracting dysplasia metrics from the same US image selected from a single US exam. Third, the intra-exam, inter-image, intra-observer variability represents the variability in the same reader extracting dysplasia metrics from two different US images captured in the same exam. Fourth, the intra-exam, inter-image, inter-observer variability represents the variability in two different readers extracting dysplasia metrics from two different US images captured in the same US exam. Fifth, the inter-exam, intra-observer variability represents the consistency with which a reader extracts dysplasia metrics on the same hip across two separate US examination sessions. Finally, the inter-exam, inter-observer variability represents the variability between two different readers extracting dysplasia metrics on the same hip based on two different US examination sessions. Each of these represent sources of potential variation, and thus, diagnostic discrepancy in DDH. When an US assessment is done at point-of-care, meaning the US exam is performed directly in the clinic either by an US technician or orthopaedic surgeon during the course of a clinical examination, the inter-exam, inter-rater variability is perhaps best reflective of the true variability.

2.2 Study Identification

Two investigators (Niamul Quader and Emily Schaeffer) independently searched two databases - Medline (1946 to September 1st 2016) and Embase (1974 to September 1st 2016) using Aromataris’ guidelines [3]. Any discrepancy between the two investigators was resolved in consensus meetings. We further consulted two librarians at the University of British Columbia to reduce any potential failure in
Table 2.1: Initial logic grid aligned with the population-intervention-comparison-outcome (PICO) elements of the review question: For infants at risk of DDH, are US imaging-based diagnoses reproducible? Note that although the comparison intervention is left empty in here, we will be comparing all the US imaging-based diagnoses with each other.

<table>
<thead>
<tr>
<th>Population</th>
<th>Intervention</th>
<th>Comparison</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infants at risk of DDH</td>
<td>Diagnosis using ultrasound imaging</td>
<td></td>
<td>Reproducibility</td>
</tr>
</tbody>
</table>

including the eligible studies.

2.2.1 Search Strategy

Our search strategy was developed to answer the following research question: for infants at risk of DDH, are US imaging-based diagnoses reproducible? We first derived a logic grid from our research question and then updated the identified concepts in Table 2.1 using synonyms and alternative words (Table 2.2). Subsequently, we updated Table 2.2 using index terms or medical subject headings (MeSH) and keywords with wildcard characters (Table 2.3). The final search strategy using keywords and MeSH derived from our research question is presented in Table 2.4.

Studies were eligible for inclusion if they fulfilled one of the following three conditions: (1) proposed a new US-based diagnosis of DDH, (2) featured a US imaging-based modification to an older method of DDH diagnosis, or (3) investigated the reproducibility of any US-based DDH diagnosis. Studies were excluded if they were based on less than 10 human patients, not reported in the English language, or deemed to be of poor methodologic quality.

2.2.2 Results

Our search strategy (Figure 2.1) retrieved a total of 427 articles from Embase and 275 articles from Medline. The combined total of 702 articles was manually de-duplicated. After sorting articles by title and removing articles only when an exact title match was found, 497 unique articles remained. Upon abstract review of these 497 remaining articles, 42 were selected for full-text review. In total, 29 articles
Table 2.2: Logic Grid with identified keywords added in each of the PICO columns.

<table>
<thead>
<tr>
<th>Population</th>
<th>Intervention</th>
<th>Comparison</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developmental dysplasia of the hip (kw) exp Hip Dislocation, Congenital/ (mh)</td>
<td>Ultrasound</td>
<td></td>
<td>Reliability</td>
</tr>
<tr>
<td></td>
<td>Ultrasonography</td>
<td></td>
<td>Repeatability</td>
</tr>
<tr>
<td></td>
<td>Measure</td>
<td>Congenital hip dislocation</td>
<td>Duplicability</td>
</tr>
<tr>
<td></td>
<td>Screening</td>
<td></td>
<td>Replicability</td>
</tr>
<tr>
<td></td>
<td>Quantitative</td>
<td></td>
<td>Variability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Interrater</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intrarater</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reproducibility</td>
</tr>
</tbody>
</table>

Table 2.3: Logic Grid with keywords and index terms or MeSH headings in each of the PICO columns [mh = mesh headings, kw = keyword, exp=explode].

<table>
<thead>
<tr>
<th>Population</th>
<th>Intervention</th>
<th>Comparison</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developmental dysplasia of the hip (kw) exp Hip Dislocation, Congenital/ (mh)</td>
<td>ultraso* (kw)</td>
<td></td>
<td>reliab*(kw)</td>
</tr>
<tr>
<td></td>
<td>exp Ultrasonography/ (mh)</td>
<td></td>
<td>repeat* (kw)</td>
</tr>
<tr>
<td></td>
<td>measur* (kw)</td>
<td></td>
<td>duplica* (kw)</td>
</tr>
<tr>
<td></td>
<td>metric* (kw)</td>
<td></td>
<td>replic* (kw)</td>
</tr>
<tr>
<td></td>
<td>screen* (kw)</td>
<td></td>
<td>varia* (kw)</td>
</tr>
<tr>
<td></td>
<td>quantit* (kw)</td>
<td></td>
<td>inter* (kw)</td>
</tr>
<tr>
<td></td>
<td>exp Neonatal screening/ (mh)</td>
<td></td>
<td>reproduc* (kw)</td>
</tr>
<tr>
<td></td>
<td>exp “Reproducibility of Results”/ (mh)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>exp observer variation/ (mh)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4: Final search strategy using keywords and MeSH [mh = mesh headings, kw = keyword, exp=explode].

<table>
<thead>
<tr>
<th>Search strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Developmental dysplasia of the hip.mp. or exp Hip Dislocation, Congenital/ [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]</td>
</tr>
<tr>
<td>2. ultraso*.mp. or exp Ultrasonography/ [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]</td>
</tr>
<tr>
<td>3. 1 and 2</td>
</tr>
<tr>
<td>4. limit 4 to humans</td>
</tr>
<tr>
<td>5. (measur* or metric* or screen* or quantit*).mp. or exp Neonatal screening/ [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]</td>
</tr>
<tr>
<td>6. (reliab* or varia* or inter* or reproduc*).mp. or (exp “Reproducibility of Results”/ or exp observer variation/ or exp “Sensitivity and Specificity”/) [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]</td>
</tr>
<tr>
<td>7. 4 and 5 and 6</td>
</tr>
</tbody>
</table>

[9, 23–25, 27, 35, 41, 64, 65, 73, 74, 79, 81, 82, 93, 94, 107, 108, 113, 114, 131, 133, 135, 138, 141, 149, 153, 157, 167] were selected for quality appraisal and data extraction following the full-text review. The reasons for excluding a study were based on the inclusion and exclusion criteria.

2.2.3 Discussion

Our systematic review was limited to articles that were published only in the English language, potentially biasing results by the omission of non-English articles with relevant variability data [83]. Inherent to all systematic reviews, there is also
risk of publication bias toward favorable results; consequently, we may be underestimating the magnitude of variability that exists in the extraction of dysplasia metrics [38].

### 2.3 Methodologic Quality Appraisal

For each eligible study, we assessed the quality of methods by using the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [87], which is a 15-item checklist providing a standard or guideline for reporting reliability or agreement in healthcare studies. For each included study, the two reviewers independently assessed for the presence or absence of each checklist item. Disagree-
ments in quality appraisal were resolved through consensus discussion. We judged a study to be poor if it had a GRRAS score of less than 8.

2.3.1 Results
The average GRRAS score of the selected articles was 10.7 (SD 1.09), with a high score of 12 achieved in five studies [41, 65, 79, 114, 139] and a low score of 8 [27, 131].

Quality of statistics: The quality of the statistics in many of the included studies was poor primarily because the statistical methods used in extracting statistics were not explicitly stated. Specifically, seven out of the ten studies that reported intraclass correlation coefficient (ICC) statistics did not provide any specific details on which model of ICC was used. In thirteen studies that reported Kappa statistics, only six studies reported both the confidence intervals of Kappa coefficient as well as the computation method of computing Kappa, two studies reported only the confidence interval but did not report the method of computing Kappa and three studies reported only the method of computing Kappa. The remaining two studies neither provided the confidence intervals nor the method of computing Kappa.

Risk of bias in individual studies: We found 20 out of 28 assessed studies that explicitly stated that measurements/ratings were conducted independently. In the remaining 8 studies, it was not clear whether there was any bias in between the repeated measurements.

2.3.2 Discussion
Overall, the quality of the included 28 studies was moderate when assessed using the GRRAS guideline, a 15 item checklist for the accurate reporting of studies of reliability and agreement (average 10.7 out of 15, range 6-12). However, we found a considerable lack of clearly reported statistics in many of the included studies. This limits the outcomes of our systematic review and meta-analysis, since they depend on the quality of study and quality of reported statistics within each of the included studies.
The first limitation relating to quality of studies was that we could not perform a meta-analysis on Cohen’s Kappa since most of the included studies in our systematic review did not mention the confidence intervals of their Kappa statistics, an essential component of Kappa statistic meta-analyses [151]. Second, we did not perform a meta-analysis on ICC measures reported within the included studies since most of those studies did not specify their ICC-calculating method. Given that there are multiple methods for ICC calculation, comparability across all studies could not be suitably established. Therefore, our meta-analysis was limited mostly to the variability measures reported within the included studies, although we have provided scatter plots summarizing the Kappa and ICC statistics found within each of the individual studies.

2.4 Meta-analysis of variability in dysplasia metrics

2.4.1 Extracting Variability and Agreement Statistics

The primary investigator (N.Q.) extracted basic demographic data and variability or agreement statistics for each selected study. Demographic data included: (1) year of publication, (2) last name of principal author, (3) country of origin for the study, (4) reported dysplasia metrics, (5) frequency setting used in ultrasound transducer, (6) statistics used in describing variability or agreement, (7) sample size, and (8) profession/expertise of clinicians. Variability or agreement statistics included: (1) intra-exam, intra-image, intra-observer, (2) intra-exam, intra-image, inter-observer, (3) intra-exam, inter-image, intra-observer, (4) intra-exam, inter-image, inter-observer, (5) inter-exam, intra-observer, and (6) inter-exam, inter-observer in estimating dysplasia metrics.

2.4.2 Summarizing Variability Statistics

To capture the variability in dysplasia metrics between repeated measurements, we categorized variability measures (i.e., standard deviation values) of each dysplasia metric separately. Specifically, we combined these variability measures by extracting the pooled standard deviation, $\sigma_p$ [14]. Next we estimated the heterogeneity statistic (i.e., between-study standard deviation, $\sigma_b$) using a Markov chain Monte
Carlo-based random effects model [115]. We then incorporated the heterogeneity statistic to estimate the total standard deviation, \( \sigma_T = \sqrt{\sigma_p^2 + \sigma_b^2} \). Finally, in order to directly compare the variability of different dysplasia metrics, we standardized each by a range, \( r = t_n - t_a \), where \( t_n \) is the threshold for normal hips and \( t_a \) is the threshold for abnormal hips for the given metric.

2.4.3 Summarizing Agreement Statistics

To capture the agreement in dysplasia metrics between repeated measurements within individual studies, we categorized agreement measures (i.e., ICC and Kappa statistic) of each dysplasia metric separately. Note that we could not impute variability measures (e.g., standard deviation values) from ICC [71] since the included studies did not report standard deviation values within each of the measurement groups separately.

2.4.4 Comparing Dysplasia Metrics

To identify the most reliable metric, we identified the dysplasia metric that had the smallest average \( \sigma_T \). Next, we tested whether that dysplasia metric had a significantly lower variability compared to the other dysplasia metrics, with a Bonferroni correction made to adjust for multiple comparisons [13].

2.4.5 Estimating Trend of Variability over Time

To investigate whether the dysplasia metric reliability trend was improving or deteriorating with time, we estimated the variability correlation coefficients of each individual dysplasia metric against year of publication for each intra-/inter-rater condition. We then separately combined the correlation coefficients using MedCalc software that first calculates the weighted summary Correlation coefficient [69] and then incorporates the heterogeneity statistic using a random effects model [32].
Table 2.5: Criteria for evaluating normal and dysplastic hips along with ranges between normal and hip dysplasia for each of the dysplasia metrics.

<table>
<thead>
<tr>
<th>Criteria, normal hip</th>
<th>Criteria, hip dysplasia</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>α &gt; 60°</td>
<td>α &lt; 43°</td>
<td>17°</td>
</tr>
<tr>
<td>β &lt; 55°</td>
<td>β &gt; 75°</td>
<td>20°</td>
</tr>
<tr>
<td>FHC &gt; 55%</td>
<td>FHC &lt; 40%</td>
<td>15%</td>
</tr>
<tr>
<td>ACA &gt; 48°</td>
<td>ACA &lt; 38°</td>
<td>10°</td>
</tr>
<tr>
<td>H &lt; 75°</td>
<td>H &gt; 85°</td>
<td>10°</td>
</tr>
<tr>
<td>PFD &gt; 5.6mm</td>
<td>PFD &lt; 6.6mm</td>
<td>1mm</td>
</tr>
<tr>
<td>M &gt; 0.18</td>
<td>M &lt; 0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>AROC &lt; 2mm</td>
<td>AROC &gt; 2.4mm</td>
<td>0.4mm</td>
</tr>
<tr>
<td>AL &lt; 2.05mm</td>
<td>AL &gt; 2.1mm</td>
<td>0.05mm</td>
</tr>
<tr>
<td>BRC &lt; 51%</td>
<td>BRC &lt; 69%</td>
<td>18%</td>
</tr>
<tr>
<td>L &lt; 0.68</td>
<td>L &gt; 0.92</td>
<td>0.24</td>
</tr>
</tbody>
</table>

2.4.6 Results

Available dysplasia metrics: The reproducibility (either agreement or variability) of the following dysplasia metrics were investigated by at least one study: (a) α angle (24 studies), (b) β angle (16 studies), (c) FHC measure (7 studies), (d) ACA angle (2 studies), (e) H angle (2 studies), (f) M measure (1 study), (g) AROC measure (1 study), (h) AL measure (1 study), (i) PFD measure (2 studies), (j) BRC measure (1 study), and (k) L measure (1 study). Variability of α angle, β angle, FHC, ACA and H angle were reported in more than one study and were included in our meta-analysis of variability measures. The resulting standard deviation measure ($\sigma_T$) for a dysplasia metric obtained from our meta-analysis were standardized by dividing $\sigma_T$ by the range of that dysplasia metric between normal and dysplastic hips. These ranges are based on the criteria for normal hips and hip dysplasia as summarized in Table 2.5 [23, 64, 82, 94, 157]. We did not find any study that investigated reproducibility of Harcke’s dynamic assessment of DDH [62].

Variability in dysplasia metrics: Variability, as denoted by standard deviation values, reported in each of the studies for the reported dysplasia metrics are shown in
α angle was investigated in the largest number of studies (12/28), followed by β angle (7/28) and femoral head coverage (5/28).

The pooled variability (σ_p) and the total variability (σ_T), for each dysplasia metric are shown in Table 2.6 and Table 2.7. Since σ_p was adjusted to σ_T based on between-study heterogeneity, the difference between σ_p and σ_T provides a measure of the heterogeneity between different studies. The standardized variability for each metric (i.e., σ_S = σ_T / r * 100%) is shown in Figure 2.9. For intra-exam, intra-image, intra-rater variability, ACA had the lowest mean variability (σ_S = 23%), followed by β (σ_S = 24%). There was no statistical difference between variability of ACA and β (p > 0.05), however both ACA and β seemed to have significantly lower variability compared to α, FHC and combined H angle (p < 0.05/4). For intra-exam, intra-image, inter-rater variability, β had the lowest variability (σ_S = 28%, p < 0.05/4), followed by ACA (σ_S = 32%). FHC is the only dysplasia metric whose intra-exam, inter-image, inter-rater variability was reported in at least two studies. For both inter-exam, intra-rater and inter-rater variability, α seemed to have the lowest variability (σ_S = 26% and σ_S = 41%, p < 0.05/4).

Agreement in dysplasia metrics: Reported ICC coefficients for each dysplasia metric are shown in Figure 2.10, Figure 2.11, Figure 2.12, Figure 2.16, Figure 2.15, Figure 2.14 and Figure 2.13. α, β and FHC were the only metrics investigated for inter-exam, inter-observer agreement. Here, inter-exam, inter-rater ICC for α varied from 0.03 to 0.45 and β varied from 0.13 to 0.45. Inter-exam, inter-observer agreement for FHC was reported in only one study (ICC=0.02).

Reported Kappa coefficients for each dysplasia metric are shown in Figure 2.17. α and β angle-based Graf classification has been investigated the most (10 studies), whereas the Kappa statistic of the other dysplasia metrics were reported in only one study each. For inter-exam, inter-observer agreement, α and β angle-based Graf classification ranged from poor-to-moderate (Kappa 0.1 to 0.6), position of cartilaginous roof-based classification was fair (Kappa 0.2 to 0.4) and shape of bony roof-based classification ranged from poor-to-moderate (Kappa 0.1 to 0.4).
Figure 2.2: Variability in $\alpha$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.3: Variability in $\beta$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.
Figure 2.4: Variability in *FHC*. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.5: Variability in *ACA* angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.
<table>
<thead>
<tr>
<th></th>
<th>Intra-examination</th>
<th>Inter-examination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>intra-rater</td>
<td>inter-rater</td>
</tr>
<tr>
<td>α in °</td>
<td>σ ( \rho )</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td>σ ( \tau )</td>
<td>4.33</td>
</tr>
<tr>
<td></td>
<td><strong>Upper limit of</strong></td>
<td>5.74</td>
</tr>
<tr>
<td></td>
<td><strong>Lower limit of</strong></td>
<td>3.48</td>
</tr>
<tr>
<td>β in °</td>
<td>σ ( \rho )</td>
<td>4.99</td>
</tr>
<tr>
<td></td>
<td>σ ( \tau )</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td><strong>Lower limit of</strong></td>
<td>5</td>
</tr>
<tr>
<td>FHC in %</td>
<td>σ ( \rho )</td>
<td>4.45</td>
</tr>
<tr>
<td></td>
<td>σ ( \tau )</td>
<td>5.26</td>
</tr>
<tr>
<td></td>
<td><strong>Upper limit of</strong></td>
<td>9.32</td>
</tr>
<tr>
<td></td>
<td><strong>Lower limit of</strong></td>
<td>4.46</td>
</tr>
</tbody>
</table>
Figure 2.6: Variability in PFD. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.7: Variability in $M$ measure. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Trend of Variability over Time: The meta-analysis results for the trend of variability in each of the individual dysplasia metrics against year of publication (i.e., correlation of variability with year of study) are shown by forest plots in Figure 2.18, Figure 2.19, Figure 2.20. With limited available data, we were only able to find the trend in variability against year of publication for $\alpha$, $\beta$, FHC and combined H angle. $\beta$ is the only dysplasia metric that shows a slight trend towards improving repeatability. All the other three dysplasia metrics show moderate to strong trends.
Figure 2.8: Variability in $H$ angle. Variability is expressed as standard deviation of discrepancy between repeated measurements.

Figure 2.9: Variability in dysplasia metrics based on our meta analysis. Here, the variability of each dysplasia metric was standardized by dividing $\sigma_T$ by the range of that dysplasia metric between normal and dysplastic hips. The inter-exam, inter-rater variability is the most clinically relevant variability.
**Figure 2.10:** Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $\alpha$.

**Figure 2.11:** Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $\beta$. 

---

36
Figure 2.12: Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $FHC$.

Figure 2.13: Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $AROC$. 
Table 2.7: Meta-analysis result for the total variability ($\sigma_T$) in ACA angle (in °) and H angle (in °) and FHC (in %).

<table>
<thead>
<tr>
<th></th>
<th>Intra-examination</th>
<th>Inter-examination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>intra-rater</td>
<td>inter-rater</td>
</tr>
<tr>
<td></td>
<td>intra-rater</td>
<td>inter-rater</td>
</tr>
<tr>
<td>Intra-image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-image</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACA in °</th>
<th>$\sigma_p$</th>
<th>$\sigma_T$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.66</td>
<td>3.89</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>2.54</td>
<td>5.42</td>
<td>6.96</td>
</tr>
<tr>
<td>Upper limit of $\sigma_T$</td>
<td>9.39</td>
<td>14.44</td>
<td>15.79</td>
</tr>
<tr>
<td>Lower limit of $\sigma_T$</td>
<td>2.66</td>
<td>2.56</td>
<td>4.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H in °</th>
<th>$\sigma_p$</th>
<th>$\sigma_T$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.18</td>
<td>5.94</td>
<td>4.06</td>
</tr>
<tr>
<td></td>
<td>4.06</td>
<td>6.76</td>
<td></td>
</tr>
<tr>
<td>Upper limit of $\sigma_T$</td>
<td>14.93</td>
<td>15.59</td>
<td></td>
</tr>
<tr>
<td>Lower limit of $\sigma_T$</td>
<td>3.2</td>
<td>4.08</td>
<td></td>
</tr>
</tbody>
</table>

towards deteriorating repeatability in repeated measurements.

2.4.7 Discussion

An US-based examination is fraught with many opportunities for operator or measurement errors to reduce its reliability. An US examination is a dynamic process involving collecting a video file that can be broken down into a series of static 2D images from which a single, optimal representative image is chosen for extraction of dysplasia metrics. Consequently, variability in metric extraction has the poten-
Figure 2.14: Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for AL.

Figure 2.15: Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for PFD.

tial to arise within the same US exam, based on both or either the image chosen, and the observer(s) performing the measurements. Additional variability is introduced by both the same observer and different observers performing measurements on images taken from different US exam sessions of the same hip.

In our meta-analysis, we were able to assess the variability of multiple dysplasia metrics commonly used for 2D US $\alpha$ and $\beta$ angles and percent femoral head coverage (FHC). Additionally, we were also able to assess the variability of a three-dimensional (3D) US dysplasia metric the ACA employed in a limited number
Agreement in dysplasia metrics, expressed as ICC coefficient between repeated measurements for $M$. The meta-analysis of variability measures reported within the studies suggested that ACA has the lowest intra-exam intra-image intra-rater variability (average 22%); however, it should be noted that only two of the 28 included studies provided data for this metric. Additionally, as a measure specific to 3D US, ACA currently has limited clinical applicability due to the lack of widespread availability of 3D US in point-of-care clinical practice. Of the more widely-used 2D US measures, the $\beta$ angle has the lowest intra-exam intra-image inter-rater variability (average 29%) while the $\alpha$ angle has both the lowest inter-exam intra-rater variability (average 25%) and the lowest inter-exam inter-rater variability (average 41%). We could not compare intra-exam inter-image variability of dysplasia metrics, since intra-exam inter-image variability of only FHC was reported in a few studies (average of 38%).

Depending upon the process-flow of the clinic setting, it is debatable which situation is the most representative of the true variability of dysplasia metrics in a real clinical setting. In some hospitals, US assessment is done at point-of-care, meaning the US exam is performed directly in the clinic either by an US technician or orthopaedic surgeon during the course of a clinical examination. In these point-of-care settings, the inter-exam, inter-rater variability is perhaps best reflective of the true variability. We found $\alpha$ (average variability of 41%), $\beta$ (average variability of 45%) and FHC (average variability of 63%) to be the only dysplasia metrics
Figure 2.17: Agreement in dysplasia metrics, expressed as Kappa coefficient between repeated measurements.
Figure 2.18: Forest plot showing correlation coefficients between variability of \( \alpha \) and year of study. Top rows show mean and 95% confidence intervals of the correlation between intra-exam, intra-image, intra-user variability and year of study for \( \alpha \). Last row show the total random effects correlation coefficient estimating the effective change in variability with year of study for \( \alpha \). Here, the summary measure is the center of diamond, and the associated confidence intervals of the summary measure are the lateral tips of the diamond.

whose inter-exam inter-rater variability has been studied. Graf’s \( \alpha \) and \( \beta \) metrics are significantly less variable compared to FHC; however, despite the lower variability in Graf’s metrics compared to other available dysplasia metrics, their near 40% variability is problematic, and is also evident in the agreement measures reported in the included studies - both \( \alpha \) and \( \beta \) showed poor-to-moderate agreement between repeated inter-exam inter-rater measurements (\( \alpha \) - range of ICC 0.03 to 0.45, \( \beta \) - range of ICC 0.13 to 0.45). With this high degree of variability in the most commonly used dysplasia metrics, it is perhaps unsurprising that a consensus clinical gold standard for DDH diagnosis has yet to be conclusively defined.
Figure 2.19: Forest plot showing correlation coefficients between variability of $\beta$ and year of study. Top rows show mean and 95% confidence intervals of the correlation between intra-exam, intra-image, intra-user variability and year of study for $\beta$. Last row show the total random effects correlation coefficient estimating the effective change in variability with year of study for $\beta$.

While $\alpha$, $\beta$ and FHC measures were proposed in 1983 [51] and 1985 [98], and despite representation from included studies across 30 years, we did not find any evidence for a trend of reduction in variability in those metrics over the years. On the contrary, the repeatability of both $\alpha$ and FHC seems to have deteriorated, while $\beta$ showed only slight improvement over this time period. This lack of improvement suggests that, even in spite of technological advances in US equipment over time, dysplasia metric measurement remains primarily subject to individual operator/observer techniques. Additionally, we found considerable heterogeneity between measurements in different studies - the between-study standard deviation...
Figure 2.20: Forest plot showing correlation coefficients between variability of FHC and year of study. Top rows show mean and 95% confidence intervals of the correlation between intra-exam, intra-image, intra-user variability and year of study for FHC. Last row show the total random effects correlation coefficient estimating the effective change in variability with year of study for FHC.

\[14, 115\] was on average 1.4 times higher than the pooled within-study standard deviation. This perhaps suggests the need for more standardized tools and process-of-care across all centers. Consequently, development of automated metric extraction processes may be the optimal way to minimize intra-observer, inter-observer and cross-centre variability.

There were several limitations to our systematic review and meta-analysis, mainly arising due to a dependence on the statistics provided within the included studies. First, we did not perform a meta-analysis on Cohen’s Kappa since most of the included studies in our systematic review did not mention the confidence intervals of their Kappa statistics, an essential component of Kappa statistic meta-
analyses [151]. Second, we also did not perform a meta-analysis on ICC measures reported within the included studies since most of those studies did not specify their ICC-calculating method. Given that there are multiple methods for ICC calculation, comparability across all studies could not be suitably established. Therefore, our meta-analysis was limited mostly to the variability measures reported within the included studies, although we have provided scatter plots summarizing Kappa and ICC statistics found within each of the individual studies. Another limitation was that, we only used Bonferroni correction in our multiple-comparison correction, which tends to be more conservative than other approaches [101]. Thus, in the only instance where we did not find a statistically significant result (comparing intra-exam, intra-image, intra-rater variability of ACA with $\beta$), there may be a possibility of a statistically significant result if we used a less conservative multiple-comparison correction. Finally, our systematic review was limited to articles that were published only in the English language, potentially biasing results by the omission of non-English articles with relevant variability data [83]. Inherent to all systematic reviews, there is also risk of publication bias toward favorable results; consequently, we may be underestimating the magnitude of variability that exists in the extraction of dysplasia metrics [38].

2.5 Summary

In this chapter, we performed a systematic review and meta-analysis of the variability of dysplasia metrics between repeated measurements for the assessment of the infant hip. We found the most commonly used measure the $\alpha$ angle to be the least variable among all dysplasia metrics; however, we found generally high variability and low agreement in all dysplasia metrics, including the $\alpha$ angle. Despite the lower variability in Graf’s $\alpha$ angle compared to other available dysplasia metrics, their near 40% inter-exam, inter-rater variability is a severe limitation to reliable DDH diagnosis. Furthermore, in the last three decades, the repeatability of dysplasia metrics has not markedly improved, indicating a genuine need for improving repeatability and reliability of US-based DDH diagnosis.
Chapter 3

Symmetry and Attenuation-based Ultrasound Bone Imaging

3.1 Introduction

Before we present our contributions towards improving reliability of diagnosing DDH, we detail our second major contribution - improvements in segmenting bone boundaries from 2D and 3D US images. This is particularly important since all of the current dysplasia metrics involve identifying or segmenting bone boundaries in US images of an infant’s hip, and then making geometric measurements on the identified bone boundaries [51, 52, 94, 155]. Furthermore, US bone imaging has important emerging uses in computer assisted orthopaedic surgery that can potentially reduce surgery time and improve outcomes from surgery [2, 59].

US bone imaging is challenging primarily because US images are typically characterized by high levels of speckle noise, reverberation, anisotropy, and signal dropout, thereby making it difficult to interpret the image and to reliably detect relevant features [103]. Over the last decade, several methods have been proposed to automatically extract bone boundaries from US image.

Bone boundaries were shown to be well captured by phase symmetry (PS) responses in US images since the beam reflection is considerably weaker both before and beyond these structures where the primary reflection occurs [2, 57, 59]. However, local symmetry features remain prone to false detection of soft-tissue
interfaces that often exhibit features similar to those of bone. Furthermore, though quite effective on relatively flat (i.e. sparsely-oriented) structures, PS responses require tedious non-intuitive parameter tuning procedures to correctly identify complex bone shapes. Attempts to automate the parameter selection process have been made [1, 2, 17, 59]. Further, to improve segmentation on non-flat, multi-oriented or curved boundaries Hacihaliloglu [58] proposed the use of isotropic phase symmetric measurements. While we did not find any direct comparison between Hacihaliloglu’s automatically-selected PS method [59] and Hacihaliloglu’s isotropic PS-based method [58], qualitative results in these works suggest that the isotropic PS-based method has considerably more false positive bone surfaces as compared to the non-isotropic PS-based method.

Apart from approaches based on local symmetric features, other bone-boundary segmentation methods exist that exploit the bone shadowing effect and local image intensity. Foroughi [46] employed dynamic programming on intensity and local gradient information to segment bone contours in 2D images. This approach has shown adequate clinical accuracy in 2D US images, but requires region-of-interest selection to remove soft-tissue interfaces near the skin surface, such that the method could only be applied to 2D US images in a semi-automated manner. A separate bone contour detection scheme depends on depth-weighted adaptive thresholding and subsequent morphological opening/closing operators to enhance segmented bone surfaces in 2D images [89]. Another study used eigen-analysis information from a multi-scale 3D Hessian matrix to enhance sheet-like surfaces for the purpose of generating 3D segmentations of large bones [42]. However, the results from these techniques remain heavily dependent on the quality of the US images used, as well as on the depth and complexity of the imaged bones due to the effects of the shadowing and attenuation of the local intensities [57, 59].

Among the strain imaging based bone segmentation techniques [76, 77, 134], the state-of-the-art [77] method combines Ultrasound (US) strain imaging and the envelope power detection of radio-frequency values in a US image to identify bone boundaries with considerably reduced false positive rates as compared to phase-symmetry-based methods. However, acquiring good quality strain images requires more skill and training than acquiring B-mode US images [161].

A recent work trained a CNN for use in segmenting bone boundaries [143]. The
authors demonstrated a significantly improved dice coefficient by switching from a random forests-based segmentation to a CNN-based segmentation (increase from 0.79 to 0.87) on a dataset of 1382 US images collected from the femur, tibia and pelvis area from multiple volunteers. While results are promising, it is not clear how many volunteers/patients were in this study, and whether the trained CNN model captures the variability in appearance of bone boundaries in US images in different types of human subjects. Another recent CNN-based implementation [66] involving transfer learning on a pre-trained CNN model [90], was evaluated on 50 US scans (2D) of infant hips. Discrepancy between their automatically segmented bone boundaries and manually segmented bone boundaries were moderate: root mean square discrepancy of 1.8 mm (SD 0.36 mm), Hausdorff distance of 2.1 mm (SD 0.45 mm).

To segment bone boundaries using local symmetry feature while remaining robust to soft-tissue outliers, we proposed to combine attenuation-related features with local symmetry feature [120]; our rationale - bone attenuates ultrasound signal considerably more than soft-tissue. Later, several other methods that combine local symmetry with attenuation-features were proposed [2, 5, 80, 109]. Jia [80] proposed combining PS with intensity gradient-derived features to identify high attenuation regions, and using that information to reduce soft-tissue outliers. Ozdemir [109] proposed quantifying the probability of a pixel in a US image as belonging to any of the following three classes (a) bone, (b) tissue, and (c) shadow. They used a binary classifier on a number of PS and attenuation-related features. Unfortunately, the authors failed to provide the details of the classifiers they utilized. Based on the probability map, Ozdemir used Markov Random Fields to label each pixel to either of the above three classes. The primary limitations for both of these methods is poor runtime (approximately 120 seconds per US image using Ozdemir’s method [109], approximately 9 seconds per US image using Hacihaliloglu’s method [56] and approximately 4 seconds per US image using Jia’s method [80]). Anas’s method is relatively faster (around 0.3 second per US slice [2], which is comparable with our 0.26 second per slice [120]). Note that the computation times reported in the different studies reported here are comparable since the implementations are all on MATLAB and the computers used are comparable (e.g. an intel i7 930 @ 2.80 GHz and 8 GB RAM for Ozdemir’s slow
2 minutes/slice \cite{109} and a Xeon(R) 3.40 GHz CPU computer with 8 GB RAM for our noticeably faster 0.25 second/slice method \cite{120}. Another potential limitation is that, the bone segmentation filters used are limited to a finite number of orientations, suggesting that it may not be well-suited in segmenting round (i.e., not sparse-oriented) bone/cartilage boundaries of the neonatal hip. Furthermore, these methods use anisotropic filters for computing symmetry features, which may not work well in segmenting the rounder bone/cartilage boundaries of the neonatal hip. We need an approach that can extract multi-oriented boundaries and yet remain robust to speckle, signal dropout and soft-tissue outliers.

In this chapter, we present two separate methods for bone boundary extraction. The first method, confidence-weighted phase symmetry (CPS), is an automatic approach that combines multiple US image features including sparsely-oriented local phase information with attenuation-based features to robustly segment bone surfaces in 3D US images in adult hips \cite{3.2,120}. On a dataset of B-mode US volumes acquired from the pelvic region in 18 adult trauma patients, we showed that a combined-feature bone extraction method is more accurate (as measured by discrepancy between surfaces of US-based bone boundaries and CT-derived ground truth bone boundaries) than using PS alone (around 50% improvement, \( p < 0.01 \)).

The second method, structured phase symmetry (SPS), is another automatic approach that extracts local symmetry features by employing isotropic filters, and is independent of the orientation of the bone and cartilage boundaries. We also present an extension to the SPS feature for use in 3D US. On a dataset of 15 US volumes acquired from 15 infant hips, we demonstrated that this technique extracts bone boundaries that are consistent to what an expert would label as bone boundaries, with a mean discrepancy between bone boundaries of \(< 1.5 \text{mm} \). This mean discrepancy is also significantly less than the mean discrepancy between expert-labelled bone boundary and a previous state-of-the-art method-based bone boundary \cite{59} \( p < 0.01 \).

### 3.2 Confidence-Weighted Local Phase Features

Bone and cartilage structures are usually considerably more hyperechoic than the neighboring structures, and are well captured using local symmetry features such
as the intensity-invariant PS feature [57][88]. Therefore, one key feature we used in our bone segmentation was the PS feature proposed by Hacihaliloglu et al. [57]. To calculate those, we used a 3D log-Gabor filter as our quadrature filter to identify the points of symmetry since it can be constructed with arbitrary bandwidth [57]. Another feature of bone material we used was its considerably higher US attenuation effect compared to other tissues, which results in the characteristic shadowing beneath the bone surface (or further away than the bone surface with reference to the US transducer) in the US image. To quantify this shadowing and attenuation feature, we used Karamalis et al.’s [84] shadow detection algorithm, which extracts a transmission model for an US image. Finally, we combine the aforementioned features into a hybrid feature that we call confidence-weighted-phase-symmetry (CPS).

Our hybrid feature, CPS, aims to improve PS-based bone surface segmentation technique presented in [57], which relies on empirically selected parameters. Given a 3D US volume, \( I(x,y,z) \), we first extract hyperechoic structures using PS information \( PS(x,y,z) \) (subsection 3.2.1). Next, to reduce outliers, we incorporate the attenuation feature, \( A(x,y,z) \), and shadowing feature, \( S(x,y,z) \) (defined in subsection 3.2.2). We combine the three measures, \( PS, A \) and \( S \) (subsection 3.2.3) to generate our hybrid feature, \( CPS(x,y,z) \).

### 3.2.1 Local Phase Symmetry Feature

To extract \( PS \), we first filter \( I \) using a bank of \( d \) orientation-weighted log-Gabor filters [57] Each of the orientation-weighted log-Gabor filters in the filter bank is defined by the transfer function:

\[
G(\omega, \phi, \theta) = L(\omega)D(\phi, \theta)
\]

(3.1)

where \( L(\omega) \) controls the frequencies to which the filter responds [hacihaliloglu 2015],

\[
L(\omega) = \exp\left(-\log^2(\omega \times p)/2\log^2(\sigma_p)\right)
\]

(3.2)

and, \( D(\phi, \theta) \) controls the orientation selectivity of the filter [Dosil et al. 2006].
\[ D(\phi, \theta) = \exp\left( -\frac{(\phi - \phi_i)^2}{2\sigma_\phi^2} - \frac{(\theta - \theta_i)^2}{2\sigma_\theta^2} \right) \]  \hspace{1cm} (3.3)

Here, \( \omega \) represents the 3D spatial frequencies along \( \omega_x, \omega_y, \omega_z \), \( ||\omega|| \) is the \( l_2 \)-norm of the spatial frequencies, \( p \) is the center-wavelength of a filter and \( \sigma_\omega \) is the standard deviation of all center-wavelengths across the filter bank, \( \phi_i \) is the azimuth angle, \( \theta_i \) is the elevation angle, \( \sigma_\phi \) determines the angular bandwidth about \( \phi_i \), and \( \sigma_\theta \) determines the angular bandwidth about \( \theta_i \).

For every filter, \( G_d \), the filtered response of \( I \) has a real component (denoted as even symmetric response, \( e_d \)) and an imaginary component (denoted as odd symmetric response \( o_d \)) (Figure 3.1). Similar to Kovesi [88], the PS feature is computed from \( e_d \) and \( o_d \) as:

\[ PS(x, y, z) = \frac{\sum_d (|e_d(x,y,z)| - |o_d(x,y,z)| - T_r)}{\sum_d \sqrt{e_d(x,y,z)^2 + o_d(x,y,z)^2 + \varepsilon}} \]  \hspace{1cm} (3.4)

where \( \varepsilon \) is a small number to prevent division by zero and \( T_r = \mu_n + \sigma_n \) is a noise threshold calculated from mean \( \mu_n \) and standard deviation \( \sigma_n \) of the smallest scale filter which is more likely to have more noise than other higher scale filters [Kovesi 1999].

### 3.2.2 Attenuation-related Features

The calculated PS captures bone/cartilage boundaries and soft-tissue interfaces, all of which exhibit ridge-like characteristics. To distinguish between bone boundaries and other tissue types, we use an attenuation-based post-processing step on every US image slice along the \( z \) direction (perpendicular to the scan planes). Specifically, in a US slice \( I(x, y) \), we make use of the property that the US signal attenuation is very high at bone boundaries. To model this phenomenon in our post-processing framework, we first estimate a relative signal strength map or a confidence map in a B-mode US image similar to that described in [84], \( m(x,y) \), which ranges from 0 to 1, where 0 corresponds to zero signal strength and 1 corresponds to the full source signal strength generated by the transducer. We then estimate the signal strength at each image pixel by calculating the probability of a random walk [49] starting from the pixel and ending at the top of the US image (i.e., location of
Figure 3.1: Flowchart for local phase analysis for a 3D volume collected around the iliac crest area of an adult trauma patient (blue box on segmented CT representing approximate location of US transducer). Here, 3D log-Gabor filtering on the 3D US image results in real and imaginary components, which are used with Equation 3.4 for computing PS response. CPS feature is computed using Equation 3.7. White arrows point to false positive bone surface points.
virtual transducers). This is computed from the graph Laplacian matrix which is defined as:

\[
L_{ij} = \begin{cases} 
  d_i & \text{if } i = j \\
  -w_{ij} & \text{if } i \text{ adjacent to } j \\
  0 & \text{otherwise}
\end{cases}
\]  

where \(d_i = \sum_j w_{ij}\). The edge weights, \(w_{ij}\), are assigned in the horizontal, vertical and diagonal directions:

\[
w_{ij}^H = \exp(-b(|c_i - c_j| + \gamma)), \quad w_{ij}^V = \exp(-b(|c_i - c_j|)), \quad w_{ij}^D = \exp(-b(|c_i - c_j| + \sqrt{2}\gamma).
\]

The term, \(c_i = g(exp(-l))\), is a depth-based intensity gradient, where \(g\) is the image intensity and \(l\) is the normalized closest distance from the pixel’s node to the top of the US image (i.e., the locations of the US transducer). The other parameters are \(\gamma\) and \(b\); \(\gamma\) represents the penalty of a horizontal and diagonal walk compared to a vertical walk, and \(b\) is a regularization term [84]. We discuss our choice of \(\gamma\) and \(b\) in more detail in results section.

Having estimated the relative signal strength \(m(x,y)\), we define two attenuation-based features to localize bone boundaries in \(PS\); our rationale - bone material tends to have a considerably higher US attenuation effect compared to other tissues [84]. The attenuation-based features are: an attenuation feature \(A(x,y) = m(x,y) - (m_{\text{min}})\), and a shadowing feature \(S = m(x,y)/(m_{\text{min}} + \epsilon)\) (both \(A\) and \(S\) are normalized to range between 0 and 1), where \(m_{\text{min}}\) is the minimum node value in a \(p\) by \(p\) window centered around \((x,y)\) and \(\epsilon\) is a small number to prevent division by zero.

### 3.2.3 Combined Feature for Bone Surface Localization

We combine the hyperechoic PS feature [3.2.1] and the attenuation-related features [3.2.2] as an arithmetic average of individual features:

\[
P(x,y) = \frac{[A(x,y) + S(x,y) + PS(x,y)]}{3} \tag{3.6}
\]

We use the bone membership value, \(P\), to enhance the bone structures in \(PS\) and also to remove outliers in \(PS\) (e.g., soft-tissue interfaces), resulting in \(CPS(x,y)\) [Figure 3.1]. Specifically, we locate the coordinates \((x_m, y_m)\) that correspond to the
maximum values of $P$ along each of the scan lines (i.e., the vertical lines in an US image). It is more likely that a pixel belonging to a bone boundary will be near $(x_m, y_m)$ having confidence map values $m(x_m, y_m)$. Leveraging $P$ and $m(x_m, y_m)$, we enhance the bone structures in $PS$ as:

$$CPS(x, y) = \begin{cases} 
P(x, y) \times PS(x, y) & \text{if } |m(x, y) - \mu| < \sigma \text{ and } P(x, y) > 0.5 \\
0 & \text{otherwise}
\end{cases}$$

(3.7)

where $\mu$ is the mean and $\sigma$ is the standard deviation of the confidence map values, $m(x_m, y_m)$. This formulation is based on our hypothesis that bone boundaries generally have confidence map values within $\mu \pm \sigma$, while away from bone boundaries (e.g., in shadows beneath bone surfaces) confidence map values tend to be noticeably different from $\mu$ (i.e., $|m(x, y) - \mu| > \sigma$).

### 3.2.4 Experiment

**Parameter Specification** In our bone pipeline, the log-Gabor filter bank used a value of $p$ ranging from $p_{\text{min}} = 1.5\text{mm}$ to $p_{\text{min}} = 4\text{mm}$ to reflect the typical widths of bone interfaces in US images. We used three sets of elevation and azimuth orientations: $45^\circ$, $90^\circ$ and $135^\circ$ to capture bone boundaries in all orientations. The standard deviation parameter was set as $\sigma_\omega = 1/3 * p_{\text{min}}$ to reduce speckle noise (in our experience, speckle in this type of image is typically under $0.5\text{mm}$). The regularization parameter in the confidence map estimation, $b$, controls the sensitivity of the random walk’s probability or the confidence value to the changes in intensity gradient along the path of the random walk. For the purpose of estimating the relative signal strength in medical ultrasound images, we set this regularization parameter as 100 throughout our experiment. We adopted this value from the original work on confidence maps [84], which suggests that a $20\%$ change around this value has relatively little effect on the confidence map results. The other parameter used in the US confidence map is $\gamma$, which represents the penalty of a horizontal and diagonal walk compared to a vertical walk. In all our experiments, we used $\gamma = 0.05$, based on the qualitative examples provided in [84].
Data  We retrieved US images with corresponding computed tomography (CT) data that were previously collected from an ex-vivo bovine femur phantom and in-vivo pelvic data (around the iliac crest area) from 18 trauma patients (obtained as part of routine clinical care under appropriate institutional review board approval, UBC CREB number: H17–01904) [57]. The ex-vivo bovine femur was placed in a polyvinyl chloride-filled cylindrical tube, with added fiducials to enable precise US-CT comparisons [57]. US images were collected using a GE Voluson 730 Expert ultrasound machine with a 3D RSP5-12 transducer (GE Healthcare, Waukesha, WI, USA) and CT images were collected using a Xtreme CT machine (HRpQCT, XtremeCT, Scanco Medical, Switzerland) [57]. The transducers center ultrasound frequency was kept at 7.5 MHz and image depth setting ranged between 1.9cm-7.2cm.

Validation Scheme  We compared our CPS method against empirical PS-based bone segmentation [57]. First, we extracted bone surfaces separately from both CPS feature and PS feature, determined as the maximum of feature responses along the scan-lines of the US transducer. Next, we extracted gold-standard bone surface from corresponding CT. To quantify the accuracy of a bone segmentation, we aimed to register the US-derived bone surface (either PS or CPS-based) to the gold-standard CT-derived bone surface. Finally, we would calculate a surface fitting error (SFE) defined as the Euclidean root mean square distance between the registered US-derived bone surface and the CT-derived bone surface.

For the in-vivo pelvic data, in the absence of fiducial markers, we used automatic Gaussian mixture model-based registration [16] to align the CT- and US-derived bone surfaces. The final bone surfaces that were used during registration algorithm were determined as the maximum of feature responses along the scan-lines of the US transducer.

Results  Qualitative results in Figure 3.2 show that combining PS response with attenuation-related features reduces false positive soft-tissue interfaces in both ex-vivo bovine femur phantom and in-vivo pelvic data. This marked decrease in false positive bone boundaries is also evident in our quantitative analysis (Figure 3.3).
CPS had a significantly reduced SFE compared to PS (mean SFE of CPS = 0.7mm, mean SFE of PS = 1.2mm, \( p < 0.01 \) using Wilcoxon signed rank test).

In terms of computational cost, for a \( 152 \times 158 \times 112 \) US volume, runtime of CPS was short (run on a Xeon(R) 3.40 GHz CPU computer with 8 GB RAM with MATLAB code) compared to the reported runtimes of other available methods (approximately 0.15 second per US slice for CPS vs. approximately 0.3 second per US slice \([2]\), approximately 3 seconds per US slice for optimized PS \([59]\), approximately 4 seconds per US slice using Jia’s method \([80]\), approximately 120 seconds per US slice using Ozdemir’s method \([109]\)). All processes were executed using MATLAB, the Mathworks Inc., Natick, MA, USA.

3.2.5 Discussion

We showed that by combining local symmetry features with attenuation-based features, bone segmentation accuracy improves significantly compared to segmenting bone boundary using symmetry feature alone \( (p < 0.01) \). We also noted that runtime of our CPS-based was comparable to Anas’s bone segmentation method \([2]\), and was around an order of magnitude faster than Jia’s and Hacihaliloglu’s \([57, 80]\), and around two orders of magnitude faster than Ozdemir’s bone segmentation method \([109]\).

One potential limitation is that some of the free parameters in our method were chosen based on empirical observations (e.g. \( \gamma = 0.05 \) was used in US confidence maps). So, there is opportunity for potentially improving reliability of bone segmentation. While we have not done a sensitivity analysis for each of the parameters, since we have successfully used the same parameter settings in all the experiments in this thesis, we do not think the sensitivity of these parameters is problematic enough to block clinical utility.

Another limitation of our CPS method is that - we used a limited number log-Gabor filters that are oriented in a finite number of orientations (three elevation angles and three azimuth angles in this case). This limited number of filter orientations means that the CPS method presented in this section is likely to work well in US images that have planar bone boundaries, but may suffer from reduced segmentation accuracy in US images that have non-planar bone boundary structures (e.g.
Figure 3.2: Qualitative results: segmented bone surfaces around bovine femur (first column) and around in-vivo pelvis in human trauma patients. First row shows segmented CT with box representing approximate location of US transducer. Second row shows corresponding US volume. Third row shows PS based on empirical parameters [57]. Last row shows the CPS response.
Figure 3.3: Quantitative results: (a) Bovine phantom. Note that our proposed CPS based segmentation resulted in a 0.302 mm reduction in error compared to PS. (b) In-vivo pelvic data across all subjects (C-1 to C-18). Note that our proposed CPS resulted in a reduction in error which is significant at \( p < 0.01 \) based on Wilcoxon signed rank test compared to PS.

small round bone boundary, etc.), which is particularly important in segmenting bone and cartilage boundaries in neonatal hips.

3.3 Structured Phase Symmetry

Bone/cartilage structures in infant hips have complex shapes (both planar and non-planar structures) and are usually considerably more hyperechoic than the neighboring structures; therefore, we introduce a orientation-independent local symmetry feature, which we call the SPS feature, to localize bone/cartilage boundaries in US images of the neonatal hip.

3.3.1 2D Filters-based Structured Phase Symmetry

To extract \( SPS(x, y) \) from a 2D US volume, \( I(x, y) \), we first enhance local PS features using responses to band-pass quadrature filters [88]. Specifically, we filter an US image, \( I(x, y) \), using a bank of \( d \) radial log-Gabor filters. Each of the 2D log-Gabor filters in the filter bank is defined by the transfer function, \( L(\omega) = \exp\left(-\log^2(\omega \times p) / 2\log^2(\sigma_p)\right) \), where \( \omega \) is the 2D spatial frequencies along \( \omega_x, \omega_y \), \( ||\omega|| \) is the \( l_2 \)-norm of the spatial frequencies, \( p \) is the center-wavelength of the filter and \( \sigma_p \) is the standard deviation of all center-wavelengths across the filter bank. To calculate the quadrature components, \( R_d(x, y) \), of the band-passed images, \( B_d(x, y) = I(x, y) \ast L_d(x, y) \), we use the Riesz transform [43], which is defined
by:

\[
R_d(x,y) = \begin{cases} 
R_{d,1}(x,y) \\ R_{d,1}(x,y) 
\end{cases}
= \begin{cases} 
h_1(x,y) * B_d(x,y) \\ h_2(x,y) * B_d(x,y) 
\end{cases}
\] (3.8)

where \(h_1(x,y)\) is the inverse Fourier transform of \(H_1(\omega) = j\omega_x/||\omega||\) and \(h_2(x,y)\) is the inverse Fourier transform of \(H_2(\omega) = j\omega_y/||\omega||\). \(R_d\) is then formulated from its components \(R_{d,1}\) and \(R_{d,2}\) as:

\[
R_d = \sqrt{R_{d,1}^2 + R_{d,2}^2} 
\] (3.9)

The resulting phase symmetry response is determined as:

\[
PS(x,y) = \frac{\sum_B(|B_d(x,y)| - |R_d(x,y)|)}{\sum_B\sqrt{B_d(x,y)^2 + R_d(x,y)^2 + \epsilon}} 
\] (3.10)

where \(\epsilon\) is a small number to prevent division by zero. This \(PS\) measure is a dimensionless, intensity invariant measure that enhances ridge-like structures [Kovesi 1996]. We further enhance the ridge-like bone/cartilage boundaries in \(PS\) using a multi-scale eigen-analysis of the Hessian of \(PS\) [47]. The Hessian of \(PS\), \(H\), is computed by convolving \(PS\) with the second order derivatives of the Gaussian filter bank,

\[
G(x,y,s) = \frac{1}{2\pi s^2} \exp\left(-\left(x^2 + y^2\right)/2s^2\right) 
\] (3.11)

where \(s\) is the scale or variance of a Gaussian filter. For the eigenvalues \(|\lambda_1| < |\lambda_2|\), a pixel on a ridge-like structure will exhibit \(|\lambda_1| \approx 0, |\lambda_2| \gg 0\). We enhance the ridge-like features in our \(PS\) as:

\[
SPS(x,y) = \begin{cases} 
0 & \text{if } \lambda_2 > 0 \\
(1 - \exp(-T)) \exp\left(-\frac{|\lambda_1|^2}{|\lambda_2|^2}\right) & \text{otherwise}
\end{cases}
\] (3.12)

59
where $\exp\left(-\frac{|\lambda_1|^2}{|\lambda_2|^2}\right)$ is a bright-ridge enhancing term, $(1 - \exp(-T))$ is a noise cancelling term with $T = |\lambda_1|^2 + |\lambda_2|^2$. SPS is normalized such that it ranges between 0 and 1.

**Results**  
We compared the qualitative performance of the SPS feature (on a sample 2D B-mode US image of the neonatal hip) against the performance of two other PS methods [43, 57], as shown in Figure 3.4. A radiologist marked the false positives, false negatives and falsely connected structures on the labrum, ilium, acetabulum, and triradiate cartilage. Our proposed method seem to extract these key structures better than the other methods (qualitative results show less false positives and false negatives along these key structures).

### 3.3.2 3D Filters-based Structured Phase Symmetry

To extract $SPS(x, y)$ from a 3D US volume, $I(x, y, z)$, we first compute the the local PS feature, $PS$, from $I(x, y, z)$ using the monogenic signal-based method described in Subsection 3.3.1. To segment the sheet-like bone and cartilage surfaces from the $PS$ feature volume, we deploy a multi-scale eigen-analysis of the Hessian matrix. For eigenvalues $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$, voxels on sheet-like structures will exhibit $|\lambda_1| \approx |\lambda_2| \approx 0$, $|\lambda_3| \gg 0$. We enhance the sheet-like features in our $PS$ volume as:

$$SPS = \begin{cases} 
0, & \text{if } \lambda_3 > 0 \\
(1 - \exp(-R_a^2))(\exp(-R_b^2))(1 - \exp(-S^2)), & \text{otherwise} 
\end{cases} 
$$

(3.13)

where $R_a = \text{abs}(2*|\lambda_3| - |\lambda_2| - |\lambda_1|)/|\lambda_3|$ is a blob eliminating term, $R_b = |\lambda_2|/|\lambda_3|$ is a sheet enhancing term and $S = \sqrt{|\lambda_1|^2 + |\lambda_2|^2 + |\lambda_3|^2}$ is a noise cancelling term.

Bone/cartilage boundaries also tend to attenuate an US beam more than other neighboring structures (e.g. soft-tissue, etc.). We further enhance the bone/cartilage structures in $SPS$ to form $CSPS$ using the attenuation-based method described in Subsections 3.2.1 and 3.2.2, i.e., we replace $PS(x, y)$ with $SPS(x, y)$ in Equations 3.7 and 3.6.
Figure 3.4: Example qualitative results. (a) B-mode US image of the femoral head, (b) monogenic signal based PS, (c) directional filter based PS, (d) SPS. Red circles point to false negatives, and green circles point to falsely straightened structures, on the labrum, ilium, acetabulum, and triradiate cartilage. Blue circles point to falsely connected structures. Along the bone and cartilages contours of labrum, ilium, acetabulum, and triradiate cartilage, SPS has less false positive and false negative.
Parameter Specification: The free parameters in our automatic algorithm are physically meaningful and can be set to reflect values characteristic of the anatomy. In the bone and cartilage extraction pipeline, the log-Gabor filter bank uses a value of $p$ ranging from $p_{\text{min}} = 1.5\text{ mm}$ to $p_{\text{min}} = 4\text{ mm}$ to reflect the typical widths of bone interfaces in US images. The standard deviation parameter was set as $\sigma_{\omega} = 1/3 \times p_{\text{min}}$ to reduce speckle noise (in our experience, speckle in this type of image is typically under 0.5 mm). These bone/cartilage width values are also reflected in setting the scale or variance parameter, $s$, of the Gaussian filter bank: $s = 1.5\text{ mm}/(4 \times \sqrt{2\ln2})$ and $s = 4\text{ mm}/(4 \times \sqrt{2\ln2})$. This relation between ridge-width and variance parameter is derived using the relation between the full width at half maximum and the variance in a Gaussian distribution.

Data We acquired ten 3D B-mode US volumes from 15 infant hips - all obtained as part of routine clinical care under appropriate institutional review board approval (UBC CREB number: H14–01448). All 3D scans were collected in the coronal plane. The scans were performed using a SonixTouch Q+ machine with a 4DL14–5/38 Linear 4D transducer (BK ultrasound, Analogic corporation, MA, USA). The transducers center ultrasound frequency was kept at 7.5 MHz and image depth setting ranged between 3.8 cm – 5 cm. While performing a 3D US scan, the operator tried to capture the entire femoral head within the scanned US volume. In each scanned volume, an orthopaedic surgeon further selected a slice near the middle of the femoral head and labeled the ilium boundary, labrum and acetabulum boundaries.

Validation Scheme We compared our 3D confidence-weighted SPS method (CSPS)-based bone segmentation against optimized PS(OPS)-based bone segmentation [59]. First, we extracted ilium bone surfaces separately from CSPS (determined as the maximum of feature responses along the scan-lines of the US transducer) and OPS (determined as the bottom-most, or furthest away from the US transducer, non-zero feature responses along the scan-lines of the US transducer) along each scan-line where the surgeon labeled an ilium boundary. Next, we separated the segment of the US image that is inferior-lateral (right-top in the image) from
the edge of the ilium (right-most ilium point in the image) - similar to extracting ilium bone boundaries, here we extracted labrum boundaries using both CSPS and OPS. We also separated the segment of the US image that is superior-lateral (bottom-top in the image) from the edge of the ilium (right-most ilium point in the image) - similar to extracting ilium bone boundaries, here we extracted acetabulum boundaries using both CSPS and OPS.

Finally, we calculated the absolute discrepancy of these ilium, labrum and acetabulum boundaries with manually labeled ilium, labrum and acetabulum boundaries, only at the scan-lines that were manually labeled to contain those structures.

**Results** Exemplary results in Figure 3.5 show that CSPS-based bone/cartilage boundaries are closer to manually-labeled bone/cartilage boundaries compared to OPS-based bone/cartilage boundaries.

When evaluated on all 15 US volumes, CSPS-based bone/cartilage boundaries had a markedly reduced absolute discrepancy with manually-labeled bone/cartilage boundaries compared to OPS (ilium: average absolute discrepancy with CSPS = 0.96mm, average absolute discrepancy with OPS = 2.91mm, \(p < 0.01\) using the Kolmogorov–Smirnov test, Figure 3.6; labrum: average absolute discrepancy with CSPS = 1.06mm, average absolute discrepancy with OPS = 2.08mm, \(p < 0.01\) using the Kolmogorov–Smirnov test, Figure 3.7; acetabulum: average absolute discrepancy with CSPS = 0.82mm, average absolute discrepancy with OPS = 2.16mm, \(p < 0.01\) using the Kolmogorov–Smirnov test, Figure 3.8). The root mean square discrepancy for CSPS was moderate: 1.23mm (SD = 1.1mm) for ilium, 1.4mm (SD = 1.2mm) for labrum, and 1.02mm (SD = 0.7mm) for acetabulum. This root mean square discrepancy values for our CSPS method was less than Hareendranathan’s CNN-based segmentation (1.8mm [66]).

In terms of computational cost, processing time for our CSPS required around 30 seconds for a 240 \(\times\) 240 \(\times\) 120 US volume (3D) and around 0.2 second for a 240 \(\times\) 240 US image (2D) (run on a Xeon(R) 3.40 GHz CPU computer with 8 GB RAM with MATLAB code). The computation time for Hareendranathan’s method was not reported [66].
Figure 3.5: Example qualitative results. (a), (h) B-mode US volumes. (b), (i) US volume segments along the posterior-anterior axis that contains the ilium, acetabulum and labrum. (c), (j) OPS responses of (b) and (i), respectively. (d), (k) CSPS responses of (b) and (i), respectively. (e), (l) US slices near the middle of the femoral head chosen by the orthopaedic surgeon for the US volumes (b) and (i), respectively. Yellow, green and red contours are manually labeled ilium, acetabulum and labrum, respectively. (f) and (m) show OPS-based contours of ilium, acetabulum and labrum. (g) and (n) show CSPS-based contours of ilium, acetabulum and labrum. White lines point to false positive bone boundaries.
Figure 3.6: Cumulative distribution function of absolute discrepancy between manually-labeled ilium and OPS- and CSPS-based ilium boundaries.

Figure 3.7: Cumulative distribution function of absolute discrepancy between manually-labeled labrum and OPS- and CSPS-based labrum boundaries.
Figure 3.8: Cumulative distribution function of absolute discrepancy between manually-labeled acetabulum and OPS- and CSPS-based acetabulum boundaries.

3.3.3 Discussion

We proposed a rotation-invariant local symmetry feature, SPS, for segmenting non-planar bone boundary structures (e.g. small round bone boundary, etc.) in both 2D and 3D US images of infant hips.

Regarding extracting bone and cartilage boundaries in infant hips, we evaluated performance of CSPS on fifteen 3D US volumes. We found a markedly reduced discrepancy between CSPS-based and manually-labeled bone and cartilage boundaries, when compared with the OPS-based method.

We use our proposed SPS and CSPS in extracting dysplasia metrics from 2D (Chapter 4) and 3D US images (Chapter 5).

3.4 Summary

In this chapter, we investigated segmenting bone boundary from US images, which is particularly important in DDH diagnosis since all qualitative and quantitative characterizations of DDH require identifying bone structures in infant hips.

We proposed combining local symmetry features with attenuation-related features - our experiments showed that combining these features produce bone seg-
mentations that are robust to soft-tissue outliers. We also presented a rotation-invariant local symmetry measure, $SPS$, that makes use of the structural information in US images (i.e., vessel-like structures of bone boundary in 2D US images and sheet-like structures of bone boundary in 3D US volumes). On a dataset of 15 US volumes collected from 15 infant hips, we showed that bone and cartilage boundaries labeled by an expert agree well (mean absolute discrepancy around $1\text{mm}$) with bone and cartilage boundaries extracted using our CSPS method.

We show in later chapters (Chapter 4 and Chapter 5) that the $SPS$ and $CSPS$ features can be used for reliable diagnosis of DDH.
Chapter 4

Characterizing Hip Joint using 2D Ultrasound

4.1 Introduction

Although current clinical practice for DDH at early infancy is to use 2D US imaging \([104, 147]\), there is continuing controversy regarding the effectiveness of US in providing accurate early diagnosis and in guiding treatment decisions \([79, 106, 147]\). This controversy regarding the effectiveness of US in DDH diagnosis is largely due to the lack of reliability of dysplasia metrics - our systematic review and metaanalysis in chapter 2 showed that the most reliable dysplasia metric, \(\alpha\) angle (angle between the acetabular roof and the vertical cortex of the ilium), suffers from poor repeatability between repeated measurements (i.e. \(\approx 40\%\) variability, where variability is measured as the ratio of standard deviation between repeated measurements to the range of \(\beta\) angle between a normal and dysplastic hip). Our goal in this chapter is to investigate methods of improving repeatability or reducing variability in repeated 2D US-based DDH diagnosis. Here, we hypothesize that an US-based automatic dysplasia metric estimator could significantly reduce measurement variability as compared to a manual dysplasia metric estimator.

In a typical 2D US-based DDH diagnosis, images used to compute dysplasia metrics are first judged by an US technologist, radiologist or orthopaedic surgeon to be adequate - i.e., to contain key hip joint structures (see example in Figure 4.1).
These images are then manually analyzed to extract dysplasia metrics. Thus, given a 2D US image, an US-based automatic dysplasia metric estimator would need to identify the adequacy of that US image, and if adequate, would need to localize bone and cartilage boundary, and measure dysplasia metrics in that image.

While we are not aware of any study that reports any method of identifying adequacy of US scans for diagnosing DDH, there has been similar work in other medical ultrasound applications, including the fetal face [44], fetal abdomen [130], fetal head [95], gestational sac [166] and phantom [91]. Previous methods utilized supervised classification approaches (e.g. random forests [102], AdaBoost [130, 166], support vector machines [95], probabilistic boosting trees [44], deep learning [164]) and are driven by structural features, which are specifically Haar-like features [44, 102, 130, 166] and dynamic texture models [91, 95] or automatically learned features for deep learning. The main limitations to these methods are that they are generally learned and validated on mostly normal patient data. For our application, we need to identify adequate images in both healthy and hip dysplasia patients; to the best of our knowledge, no methods have been yet proposed to automatically identify adequate US images of the neonatal hip.

After identifying the adequacy of an US scan, the next step is to localize bone and cartilage boundaries that are necessary for extracting dysplasia metrics (e.g. vertical cortex of the ilium, etc.). A few researchers have proposed semi-automatic methods to segment bone/cartilage boundaries in US images of the neonatal hip, which they have used to extract dysplasia metrics. For example, Hareendranathan et al. [64] have recently proposed measuring a contour $\alpha$ angle that is derived from segmented ilium and acetabulum boundaries in a 2D US image. Their segmentation uses a graph-search-based technique that requires input of two seed points by an expert clinician. They found that this contour $\alpha$ has slightly lower intra-exam variabilities compared to the standard $\alpha$ angle ($\Delta\sigma = 0.2^\circ$ for the first rater and $0.4^\circ$ for the second rater). In 2016, Golan proposed [48] proposed using CNN to automatically segment ilium and acetabulum boundary in US images of the neonatal hip, and used the segmented bone boundary to estimate $\alpha$ angle values. The authors demonstrated that the Graf angles extracted with their method correlated well with an expert’s annotation (correlation coefficient, $r = 0.76$, CI not reported), but the authors did not report variability of these dysplasia metrics. A potential lim-
itation of Golan’s method is that they used a human expert’s annotation for training their CNN - this can be problematic since the variability of labeling/measuring is markedly high even when the labeling/measuring is done by experts [79, 94, 106]. So, Graf angles extracted by Golan’s method may correlate well with an expert, but that does not necessarily mean that the variability in the automatically extracted angles will be significantly smaller as compared to the variability in the manually estimated angles. In an earlier work, de Luis-Garcia [31] attempted to segment the femoral head using an energy function that uses both gray level and texture information from 3D US image and requires an user to initialize the center of the femoral head. The authors presented only one qualitative example and did not report any quantitative analysis.

Figure 4.2 shows a flow-chart outlining the steps involved in our computer-assisted 2D US-based DDH diagnosis. In our work, to localize bone and cartilage boundaries in a 2D US scan, we used the 2D local symmetry enhancing method and bone boundary segmentation method described in Chapter 3, and prior geometric information regarding the relative geometry of bone and cartilage boundaries (e.g. ilium superior to the femoral head) (section 4.2). Next, we deployed a random forests-based approach to identify adequacy of a US scan (section 4.3). Finally, after localizing the bone and cartilage boundary and identifying adequacy of the 2D US scan, we presented methods to automatically extract the standard clinically-used dysplasia metrics: the $\alpha$ angle, $\beta$ angle and FHC (Figure 4.1) (section 4.4). The $\beta$ angle is the angle between the vertical cortex of the ilium and triangular labral fibrocartilage, and FHC is the ratio of the acetabular width to the maximal femoral head diameter.

4.2 Localizing Bone and Cartilage Boundaries

To extract bone/cartilage boundaries in a 2D US image ($U$) of an infant hip, we used a rotation-invariant local symmetry feature, structured phase symmetry (SPS) (sub-section 3.3.1) (Figure 4.3b), which extracts vessel-like hyperechoic responses in $U$. After segmenting vessel-like hyperechoic responses in $U$, which includes bone boundaries, cartilage boundaries and soft-tissue interfaces, we use the relative signal strength map-based (Figure 4.3c) attenuation features to enhance bone
Figure 4.1: Example coronal US images of the neonatal hip. (a) Adequate US image showing key hip joint structures. (b) Inadequate US image (missing ilium and labrum). (c), (d) The $\alpha$, $\beta$ angles and FHC measurements extracted manually from the adequate US image in (a).

Figure 4.2: A flow-chart outlining the steps involved in our computer-assisted 2D US-based DDH diagnosis.

boundaries and remove soft-tissue outliers [subsection 3.2.2](Figure 4.3(d)). While our extracted bone and cartilage boundaries approach do have false positive and false negative, our dysplasia metric extraction is based on straight line approximations [section 4.4](that are robust to such outliers and missing boundary points (i.e., false positive and false negative).

4.2.1 Ilium, Labrum and Acetabulum

Figure 4.4 shows a block diagram outlining the extraction of ilium, labrum and acetabulum. In adequate scans, the bony ilium ($I$), is oriented approximately parallel to the US transducer. To localize $I$, we first use the Radon transform ($R(\theta,x_p)$) of
CPS, where $\theta$ is the counter-clockwise angle from downwards vertical line and $x_p$ is the distance from the center of the image) bounded between $\theta = 85^\circ$ to $\theta = 95^\circ$ and find the straight line corresponding to the maximum response of $R$ - this straight line provides a straight line approximation of $I$. We remove all bone responses that are 1 mm away from $I$. Next, we use a binary image formed by CPS measures with eigenvectors of $\lambda_1$ (minor eigenvalues of the Hessian matrix [subsection 3.3.1]) within $\pm 15^\circ$ of any lines parallel to the US transducer - the resulting binary image includes bone/cartilage boundaries that are reasonably parallel to the US transducer. Next, we thin the responses in the binary image to widths of 1 pixel, and calculate the areas of all the connected structures. Since widths of the structures are 1 pixel, the area responses are approximately equal to the lengths of the structures. We identify the ilium as the contour that has the highest length, and label the inferior edge of the ilium as $i$ (Figure 4.3(d), Figure 4.4).

Since the labrum ($L$), acetabular roof ($A$) and $I$ branch off from around $i$, we isolate the responses of SPS belonging to $A$, $L$ and $I$ by masking SPS with a circle having center $i$ and radius $r_f$, with $r_f$ being the radius of a typical femoral head (initialized as 9 mm based on empirical observations [51]). We estimate best-fit straight line characteristics of $I$, $L$ and $A$ using the peak responses of the Radon transforms on the regions of the masked SPS as shown in Figure 4.4. We refine the estimated ilium edge point, $i$ based on the intersection of approximated lines of $I$ and $A$. Afterwards we extract center and radius of femoral head using the method outlined in subsection 4.2.2, and use the updated estimate of femoral head radius $r$ to recompute $I$, $L$ and $A$.

4.2.2 Femoral Head

To locate the center of the femoral head in $U$, $F$ (radius $r_f$, center $c$), we extract a total of 28 features that capture orientation, texture and geometrical features of the femoral head, and input them to a classifier that identifies whether a pixel location in the US image belongs to the femoral head or not. We use random forests [15] as our femoral head classifier model to learn the probability $p(y_i|\phi(x_i))$ of a pixel location in the US image being part of the femoral head. The random forests classifier uses an ensemble of decision trees (weak learners), each grown using a
bootstrap sample of the training data, and randomly selected subsets of predictor variables as candidates for splitting tree nodes [15]. We chose random forests due to their superior performance and robustness to parameter tuning compared to other machine learning approaches such as artificial neural network, k-nearest neighbors, Naive Bayes and support vector machine [15, 20, 26, 45]. The input feature vector $\phi(x_i)$ of 28 features (Figure 4.5) for our femoral head random forests classifier consist of seven geometric features, twelve texture features and nine orientation features. The lowest out-of-bag feature importance among all these features was 0.6, suggesting that all of these features are important in segmenting the femoral head. These 28 features (shown in Figure 4.5) are listed as follows:

- Geometric features: we extract the geometric features primarily from the identified locations of $I$, $L$, $A$ and edge of ilium, $i$: (1) distance from $I$, (2) distance from $A$, (3) distance from $L$, (4) distance from the line formed by an angular bisection of $A$ and $L$, (5) vertical distance from $i$, (6) horizontal distance from the edge of ilium, $i$, and (7) depth (Figure 4.5).

- Texture features: (1) median-filtered (filter size 1mm by 1mm) intensity of $U$, (2) range-filtered (filter size 1mm by 1mm) intensity of $U$, (3) range-filtered (filter size 2mm by 2mm) intensity of $U$, (4) range-filtered (filter size 3mm by 3mm) intensity of $Ue$, (5) entropy-filtered (filter size 2mm by 2mm)
Figure 4.4: Block diagram showing extraction of $I$, $L$ and $A$. 
intensity of $U$, (6) median-filtered (filter size 1mm by 1mm) confidence map value of $U$, (7) range-filtered (filter size 1mm by 1mm) confidence map value of $U$, (8) range-filtered (filter size 2mm by 2mm) confidence map value of $U$, (9) range-filtered (filter size 3mm by 3mm) confidence map value of $U$, (10) entropy-filtered (filter size 2mm by 2mm) confidence map value of $U$, (11) circular mask (radius 2mm)-filtered intensity of $U$, and (12) a blobness measure which we define as $blob = (1 - \exp(-\frac{|\lambda_1|^2}{|\lambda_2|^2}))$ (Figure 4.5). Here, $|\lambda_1|$ is the minor eigenvalue and $|\lambda_2|$ is the major eigenvalue in the Hessian-based analysis in subsection 3.3.1. The rationale for the blobness measure is that a blob is characterized by high minor and major eigenvalues with $|\lambda_1|$ close to $|\lambda_2|$, leading to a low value for $\exp(-\frac{|\lambda_1|^2}{|\lambda_2|^2})$ and a high value for $blob$.

- Orientation features: (1) we extract nine orientation features from the histogram of oriented gradients (HOG) (Figure 4.5) [29].

After extracting the probability map $p$ of the femoral head (Figure 4.6 b), we apply a threshold of $p > 0.5$ and then a Sobel filter to find edges of the femoral head (Figure 4.6 c) [19]. Next, we apply the circular Hough transform [8], and identify the peak to estimate the center coordinates and radius of the circular femoral head (Figure 4.6 d). We use these center coordinates and radius in evaluating adequacy and in localizing ilium and acetabulum.

### 4.3 Scan Adequacy Classification

An adequate US image in DDH diagnosis includes the following five key bone/cartilage structures: the acetabular roof, ilium, labrum, ischium and femoral head. Given the extracted bone/cartilage boundaries obtained using the methods outlined in the previous section, we extract 26 geometrical features pertaining to these key structures and input them to our adequacy classifier. We also use HOG [29] and local binary patterns (LBP) [105] features since using these two features together can significantly improve object detection performances as compared to using only one of those features in object detection [162]. A complete list of the features used in our adequacy classifier are shown in Table 4.1.
Figure 4.5: Features used in localizing femoral head. Note that only the first HOG feature is shown in here.
Figure 4.6: (a) B-mode US image. (b) Output of femoral head random forests classifier, $p$. (c) Sobel filtering on $p$ generates edges of the femoral head. (d) Circular Hough transform applied on (c). Peak of circular Hough transform (yellow arrow) provides an estimate of the center coordinates and radius of the femoral head.

4.3.1 Experiment

Parameter Specification For our random forest classifier, we selected the number of trees of the random forest as 70 (selected based on out-of-bag classification error analysis - the out-of-bag classification error does not seem to decrease considerably after number of trees $= 70$, Figure 4.7). Similar to the original implementation of random forests, the trees were not pruned and the number of features to select at random for each decision split was selected as the square root of the number of features [10][15].
Table 4.1: Features used in adequacy classifier.

<table>
<thead>
<tr>
<th>Type of features</th>
<th>Feature details</th>
</tr>
</thead>
</table>
| Geometric features | (1) length of ilium  
(2) horizontal length of ilium  
(3) vertical length of ilium  
(4) aspect ratio of ilium (ratio of vertical length and horizontal length)  
(5) $x$ and $y$ coordinates of edge of ilium (both initial and final estimation, 4 features)  
(6) gradients of the approximated lines of $I$, $L$ and $A$ (both initial and final estimation, 6 features)  
(7) peak Radon transform responses corresponding to lines of $I$, $L$ and $A$ (both initial and final estimation, 6 features)  
(8) center coordinates $c$ and radius $r$ of the femoral head (3 features)  
(9) maximum US image intensity values inside each of the three circles having center $c$ and radius $r$, $r/2$ and $r/3$ |
| Texture features | (10) LBP (59 features) |
| Orientation features | (11) HOG features (144 features) |

Figure 4.7: Out-of-bag classification error vs. number of grown trees in random forest classifiers for adequacy classification. Out-of-bag classification error does not seem to decay considerably after number of trees $= 70$. 
Data  We obtained access to 885 US images from both left and right hips (obtained as part of routine clinical care with appropriate institutional review board approval and with informed consent from each participant’s parents) of 69 infants in multiple US sessions (total of 82 US sessions with both hips and 1 US session with 1 hip, 82*2+1*1=165 hip examinations) at British Columbia Children’s Hospital (UBC CREB number: H14–01448). The scans were performed by multiple clinicians from the health center as part of routine clinical care using a LOGIQ E9 with XDclear premium ultrasound system and a GE ML6-15-D Matrix Linear Probe (LOGIQ E9 with XDclear, General Electric Healthcare, Waukesha, WI, USA). The transducers center ultrasound frequency was kept within 7 MHz – 11 MHz. Each patient underwent a minimum of 1 and a maximum of 4 sessions of US scanning based upon the recommendation of an orthopaedic surgeon or radiologist. The available US images in the health record consists of images of the femoral head that were taken from a range of orientations. The available US images in the health record consisted of images of the femoral head that were taken from a range of orientations. A subset of these US images (444/885) were judged by the radiologist to be adequate for use in extracting dysplasia metrics (i.e., $\alpha$ and $\beta$ angles). An orthopedic surgeon reviewed the remaining set of US images to determine if any additional images might be considered adequate for extracting dysplasia metrics. A subset of 20 possibly adequate images were identified and submitted to a radiologist for definitive classification. Of these, an additional 3 images were accepted as adequate and added to the original 444, producing a total of 447 images that were deemed adequate by a radiologist; the radiologist also extracted dysplasia metrics from these three newly added images.

Validation Scheme  The adequate/inadequate labels for the 885 US images (collected from 69 infants) were used as the gold standard for analyzing performance of the automatic scan adequacy method. We randomly selected 60 patients’ data for training and cross-validation, and kept the remaining 9 patients’ data for testing. The test data contained 47 adequate images and 53 inadequate images. The cross-validation dataset consisting of 60 patients’ data was randomly split into five sets of 12 patients’ data. Four of these sets were used in training and the remaining one was kept for validation. This was repeated five times to evaluate classifier perfor-
mance on the entire cross-validation dataset. We also evaluated the performance of our 229 features against only HOG features by using both of them separately as input feature vectors to adequacy random forests classifiers. We chose HOG features on the basis of their recent success in binary classification of ultrasound images [85], and their extensive use in image classification applications [29]. We calculated the difference between receiver operating characteristics (ROC) curves of different classifiers [60] to evaluate the null hypothesis that there was no change in classification performance between the different classifiers.

**Results** Figure 4.8 shows sample classification results of our proposed random forests classifier (probability$_{RF} > 0.5$: adequate) from amongst the set of 885 US images, along with the corresponding labels provided by the clinicians. The majority of the automatic assessments agreed with the manual assessments (average area under the ROC curve of around 99.5% on both the cross-validation dataset and the test dataset - see Figure 4.9). This is significantly better than our previously reported value of 95% for a classifier based on a single feature (horizontal length of the ilium), $p < 0.01$ [124]. Furthermore, the random forests classifier is significantly better than the HOG and LBP features-based random forests classifier (area under the ROC curve: 93%, $p < 0.01$).

On the 20 borderline adequate images that were later submitted to the radiologist for adequacy assessment, the agreement between the adequacy classifier’s output and the expert’s assessments was low (35% agreement).

### 4.3.2 Discussion

The excellent agreement between the 229 features-based adequacy classifier output with that of radiologists’ assessments suggests that the adequacy classifier’s assessments can be similar to that of an expert clinician. However, the dataset we used mostly involved clearly inadequate (424 US images) and clearly adequate US images (441 US images), and only a small portion of borderline US images (20 US images), so the excellent agreement between the adequacy classifier and an expert assessment is primarily indicative of differences in the discriminating abilities of the different classifiers we have evaluated on clearly adequate and clearly inade-
Figure 4.8: Example scan adequacy classification results. (a) Typical 2D B-mode images that were judged adequate by a radiologist (i.e., identified manually). Arrows in the images point to the key structures that were identified manually. In an adequate image, all of the key five structures need to be present. The first four images were classified as adequate by our method, while the last image was classified as inadequate (probability metric = 0.46). (b) Images that were judged to be inadequate by the radiologist. The first four images were classified as inadequate by our method, while the last image was classified as adequate (probability metric = 0.51). Based on the probability metrics produced by our classifier, both of these disagreements in classification could be considered to be borderline cases.

In the borderline US images, the agreement between the adequacy classifier’s output and the expert’s assessments was low (35% agreement). Thus, to better understand the adequacy classifier’s performance in a live clinical setting, it would be necessary to collect a broader range of US images along with reference labels provided by experts. Correct labels are particularly important since it can directly affect the learning of the classifier and subsequently the classifier’s performance in identifying adequate scans. This is currently a challenging task since manual
Figure 4.9: ROC curves of our 229 features-based random forests classifier. (a) The five different plots correspond to the five cross-validation experiments we conducted on the validation dataset. (b) The five different plots correspond to the five classifiers each from a different cross-validation dataset and evaluated on the test dataset.

adequacy classification is prone to mistakes, even when it is done by experts. For example, one study reported that around 44% of medical doctors in one German state made mistakes in localizing the bone/cartilage boundaries and adequate US scans [53]. One possible approach for generating more reliable adequacy labels is to combine labels obtained from multiple experts and then use the label that most of the experts agreed with.

4.4 Extraction of 2D Dysplasia Metrics

In US-based dysplasia assessment, three measures are most commonly used by clinicians: the $\alpha$ angle, $\beta$ angle and $FHC$ (Figure 4.1). The $\alpha$ angle is the angle between the acetabular roof and vertical cortex of the ilium, whereas the $\beta$ angle is the angle between the labrum and vertical cortex of the ilium. We calculate $\alpha = R_I - R_A$ and $\beta = R_L - R_I$, where $R_I$, $R_A$ and $R_L$ are the characteristic angles of the ilium, acetabular roof and labrum. The $FHC$ is defined as the ratio of the acetabular width to the maximal femoral head diameter (Figure 4.1). We calculate $FHC$ as $FHC = \frac{\sum M}{\sum X} P$, where $P$ represents the output of the femoral head
random forests classifier, $X = [x, y]$ and $M$ represents all $X$ locations that are medial to $I$.

### 4.4.1 Experiment

**Data:** We used the same dataset that we described in validating scan adequacy.

**Results for extracting dysplasia metrics in all hips:** Results for extracting 2D dysplasia metrics in all hips are shown in Appendix A.

**Validation Scheme:** We compared the manual measurements, $\alpha_m, \beta_m$ and $FHC_m$, with the automatically extracted measurements, $\alpha_a, \beta_a$ and $FHC_a$, in the same images that were judged adequate by the radiologists. Specifically, we analyzed the following:

- The discrepancies between the two sets of measurements (i.e., $\alpha_m$ vs. $\alpha_a$, $\beta_m$ vs. $\beta_a$, and $FHC_m$ vs. $FHC_a$).
- The within-hip variability in the measurements (i.e., $\text{std}(\alpha_m)$, $\text{std}(\alpha_a)$, $\text{std}(\beta_m)$, $\text{std}(\beta_a)$, $\text{std}(FHC_m)$ and $\text{std}(FHC_a)$) for each hip of the 165 hip examinations.
- We evaluated our results in terms of Graf-DDH classifications, i.e., type-I ($\alpha > 60^\circ$, mature hip), type-II ($43^\circ < \alpha < 60^\circ$, physiological immature hip) and type-III ($\alpha < 43^\circ$, eccentric hip) \[79\]. We used maximum values of $\alpha$ for each of the classifications based on the assumption that a mature (normal) hip will have at least some orientations that show a high value of $\alpha$, but eccentric hips usually do not have orientations that will show a high value of $\alpha$ \[4\].

**Discrepancy between Manual and Automatic Metrics:** Figure 4.10 shows example images with low (Figure 4.10 a, c and e) and high discrepancies (Figure 4.10 b, d and f) between manual and automatic measurements (i.e., $\alpha_a$ vs. $\alpha_m$, $\beta_a$ vs. $\beta_m$, $FHC_a$ vs. $FHC_m$).
\( \beta_M \), and \( FHC_A \) vs. \( FHC_M \). Overall, the discrepancies (i.e., automatic-manual measurements) for \( \alpha \) angles and \( FHC \) were high and statistically significant (\( \Delta \alpha \): mean 4.64°, SD 6.1°, \( p < 0.01 \) (Figure 4.11); \( \Delta FHC \): mean 5.94%, SD = 7.0%, \( p < 0.01 \) (Figure 4.13)) compared to a larger and also statistically significant discrepancy for \( \beta \) (\( \Delta \beta \): mean 8.14°, SD 7.2°, \( p < 0.01 \) (Figure 4.12)). The correlation (\( R \)) between manual and automatic measurements was high and statistically significant, suggesting the presence of a relationship between the manual and automatic measures (\( \alpha \): \( R = 0.8 \) [CI 0.76 and 0.83], \( p < 0.01 \) (Figure 4.11); \( \beta \): \( R = 0.72 \) [CI 0.67 and 0.76], \( p < 0.01 \) (Figure 4.12); \( FHC \): \( R = 0.77 \) [CI 0.72 and 0.80], \( p < 0.01 \) (Figure 4.13)).

Appendix A shows other examples for automatic extractions of 2D US-based dysplasia metrics in a healthy hip, a borderline hip and a dysplastic hip – we found that our approximations of the ilium, labrum, acetabulum and femoral head and our estimation of the 2D US–based dysplasia metrics are realistic.

**Variability of Metrics** Since an ideal dysplasia metric will have minimal variability across multiple images in a hip examination, we calculate the standard deviations within each of the 165 hip examinations to investigate the overall reproducibility of each method. Box plots of all the standard deviation values are shown in Figure 4.14, Figure 4.15 and Figure 4.16. We found that the automatic method results in modest, but statistically reduced variability for all the dysplasia metrics (\( \alpha \), \( \beta \) and \( FHC \)) compared to variability of manually-extracted dysplasia metrics (Wilcoxon signed rank test: \( p < 0.01 \) for \( \alpha \), \( \beta \) and \( FHC \); \( \text{mean}(\sigma_{\alpha_M}) = 3.09^\circ \) and \( \text{mean}(\sigma_{\alpha_A}) = 2.45^\circ \), \( \text{mean}(\sigma_{\beta_M}) = 4.4^\circ \) and \( \text{mean}(\sigma_{\beta_A}) = 3.6^\circ \), \( \text{mean}(\sigma_{FHC_M}) = 5.27\% \) and \( \text{mean}(\sigma_{FHC_A}) = 4.31\% \).

We further evaluated whether there is any association in the variability of dysplasia metrics with the severity of hip dysplasia. We did not find any statistically significant association in the variability of dysplasia metrics with the severity of hip dysplasia for \( \alpha_M \) (\( R = -0.08 \) [CI -0.23 and 0.07], \( p = 0.3 \) (Figure 4.17) and \( \alpha_A \) (\( R = -0.03 \) [CI -0.18 and 0.12], \( p = 0.69 \) (Figure 4.18). We also did not find any statistically significant association in the variability of dysplasia metrics with the severity of hip dysplasia for measuring \( \beta_M \) (\( R = 0.12 \) [CI -0.04 and 0.14], \( p = 0.14 \) (Figure 4.19) and \( \beta_A \) (\( R = 0.14 \) [CI -0.01 and 0.29], \( p = 0.07 \) (Figure 4.20). However,
Figure 4.10: Example qualitative results of manual (labelled as 'M' at right top corners) and automatic measurements (labelled as 'A' at right top corners). (a), (c) and (e) show examples where manual and automatic measurements produced similar values for $\alpha$, $\beta$ and $FHC$, whereas (b), (d) and (f) show examples where the manual and automatic measurements differed noticeably. The white dotted lines represent the manual measurements, while the blue dotted lines represent the automatic measurements and the green in (e) and (f) represent femoral head random forests’ probability output.
we did find a statistically significant association in the variability of dysplasia metrics with the severity of hip dysplasia for $FHC_M (R = -0.28 \ [CI -0.42 \ and \ -0.13], \ p = 0.0003$ [Figure 4.21]) and $FHC_A (R = -0.25, \ C.I. -0.38 \ and \ -0.1, \ p = 0.001$ [Figure 4.22]), where the variability of $FHC$ tends to increase in patients with hip dysplasia.

**Graf Classification Agreement:** The Graf classifications for the manual and automatic methods are shown in Table 4.2 Overall, 148 of the 165 hip examinations were classified identically by the two methods (84% agreement, substantial Kappa coefficient agreement of $= 0.61$ (CI:0.48 and 0.75) [97]). We looked into the first
Figure 4.12: Scatter plot of automatic versus manual measurements $\beta$ angle. The red line is the equality line. Blue data points correspond to lower $\beta_A$ and yellow data points correspond to higher $\beta_A$, so blue correspond to hips that would be judged as normal whereas yellow correspond to hips that would be judged as hip dysplasia.
Figure 4.13: Scatter plot of automatic versus manual measurements $FHC$ angle. The red line is the equality line. Blue data points correspond to lower $FHC_A$ and yellow data points correspond to higher $FHC_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal.
Figure 4.14: Box-plot of the within-hip standard deviations among the manual and automatic measurements of $\alpha$ angles within all of the 165 hip examinations (i.e., 165 values of $\text{std}(\alpha_M)$ and 165 values of $\text{std}(\alpha_A)$). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The ‘+’ points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in the automatic $\alpha$ angle measurements ($p < 0.01$).
Figure 4.15: Box-plot of the within-hip standard deviations among the manual and automatic measurements of $\beta$ angles within all of the 165 hip examinations (i.e., 165 values of $\text{std}(\beta_M)$ and 165 values of $\text{std}(\beta_A)$). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The ‘+’ points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in the automatic $\beta$ angle measurements ($p < 0.01$).
Figure 4.16: Box-plot of the within-hip standard deviations among the manual and automatic measurements of $FHC$ angles within all of the 165 hip examinations (i.e., 165 values of $\text{std}(FHC_M)$ and 165 values of $\text{std}(FHC_A)$). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The ‘+’ points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in the automatic $FHC$ angle measurements ($p < 0.01$).
Figure 4.17: Scatter plot of variability of $\alpha_M$ vs. mean $\alpha_M$. Blue data points correspond to lower mean values of $\alpha_M$ and yellow data points correspond to higher mean values of $\alpha_M$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\alpha_M$ with the severity of hip dysplasia.

nine cases of disagreement. In these cases, the automatic method graded the hip one class more severely than the manual method. In three hip examinations from two patients, the automatic method found a type-II hip while the manual method found a type-I hip at their first US sessions (both hips for patient 1 and only the right hip for patient 2). Although these patients were deemed Graf type I based on the manual assessment, their $\alpha$ angles were borderline (61°, 62° and 60°) and they subsequently had a follow-up US session. In another two patients, the automatic method classified the patients’ hips as type III while the manual approach classi-
Figure 4.18: Scatter plot of variability of $\alpha_A$ vs. mean $\alpha_A$. Blue data points correspond to lower mean values of $\alpha_A$ and yellow data points correspond to higher mean values of $\alpha_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\alpha_A$ with the severity of hip dysplasia.

fied them as type II in early US sessions (both hips for patient 1 in the first two US sessions, and the left hip for patient 2 in the first and third US sessions). Both of these patients were treated using a Pavlik harness based on the manual US findings, but were later both booked for surgery, indicating that the clinicians regarded these patients as having relatively severe dysplasia.

Computational Considerations: The process of image classification and dysplasia metric extraction is near real-time (approximately 2 seconds) for a B-mode
Figure 4.19: Scatter plot of variability of $\beta_M$ vs. mean $\beta_M$. Blue data points correspond to lower mean values of $\beta_M$ and yellow data points correspond to higher mean values of $\beta_M$, so blue correspond to hips that would be judged as normal whereas yellow correspond to hips that would be judged as hip dysplasia. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\beta_M$ with the severity of hip dysplasia.

Table 4.2: Graf-DDH classification of the neonatal hips. Here, type-I ($\alpha > 60^\circ$, mature hip), type-II ($60^\circ < \alpha < 43^\circ$, physiological immature hip) and type-III ($\alpha < 43^\circ$, eccentric hip). The agreement between automatic and manual classification is fair (Cohen’s kappa coefficient=0.61 (CI:0.48 and 0.75)).

<table>
<thead>
<tr>
<th></th>
<th>Automatic</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type I</td>
<td>Type II</td>
<td>Type III</td>
</tr>
<tr>
<td>Type I</td>
<td>114</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Manual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type II</td>
<td>0</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Type III</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>
Figure 4.20: Scatter plot of variability of $\beta_A$ vs. mean $\beta_A$. Blue data points correspond to lower mean values of $\beta_A$ and yellow data points correspond to higher mean values of $\beta_A$, so blue correspond to hips that would be judged as normal whereas yellow correspond to hips that would be judged as hip dysplasia. The red line is the best fit line. Correlation and p values suggest that there is no significant association in variability of $\beta_A$ with the severity of hip dysplasia.

US image, when run on a Xeon(R) 3.40 GHz CPU computer with 8 GB RAM. All processes were executed using MATLAB 2015b (MATLAB 2017a, the Mathworks Inc., Natick, MA, USA).

4.4.2 Discussion

Discrepancy between Manual and Automatic Metrics: Overall, the discrepancies for $\alpha$ and $\beta$ angles seem to be highest in images in which the relevant structures
Figure 4.21: Scatter plot of variability of $FHC_M$ vs. mean $FHC_M$. Blue data points correspond to lower mean values of $FHC_M$ and yellow data points correspond to higher mean values of $FHC_M$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is significant association in the variability of $FHC_M$ with the severity of hip dysplasia.

(acetabular roof and labrum, respectively) do not appear to be simple line segments, but instead have multiple sections and/or more complex shapes (Figure 4.10(b, d)). In such instances, the operator appears to focus more on the portions of the line segments comprising the labrum and acetabulum that are more to the right in the images, whereas the automatic process seems to be more influenced by the first strong deviation from the ilium, so both $\alpha$ and $\beta$ angles tend to be higher in manual measurements than in automatic measurements (Figure 4.10(b, d)). The discrepancy between automatic and manual $FHC$ measurements seem to be higher where
Figure 4.22: Scatter plot of variability of $FHC_A$ vs. mean $FHC_A$. Blue data points correspond to lower mean values of $FHC_A$ and yellow data points correspond to higher mean values of $FHC_A$, so blue correspond to hips that would be judged as hip dysplasia whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and $p$ values suggest that there is significant association in the variability of $FHC_A$ with the severity of hip dysplasia.

Our automatic method produces a non-circular probability map (Figure 4.10(f)). Since femoral heads can be aspherical in shape, with a tendency for heightened aspherical shape in hip dysplasia [140], presuming that the femoral head is circular is perhaps a limitation of the manual measurement process.

The $4.6^\circ$ bias of the automatic $\alpha_A$ angle towards smaller values than the manual angles means that the automatic method will tend to classify more patients as having DDH than would be identified as such by the current manual method. This tendency of heightened suspicion may perhaps be acceptable from a clinical
perspective since we know that a large percentage of patients who initially had a diagnosis of DDH that subsequently resolved nonetheless had radiological signs of DDH a few months later (e.g., around 33% of such patients in Sarkissian et al.’s study [144]), though it is difficult to tell in such cases whether there was residual DDH not detected by 2D US or the DDH redeveloped post-initial exam.

In contrast, the more noticeable bias of 8.1° for the \( \beta \) angle (\( \beta_A < \beta_M \)) means that the automatic method’s measure would be interpreted as indicating greater stability in the labrum than if the same measure was obtained manually [52]. Likewise the automatically-extracted \( FHC \) values with a bias of \(-5.9\%\) (\( FHC_A > FHC_M \)) would similarly be interpreted as more indicative of relative ‘acetabulum-femoral head’ stability than would the corresponding manually-extracted values. Overall, the automatic method tends to classify more patients as having problematic shape of acetabulum, less patients having problematic shape of labrum and less patients having problematic (e.g. subluxed, dislocated) femoral head.

**Variability of Metrics:** The variabilities in the automatic measurements using our proposed method are statistically lower than those of the manual measurements (\( \approx 21\% \) reduction for \( \alpha \), \( \approx 18\% \) reduction for \( \beta \) and \( \approx 18\% \) reduction for \( FHC \)). However, the clinical significance of this reduction remains uncertain. This variability may still lead to considerable misdiagnosis rates since the difference in dysplasia metrics between mature and severely dysplastic hips can be around 17°. To further reduce the variability in dysplasia metrics, one would need to also address the more dominant transducer-orientation-dependent variability [79], which we have not taken into account in our current 2D US-based approach. Our proposed approach is therefore still susceptible to variability arising from images taken from different orientations. Thus, a potential direction to further reduce the variability in the measured dysplasia measures would be to implement a 3D US-based system that would be more robust to variations in the orientations/placements of the US transducer during image acquisition.

Regarding the variability in healthy and hip dysplasia, we found that the variability of \( \alpha \) measurements are independent of the degree of dysplasia since there was no association in the variability of \( \alpha \) measurements with the severity of hip dysplasia. For \( \beta \) measurements, variability seemed to increase when measured on
dysplastic hips, though the increase was not statistically significant. However, variability in both automatic and manual $FHC$ measurements to increased ($p < 0.01$) in dysplastic hips. This worsening reproducibility in dysplasia metric was also observed in Bar-On’s study [9], where the authors showed poor agreement between raters classifying hips based on Graf classification.

**Graf Classification Agreement:** We found a high agreement between manual and automatic Graf classifications (84% agreement). In nine cases of disagreement between the manual approach and our automatic method in the Graf-DDH type assignment, process-of-care (e.g. patient booked for follow-up US session, patient booked for surgery, etc.) support the classification decisions made by our proposed automatic method, suggesting it could potentially be of value in reducing missed early diagnosis rate (i.e., false negative rate) without increasing over-treatment (i.e., false positive rate). However, the number of such disagreements here is too small to state definitively that our classifier is more often in agreement with the ultimate clinical decisions than the current manual approach. More importantly, recent research suggests that using maximum values of $\alpha$ for Graf classifications is not reliable. The primary assumption behind using maximum values of $\alpha$ is that a mature (normal) hip will have at least some orientations that show a high value of $\alpha$, but eccentric hips do not have orientations that will show a high value of $\alpha$ [4]. Recent studies identified that eccentric hips can also result in high $\alpha$ values [79, 86].

### 4.5 Summary

In this chapter, we presented a novel, automatic, near real-time (around 2 seconds/US image) method to assess the adequacy of a 2D US image of the neonatal hip, and, when found adequate, to subsequently extract the patient’s dysplasia metrics. The proposed method produced excellent agreement with radiologists in scan adequacy classification. Furthermore, the automatic method reduced the variability of manually measuring dysplasia metrics by around 20%. Finally, one of our automatically extracted metrics, $\alpha_A$, more consistently provided reproducible measures in both healthy and dysplastic hips when compared with the other 2D dysplasia
A 2D US-based method such as ours, however, is not robust to variations in the orientation and placement of the US transducer during image acquisition. This variability may perhaps be reduced using a 3D-US based system, which can better capture the 3D morphology of an infant’s hip while remaining robust to transducer orientation during image acquisition [64, 79]. We address this issue in the next chapter.
Chapter 5

3D Ultrasound-based Hip Dysplasia Assessment

5.1 Introduction

We have demonstrated in [chapter 4] that automatically extracting 2D dysplasia metrics can potentially reduce between-user variability, however, it continues to be sensitive to the orientation and placement of the US transducer during image acquisition. This variability due to the 2D transducer location/orientation is particularly important since it can cause a normal hip to appear dysplastic and a dysplastic hip to appear normal [79]. This probe-orientation-dependent variability problem can potentially be resolved using a 3D US transducer. In contrast to a 2D US transducer, a 3D US can capture the entire femoral head and its neighboring structures (i.e., the vertical cortex of the ilium, acetabulum, labrum). For 3D US volumes that capture the entire femoral head and its neighboring structures in a normal infant hip, we hypothesize that this transducer can consistently extract the acetabular and femoral head morphology regardless of its location/orientation during acquisition. Dysplasia metrics derived from 3D hip morphology will thus be markedly less variable than the conventional dysplasia metrics derived from 2D US images.

To the best of our knowledge, only one previous work [64] has proposed the use of an intrinsically 3D dysplasia metric - the ACA. Similar to the $\alpha_{2D}$, the ACA represents the angular separation between the acetabular roof ($A$) and the lateral
iliac (I), except that the ACA is based on the segmented 3D surfaces of A and I. Hareendranathan et al.’s method [64] involves a slice-by-slice analysis process that requires manually selecting at least three seed points in three 2D US slices in a 3D US volume and manually separating A from I. Using such an interactive method would require valuable clinician time, and the manual operations introduce an approximately 1° within-image measurement variability [64] and approximately 4° inter-scan variability [94].

In this chapter, we define a form of 3D US-derived metrics analogous to the 2D α angle and FHC ratio. α3D is the angle between the normals to the fitted planar surfaces of A and I (Figure 5.1) and FHC3D is the ratio of the femoral head portion medial to the plane of I (Figure 5.1). Manually segmenting the femoral head, I and A are difficult, so we provide a fully automatic method that approximates the femoral head, I and A, and then uses the approximated structures to estimate α3D and FHC3D. We validate our method on US data collected from 40 hips from 25 infants.

5.2 Defining 3D Hip Morphology-derived Dysplasia Metrics

In a 3D B-mode US image of an infant’s hip, U(x,y,z), we define α3D as the angle between the planar approximations of A and I (section 5.3, Figure 5.1).

We define FHC3D as the ratio of the volume of the femoral head portion medial to I to the total volume of the femoral head (section 5.4, Figure 5.1).

5.3 Extracting α3D

Figure 5.2 presents a block diagram outlining the extraction of A, I and α3D. To compute α3D, we first identified U(x,y,zA,I) (subsection 5.3.1), where zA,I represents all the coronal plane slices containing A and I. Next, we identified the bone boundaries in U(x,y,zA,I) (subsection 5.3.2), and then we finally used geometrical priors to find the planar approximations of I and A (subsection 5.3.3). Once we had approximated the planes for I and A, we extracted α3D (subsection 5.3.4).
Figure 5.1: (a) A rendering of the anatomy of a hip joint showing $A$ (red), $I$ (blue) and the femoral head. (b) A schematic illustration of $\alpha_{3D}$ - the angle between the normals to the fitted planar surfaces approximating $A$ and $I$. (c) A schematic illustration of $FHC_{3D}$, which is defined as the ratio of the volume of the femoral head portion that is medial to $I$ to the total volume of the femoral head.

Figure 5.2: Block diagram showing the extraction of $I$, $A$, femoral head, $\alpha_{3D}$ and $FHC_{3D}$. 
5.3.1 Localizing I and A Across the Coronal Plane

Figure 5.3 shows a block diagram outlining the steps involved in identifying coronal plane slices that contain I and A. To identify these slices, we began by training a random forest classifier, $R_1$, to distinguish the US slices containing I and A from the other US slices that do not (all the slices being labeled manually by a graduate student based on the presence of I and A in those slices, Figure 5.4). This classifier $R_1$ (70 decision trees, minimum number of observations per tree set equal to 1, no pruning applied as similar to Breiman’s implementation [15]), was trained using the HOG [29] and LBP [105] features. HOG and LBP were employed since their use in conjunction can significantly improve object detection performance as compared to using only one of those features in object detection [162].

In a test US volume, the likelihood of each of the coronal plane slices ($i$) to contain I and A is evaluated using the trained classifier, $R_1$. The likelihood scores of $i$ are median filtered (resulting in $p_1(i)$) and normalized with the unity-based normalization, $p_{1,n} = \text{norm}(p_1)$. We defined the center slice as the median of all the slices that satisfy the criteria, $p_{1,n} > 0.5$. Around this center slice, we cropped out all slices that are greater than $s_d$ mm away from it. We call this $s_d$ parameter the span and discuss it in more detail in results section. Finally, in this cropped US volume, we defined our localized region across the coronal plane $(U(x,y,z_{A,I}))$ as originating from the most anterior position having $p_{1,n}(i) > 0.5$ and ending at the most posterior position having $p_{1,n}(i) > 0.5$.

5.3.2 Extracting 3D Bone Boundary

We employed the bone segmentation technique described in subsection 3.3.2 to extract bone boundaries in $U(x,y,z)$ and identify bone boundaries in the region $(U(x,y,z_{A,I}))$ (Figure 5.5). Here, we use a rotation-invariant local symmetry feature, structured phase symmetry (SPS) (subsection 3.3.2), which extracts sheet-like hyperechoic responses in a 3D US image, SPS. After segmenting these sheet-like hyperechoic responses in the 3D US image, which includes bone boundaries, cartilage boundaries and soft-tissue interfaces, we extracted bone boundaries, $B$, using the attenuation-based post-processing step described in subsection 3.2.2 (Figure 5.5).
5.3.3 Approximating Planes of I and A

Once we had extracted the bone boundary $B$ in $U(x, y, z_{A,I})$, we used geometric priors to find planes that approximate $I$ and $A$. Since $I$ and $A$ are bone boundaries, we did not expect to find any other bone boundary beyond these structures in US infant hip images. So, we selected the bottom-most non-zero $B$ point along each scan line for further processing to locate $I$ and $A$ (note that the lowest non-zero $B$ value is 0.5).

Since $I$ is approximately perpendicular to the US beam in coronal scans of infant hips, we approximated $I$ by fitting a plane that is within $15^\circ$ of the $x,y$ plane. To fit planes onto bone boundary points, we used a M-estimator SAmple Consensus (MSAC) algorithm [156]: this iteratively identifies the parameters of a plane
Figure 5.4: (a), (b) and (c) show example US slices within an US volume that contain the $I$ and $A$ structures, whereas (d), (e) and (f) show example US slices in the same US volume that does not contain at least one of $I$ and $A$ structures. (g) Plot showing the slice numbers in the US volume that contain $I$ and $A$. 
from the bone boundary points by identifying the inliers (points close to the fitted model during iterations), outliers (points far away from the fitted model during iterations) and a likelihood measurement of the amount of consensus extant between the inliers and plane model.

To identify $A$, we used a prior based on the fact that $A$ is medial to $I$ (or downwards in US images collected in the coronal plane). More specifically, we approximated the plane of $A$ as fitting onto bone boundary points that were medial to $I$ and oriented medially away from the plane of $I$.

### 5.3.4 Calculating $\alpha_{3D}$

We calculated $\alpha_{3D} = \cos^{-1}(n_A \cdot n_I) / (||n_A|| \cdot ||n_I||)$, where $n_A$ and $n_I$ are the unit normal vectors of $A$ and $I$.

### 5.4 Extracting $FHC_{3D}$

To estimate $FHC_{3D}$, we first extracted a voxel-wise probability map, $P(x,y,z)$, characterizing the likelihood of a voxel belonging to the femoral head (subsection 5.4.1). Next, we used $P(x,y,z)$ and $I$ to calculate $FHC_{3D}$ (subsection 5.4.3).
5.4.1 Femoral Head Segmentation

The femoral head in infant hips is unossified and appears hypoechoic in US images (Figure 5.6(a)). It is surrounded by anatomical structures with distinctive sonographic properties, e.g. the ilium which has a high sonographic response at its boundary and a shadow region beneath it, the labrum, the triradiate cartilage, the greater trochanter, etc. A cross-section of $U$, $C(d, \theta, \phi)$, ($d$ is the shortest distance of $C$ from the origin of $(x, y, z)$, and $\theta$, $\phi$ are rotations about $x$ and $y$ axis, respectively, of a reference plane defined by $z = 0$ to a plane parallel to $C$, Figure 5.6(b)) that intersects the femoral head is expected to include a hypoechoic region surrounded by cross-sections of the neighbouring anatomical structures. In order to segment a femoral head, we therefore started by training a random forest classifier, $R_2$, to distinguish slices that intersect the femoral head from those that do not. The classifier, $R_2$ (70 decision trees, minimum number of observations per tree set equal to 1, no pruning applied similar to Breiman’s implementation [15]), is trained using HOG and LBP features extracted from samples of cross-sections from US volumes. In a test US volume, the likelihood of $C$ intersecting the femoral head is evaluated using the trained classifier, $R_2$. We encoded this measurement throughout the coordinate space of $C$, and back-projected or interpolated it onto the $(x, y, z)$ coordinates, $L(x, y, z)$ (Figure 5.6(c)). For $N$ cross-sections sampled within the test US volume, we constructed a slice-based or tomographic voxel-wise probability map as being: $p_2 = \text{norm}(\sum_{N} L_N)$ (Figure 5.6(d)), where $\text{norm}(.)$ is the unity-based normalization.

Although $p_2$ alone does not provide an accurate segmentation of the femoral head (Figure 5.6(d)), we hypothesized that we could reliably segment the femoral head when we used $p_2$ with other voxel-wise features.

5.4.2 Enhancing Femoral Head Segmentation

To further enhance the slice-based voxel-wise probability map $p_2$ of the femoral head, we combined this map $p_2$ with twenty-four voxel-wise features (Figure 5.7). The complete list of features used are: (a) depth of voxel, (b) location of voxel along the inferior-superior axis, (c) $p_2$, (d) $U$, (e) range filtered (3x3 kernel size) $U$, (f) range filtered (5x5 kernel size) $U$, (g) entropy filtered (3x3 kernel size) $U$, (h)
entropy filtered (5x5 kernel size) $U$, (i) confidence map, $M$ \cite{84}, (j) range filtered (3x3 kernel size) $M$, (k) range filtered (5x5 kernel size) $M$, (l) entropy filtered (3x3 kernel size) $M$, (m) entropy filtered (5x5 kernel size) $M$, (n) nine HOG features, (o) convolution of $U$ with a sphere of radius 5mm, (p) distance from the approximated ilium plane, and (q) distance from the approximated acetabulum plane \cite{Fig5.7}. We feed these features into another RF classifier, $R_3$, (25 decision trees, minimum number of observations per tree equal to 1, no pruning applied), to estimate a probabilistic score, $P(x,y,z)$, of each voxel belonging to the femoral head \cite{Fig5.8}. In subsequent steps, we used both $P$ and the center of the femoral head, which we calculated as: $c = [c_x, c_y, c_z] = \sum_X (P \times X) / \sum P$, where $X = [x, y, z]$.

### 5.4.3 Estimating $FHC_{3D}$

Finally, we estimated $FHC_{3D}$ as $FHC_{3D} = \sum_X (P_M / \sum_X P)$, where $X = [x, y, z]$ and $P_M$ represents all $P$ values that are medial to $I$.

### 5.5 Experiment, Results and Discussion

**Parameter Specification:** For our RF classifiers, we selected the number of trees based on out-of-bag classification error analysis - we identified the number of trees beyond which the out-of-bag classification error does not seem to decrease considerably. The number of trees for $R_1$, $R_2$ and $R_3$ was 70, 70 and 25, respectively \cite{Fig5.9} \cite{Fig5.11} \cite{Fig5.11}. The trees were not pruned and the number of features to select at random for each decision split was selected as the square root of the number of features [10, 15].

The span parameter $s_d$ in subsection 5.3.1 is an empirical parameter describing the distance between the furthest slice containing $I$ and $A$ and the middle slice that contains $I$ and $A$. We empirically set $s_d = 3.5 \text{mm}$ based on the highest value of $s_d$ we could find in our data \cite{Fig5.12}.

**Data** We acquired 3D and 2D B-mode US images from 40 infant hips (25 infants) - all obtained as part of routine clinical care under appropriate institutional review board approval (UBC CREB number: H14–01448). All 2D and 3D scans
Figure 5.6: Overview of our slice-based probability map extraction. (a) Overlay of example US volume and a manually segmented femoral head. (b) Number of \( C \) were evaluated using classifier \( R_2 \) to determine their likelihood of intersecting the femoral head. (c) Back-projected likelihood scores, \( L \), for each of the cross-sections \( C \). (d) Back-projected responses were summed and normalized to construct voxel-wise probability map, with an overlay of the manually segmented femoral head (yellow).

were collected in the coronal plane. The scans were performed using a SonixTouch Q+ machine with a 4DL14–5/38 Linear 4D transducer (BK ultrasound, Analogic corporation, MA, USA). The transducers center ultrasound frequency was kept at 7.5 MHz and image depth setting ranged between 3.8 cm – 5 cm. While performing a 3D US scan, the operator tried to capture the entire femoral head within the scanned US volume. A maximum of two surgeons each performed two 3D US scans and two 2D US scans. The repetition was designed to allow us to evaluate inter-operator repeatability. The surgeon who acquired each of the 2D US scans further chose a 2D US image from all the 2D US images collected during that ex-
Figure 5.7: Features used in the classifier $R_3$ to estimate the probabilistic score, $P(x,y,z)$, of each voxel belonging to the femoral head.
Figure 5.8: (a) Overlay of an example US volume and a manually segmented femoral head. (b) Overlay of the example US volume and its automatically extracted femoral head voxel-wise probability map $P$.

Figure 5.9: Out-of-bag classification error for various numbers of grown trees used by the $R_1$ classifier. We selected 70 as being our number of trees for $R_1$. 

112
Figure 5.10: Out-of-bag classification error for various numbers of grown trees used by the $R_2$ classifier. We selected 70 as being our number of trees for $R_2$.

Figure 5.11: Out-of-bag classification error for various numbers of grown trees used by the $R_3$ classifier. We selected 25 as being our number of trees for $R_3$. 
amination and measured $\alpha_{2D}$ and $FHC_{2D}$. If the surgeon did not find any adequate images in a 2D US imageset belonging to a hip examination, that hip examination data was labeled as an inadequate imageset. Furthermore, we noted the acquisition time for acquiring both 2D US and 3D US scans in all the infants.

**Validation Scheme** To avoid overfitting while using classifiers $R_1$, $R_2$ and $R_3$, we avoided training and testing on the same patient. We randomly split the data into two halves - 13 patients in the first half and 12 patients in the second half. When we trained on the first half of the data, we tested on the second, and vice versa. Since $R_2$ is an independent variable while $R_3$ is dependent (i.e., output of $R_2$ is fed into $R_3$), we further split the data to avoid any over-fitting while training $R_3$.

To investigate the reliability of using 3D US in diagnosing DDH, we first investigated the reliability in acquiring adequate 3D US volumes. A graduate student labeled adequacy of 3D US volumes based on the criteria that an US volume is adequate only when it has the femoral head, ilium, acetabulum and labrum. We compared this with the reliability in manually acquiring adequate 2D US imagesets. We also compared acquisition times for both 2D and 3D US images by the same surgeons.
Table 5.1: Success rates in acquiring adequate 3D US images and 2D US imageset.

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequate 2D</td>
<td>37</td>
<td>21</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Adequate 3D</td>
<td>37</td>
<td>21</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

We reported both the inter-rater and intra-rater variability of 3D US based dysplasia metrics, \( \alpha_{3D} \) and \( FHC_{3D} \). We compared these variability measures with the inter-rater and intra-rater variability of their their 2D counterparts, \( \alpha_{2D} \) and \( FHC_{2D} \).

We estimated intra-rater variability by calculating the standard deviation within repeated measurements made on the same hip by the same rater. We estimated inter-rater variability between two raters by calculating the standard deviation between the mean dysplasia metric values measured by the two raters on the same hip.

Finally, we investigated the agreement between \( \alpha_{3D} \) and \( FHC_{3D} \) with their 2D counterparts \( \alpha_{2D} \) and \( FHC_{2D} \). Here, we extracted correlation between the 2D and 3D US-based measures collected by the same raters.

Acquiring 3D US Images: In total, 6 orthopaedic surgeons participated in acquiring both 3D and 2D US images. The success rate of each of the surgeons in acquiring adequate 3D US volumes (a volume is adequate if it has the ilium, acetabulum, labrum and femoral head in it) and 2D US imagesets (a 2D US is adequate if it has the ilium, acetabulum, labrum and femoral head in it) is summarized in Table 5.1.

There was no marked difference between the success rates for acquiring adequate 3D US vs. the success rates for acquiring adequate 2D US imagesets.

We did not find any significant difference between the acquisition time for 3D US images vs. acquisition time for 2D US image (23 seconds for 3D US, 25.4 seconds for 2D US, \( t_{3D} - t_{2D} = -2.4 \) seconds \( \pm 21.5 \) seconds, \( p = 0.4 \) [Figure 5.13, Figure 5.14]).
Figure 5.13: Box-plot showing the distribution in acquisition time for both 2D and 3D US for all the six participating surgeons.

Figure 5.14: Box-plot showing distribution of acquisition time of both 2D and 3D US. Mean acquisition time of 3D US is slightly less than that of 2D US, however the change was not statistically significant.
Results for extracting dysplasia metrics in all hips: Results for extracting 3D dysplasia metrics in all hips are shown in Appendix B.

Discrepancy between 2D US and 3D US-based Dysplasia Metrics: The discrepancy between $FHC_{2D}$ and $FHC_{3D}$ (i.e., $FHC_{2D} - FHC_{3D}$) was not statistically significant (mean 0.15%, SD = 11.1%, $p > 0.05$) Figure 5.15, however the discrepancy between $\alpha_{2D}$ and $\alpha_{3D}$ (i.e., $\alpha_{2D} - \alpha_{3D}$) was statistically significant (mean 2.88°, SD = 11.7°, $p < 0.05$) Figure 5.16. Agreement between the $FHC_{2D}$ and $FHC_{3D}$ of each hip examination was moderate (correlation coefficient, $r = 0.4$ (95% confidence interval 0.23 and 0.55) Figure 5.15). We found similar agreement between $\alpha_{2D}$ and $\alpha_{3D}$ (correlation coefficient, $r = 0.41$ (CI 0.24 and 0.56) Figure 5.16).

Appendix B shows other examples for automatic extractions of 3D US-based dysplasia metrics in a healthy hip, a borderline hip and a dysplastic hip – we found that our approximations of the ilium, labrum, acetabulum and femoral head and our estimation of the 3D US-based are realistic.

Variability of Dysplasia Metrics, $\sigma$: Figure 5.17 illustrates how multiple 3D US images acquired from the same hip appear to be more similar to each other compared to the similarity between multiple 2D US images collected from the same hip. Figure 5.17 is also an example where the variability of $FHC_{3D}$ and $\alpha_{3D}$ was much lower than its manual counterparts.

Box plots in Figure 5.18 and Figure 5.19 summarize variability of the dysplasia metrics over all hip examinations. We found that the 3D-based dysplasia metrics have significantly lower variability compared to their 2D counterparts - intra-rater variability of $FHC_{3D}$ was 63% lower than $FHC_{2D}$ (3.19% for $FHC_{3D}$ vs. 8.56% $FHC_{2D}$, $p < 0.01$), inter-rater variability of $FHC_{3D}$ was 65% lower than $FHC_{2D}$ (2.72% for $FHC_{3D}$ vs. 7.79% $FHC_{2D}$, $p < 0.01$), intra-rater variability of $\alpha_{3D}$ was 73% lower than $\alpha_{2D}$ (2.2° for $\alpha_{3D}$ vs. 8.33° $\alpha_{2D}$, $p < 0.01$), and inter-rater variability of $\alpha_{3D}$ was 78% lower than $\alpha_{2D}$ (2.35° for $\alpha_{3D}$ vs. 10.63° for $\alpha_{2D}$, $p < 0.01$). Also, comparing with ACA [94] which is the only other 3D metric available, the intra-rater variability of $\alpha_{3D}$ is around 45% lower than the variability of ACA (2.2° for $\alpha_{3D}$ vs. 4.1° for ACA) [94].
Figure 5.15: Scatter plot of $FHC_{2D}$ vs. $FHC_{3D}$. The red line is the equality line. Blue data points correspond to lower $FHC_{3D}$ and yellow data points correspond to higher $FHC_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal.
Figure 5.16: Scatter plot of $\alpha_{2D}$ vs. $\alpha_{3D}$. Blue data points correspond to lower $\alpha_{3D}$ and yellow data points correspond to higher $\alpha_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal.

We further evaluated whether there is any association in the variability of measuring the 3D US-based dysplasia metrics with the severity of hip dysplasia. We did not find any statistically significant association in the variability of dysplasia metrics with the severity of hip dysplasia for either $\alpha_{3D}$ or $FHC_{3D}$ ($p > 0.05$, Figure 5.20, Figure 5.21).
Figure 5.17: Qualitative results. (a), (b), (d) and (e) show example variability of $\alpha_{2D}$, $\alpha_{3D}$, $FHC_{2D}$ and $FHC_{3D}$ from two 2D and two 3D US images from a hip examination ($\alpha_{2D} = 47^\circ$ and $56^\circ$, $\alpha_{3D} = 45.1^\circ$ and $45.9^\circ$, $FHC_{2D} = 51\%$ and $71\%$, $FHC_{3D} = 46.6\%$ and $47.8\%$). The higher variability in the input 2D US images (and $\alpha_{2D}$ and $FHC_{2D}$ values) can be seen in the manually aligned 2D US images in (c) compared to the lower variability in the manually aligned 3D US images (and $\alpha_{3D}$ and $FHC_{3D}$ values) in (f).

Computational Considerations: It takes around 58 seconds to compute CSPS, the planes of ilium, acetabulum and $\alpha_{3D}$, and another 62 seconds to compute the probability map of the femoral head and to compute $FHC_{3D}$, when run on a Xeon(R) 3.40 GHz CPU computer with 12 GB RAM. All processes were executed using MATLAB 2017a (MATLAB 2017a, the Mathworks Inc., Natick, MA, USA).
Figure 5.18: Box-plot of the within-hip standard deviations among $FHC_{2D}$ vs. $FHC_{3D}$ values within all of the hip examinations. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The '+' points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in $FHC_{3D}$ compared to $FHC_{2D}$ ($p < 0.01$).

5.5.1 Discussion

**Acquiring 3D US Images:** Overall, the acquisition time for 3D US images was similar to those of 2D US images. Furthermore, the rate of 3D hip examinations where adequate 3D US images were missed was similar to the rate of 2D hip examinations where adequate 2D US images were missed. These suggests that there is no noticeable difference between the difficulty in acquiring 3D and 2D US images for neonatal hip examination.

**Variability of Metrics:** The variabilities in the automatic measurements using our proposed 3D US-based dysplasia metrics were significantly lower than those of their 2D US-based counterparts ($\approx 75\%$ reduction for $\alpha$ and $\approx 65\%$ reduction for
Figure 5.19: Box-plot of the within-hip standard deviations among $\alpha_{2D}$ vs. $\alpha_{3D}$ values within all of the hip examinations. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 1st and 3rd quartiles, respectively. The '+' points indicate outliers or data points that are outside the range of whiskers, where the whiskers correspond to 99.3% coverage if the data are normally distributed. The within-hip standard deviations are significantly less in $\alpha_{3D}$ compared to $\alpha_{2D}$ ($p < 0.01$).

These reductions in variability in 3D US-based dysplasia metrics suggest that probe position variation has a much larger effect on variability in the dysplasia metrics than manual processing of the 2D US (approximately 20% improvement with automatic image processing within a 2D US; see chapter 4).

Discrepancy between 2D US and 3D US-based Dysplasia Metrics: While the discrepancy between $FHC_{2D}$ and $FHC_{3D}$ was not statistically significant, there is a significant bias of $\alpha_{3D}$ to be larger than $\alpha_{2D}$ by 2.88°. This bias and the thresholds for $\alpha_{2D}$ to classify hips as normal or dysplastic (i.e., $\alpha_{2D} < 43°$: dysplastic, $\alpha_{3D} > 60°$: healthy), can potentially be used for setting thresholds for $\alpha_{3D}$ to classify healthy and dysplastic hips (i.e., $\alpha_{3D} < 45.88°$: dysplastic, $\alpha_{3D} > 62.88°$: healthy). However, the validity of such thresholds is limited since the thresholds used for
Figure 5.20: Scatter plot of variability of $FHC_{3D}$ vs. mean $FHC_{3D}$. Blue data points correspond to lower mean values of $FHC_{3D}$ and yellow data points correspond to higher mean values of $FHC_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $FHC_{3D}$ with the severity of hip dysplasia.
Figure 5.21: Scatter plot of variability of $\alpha_{3D}$ vs. mean $\alpha_{3D}$. Blue data points correspond to lower mean values of $\alpha_{3D}$ and yellow data points correspond to higher mean values of $\alpha_{3D}$, so blue correspond to hips that would be judged as dysplastic hips whereas yellow correspond to hips that would be judged as normal. The red line is the best fit line. Correlation and p values suggest that there is no significant association in the variability of $\alpha_{3D}$ with the severity of hip dysplasia.
\( \alpha_{2D} \) are debatable (e.g. \( \alpha_{2D} > 60^\circ \): healthy in one study \cite{79}, whereas \( \alpha_{2D} > 55^\circ \): healthy in another study \cite{82}).

**Computational Considerations:** The complete process of extracting \( FHC_{3D} \) and \( \alpha_{3D} \) from an US volume took approximately 170 seconds, when run on a Xeon(R) 3.40 GHz CPU computer with 8 GB RAM. All processes were executed using MATLAB 2017b. Current practice has a sonographer process the images post-acquisition, so this computation time is not a critical barrier to implementation. Although not critical for clinical use, real-time feedback during acquisition on image adequacy could help prevent incidents in which a sonographer fails to acquire an adequate 3D US volume. While our focus thus far was on automatically extracting 3D US-based dysplasia metrics, we have earlier developed a fast 2D image adequacy classifier (chapter 4) that we hope in future to extend to 3D. We believe that adequacy classification could be performed in real-time and provided as feedback during acquisition.

### 5.6 Summary

In this chapter, we presented novel 3D hip morphology-based dysplasia metrics, and we provided automatic methods for extracting those dysplasia metrics from 3D US B-mode images. To the best of our knowledge, this is currently the only method that automatically characterizes 3D hip morphology from 3D US images. We showed that these dysplasia metrics are significantly less variable than their 2D counterparts (around 65% reduced variability for \( FHC_{3D} \) and around 75% reduced variability for \( \alpha_{3D} \)). This suggests that this 3D morphology-derived dysplasia metric could be valuable in improving the reliability in diagnosing DDH, which may lead to a more standardized DDH assessment with better diagnostic accuracy than the current 2D assessment.

Notably, the improvement in reliability associated with the 3D scans was achieved by orthopaedic surgeons, who have limited training in performing US examinations in comparison to a radiologist. This strongly suggests that we may, in future, be able to train personnel other than radiologists to obtain reliable and reproducible dysplasia metrics using 3D ultrasound machines, potentially reducing the costs as-
associated with screening for DDH.
Chapter 6

Conclusions

DDH is an important clinical problem, impacting infants and families worldwide [7, 55, 147]. As DDH is the most common pediatric hip condition and can cause considerable long-term debilitation if left untreated, screening and detection efforts have resulted in universal clinical examinations of all newborns [147]. Consequently, every child born has the potential to be affected by both limitations to and advances in DDH diagnosis.

6.1 Thesis Contributions

This thesis contributes towards providing a comprehensive analysis of all the currently available dysplasia metrics. We demonstrate the relative usefulness of each of these dysplasia metrics in terms of their reproducibility in DDH diagnoses.

This thesis further contributes towards inventing automatic and reliable methods for diagnosing DDH using both 2D and 3D US imaging. We demonstrate the advantages and limitations of our methods in comparison to the currently used standard US imaging-based approach for diagnosing DDH.

The main contributions of this thesis (Figure 6.1) are summarized as follows.
6.1.1 A Systematic Review and Meta-analysis of Variability in Dysplasia Metrics

In chapter 2, we conducted a systematic review and meta-analysis to identify the most reliable dysplasia metric for diagnosing DDH. Our findings suggested that the $\alpha$ angle extracted from 2D US images of neonatal hips is currently the most reliable metric. However, we also determined that even $\alpha$, with the best repeatability amongst the assessed metrics, can be problematic and can lead to high rates of misdiagnosis. We also found that all the dysplasia metrics rely on first approximating the bone and cartilage boundaries from US images. A list of our contributions in chapter 2 are:

- We performed a systematic review and meta-analysis of variability in various dysplasia metrics used in US-based DDH diagnoses [P1].
- We provided systematic review-based estimates (considered better than individual studies [75]) for variability estimates in measuring dysplasia metrics. We demonstrated that the variability in these dysplasia metrics are large [P1].
- We also showed that, among all the currently available dysplasia metrics, Graf’s $\alpha$ angle measurements are most reproducible [P1].
- We demonstrated that the reproducibility of dysplasia metrics have not improved since their advent in 1983, despite decades of advancements on US imaging technologies [P1].

6.1.2 Symmetry and Attenuation-based US Bone Imaging

To reliably extract dysplasia metrics, our first work in chapter 3 focused on an automatic and robust technique to extract bone and cartilage boundaries. We proposed new attenuation-based features and presented a method for combining them with local symmetry features to result in a bone segmentation that was significantly more robust compared with using only local symmetry features. We also presented a rotation-invariant local symmetry measure, SPS, which we later used along with our bone segmentation method for extracting dysplasia metrics. In an experiment
including US data from 15 infant hip exams, we found that our proposed bone segmentation agreed well with manually-labelled bone boundaries (mean discrepancy between segmentations around 1mm). A list of our contributions in chapter 3 are:

- We proposed novel attenuation-based features that are useful for extracting bone boundaries in US images. We showed that these features can be combined with local symmetry features towards extracting bone boundaries in US images [P2 [121], P3 [122]].

- We demonstrated that the combination of attenuation and local symmetry features result in a more robust segmentation (i.e., more robust to soft-tissue outliers and US artifacts) than using a local symmetry feature alone [57]. We demonstrated this in terms of surface fitting error when comparing with CT-derived ground truth (∼0.24mm with combined features-based method vs. ∼0.5mm with local symmetry-based method [57], p < 0.01) [P2 [121], P3 [122]].

- We presented a second bone segmentation method that employs isotropic filters and is independent of the orientation of the bone boundaries [P4 [124], P7 [129]].

- We extended the orientation-independent bone segmentation method for use in 3D US [P5 [125]].

- We demonstrated that this orientation-independent bone segmentation method has lower discrepancy with manually-labelled bone boundaries compared to a previous state-of-the-art [59] (mean discrepancy: around 1 mm for our method vs. around 3 mm for method in [59]) [P6 [128]].

6.1.3 Characterizing Hip Joint using 2D US

In chapter 4, we developed a machine learning-based method for identifying whether a 2D US image is adequate for dysplasia metric measurements. We also presented methods for automatically extracting the three commonly used dysplasia metrics [104], α, β and FHC. We demonstrated on data from 69 patients that automatic
extraction of these dysplasia metrics reduces variability by around 18% to 21% as compared to that of their manual extraction counterparts. We also demonstrated a high agreement between manual and automatic Graf classifications (84% agreement). In nine cases of disagreement between the manual approach and our automatic method in the Graf-DDH type assignment, process-of-care supported the classification decisions made by our proposed automatic method. Our classification tends to classify more patients as having DDH than would be identified as such by the current manual method. This tendency of heightened suspicion may perhaps be acceptable from a clinical perspective since we know that a considerable percentage of patients who initially had a diagnosis of DDH that subsequently resolved nonetheless had radiological signs of DDH a few months later (e.g., around 33% of such patients in Sarkissian’s study [144]).

A list of our contributions in chapter 4 are:

- We have developed a novel automatic method that identifies the US images suitable for making measurements [P4 [124], P7 [129]].

- We demonstrated that this identifier is in excellent agreement with identifications conducted by experienced clinicians (the area under the receiver operating characteristics curve > 98%) [P7 [129]].

- We have developed a novel automatic method that computes dysplasia metrics from US images of the neonatal hip [P4 [124], P7 [129], P8 [123], P9 [126]].

- We demonstrated that this automatic computation can improve the reproducibility of dysplasia metrics when compared with the dysplasia metrics that are manually extracted by expert clinicians (p < 0.01, an approximately 18% to 21% improvement) [P7 [129]].

- We demonstrated a high agreement between manual and automatic Graf classifications (84% agreement) [P7 [129]].

### 6.1.4 A 3D US-based Hip Dysplasia Assessment

In chapter 5 we addressed perhaps the most crucial problem in diagnosing DDH with US imaging: the variability of dysplasia metrics and in turn diagnosis and
treatment decisions is very sensitive to the orientation of the US transducer during acquisition [79]. To reduce sensitivity to transducer orientation, we proposed (like others [64, 94]) to use 3D US. Here, we investigated new dysplasia metrics that could be extracted from 3D US and that characterize 3D hip morphology in infants. We presented an automatic method for extracting these dysplasia metrics, which included a novel tomographic-reconstruction based segmentation of the femoral head. We demonstrated on data from 25 patients that these 3D hip morphology-derived dysplasia metrics provide a significant reduction in variability as compared to 2D dysplasia metrics: an approximately 75% reduction for $\alpha^{3D}$ and an approximately 65% reduction for $FHC^{3D}$). A list of our contributions in chapter 5 are:

- We have proposed two novel 3D hip morphology-derived dysplasia metrics: $\alpha^{3D}$ and $FHC^{3D}$ [P5 [125], P6 [128], P10 [127]].
- We have developed automatic image processing pipelines to compute each of these dysplasia metrics [P5 [125], P6 [128], P10 [127]].
- We demonstrated that both $\alpha^{3D}$ and $FHC^{3D}$ provide significantly more consistent diagnoses than their 2D counterparts: they are significantly less variable ($p < 0.01$ and approximately 75% reduction in variability for $\alpha^{3D}$, $p < 0.01$ and approximately 65% reduction in variability for $FHC^{3D}$) [P5 [125], P6 [128], P10 [127]].

6.1.5 Summary of Technical Contributions

We conclude with a list of our key technical contributions in this thesis:

- We developed an intensity-invariant and rotation-invariant local symmetry feature, the SPS. We demonstrated in our experiments that the SPS feature is effective in identifying bone boundaries in US images of an infant hip [125, 127, 129].
- We combined local symmetry feature with attenuation-related features [120], and demonstrated that the combination of attenuation and local symmetry
features result in a more robust segmentation (i.e., more robust to soft-tissue outliers and US artifacts) than using a local symmetry feature alone [57].

- We developed the concept of image adequacy in the context of diagnosing DDH using US imaging [129]. An inadequate US image can be due to the acquisition of the US image from an improper transducer position or due to a sub-standard quality of the US image, both of which can result in an absence of key anatomical landmarks in the image. By using an adequacy classifier before computing our dysplasia metrics, we can avoid erroneous
measurements of dysplasia metrics in inadequate images that do not contain all the anatomical landmarks necessary for making a dysplasia metric measurement. We implemented an automatic 2D US image adequacy classifier that has high specificity and sensitivity (area under the ROC curve around 99.5%).

- We proposed a fully automatic method for extracting 2D dysplasia metrics \cite{124,129}. We demonstrated that the automatically extracted 2D dysplasia metrics using our method had a significantly lower intra-exam variability in repeated diagnosis compared to a manual 2D diagnosis (approximately 20% reduction in variability, \( p < 0.01 \)) \cite{129}.

- We developed a multi-oriented slice-based learning approach (i.e., a random forests based learning for identifying presence of an object in a slice, followed by a 3D reconstruction of the probability map of the object based on the random forests outputs and the corresponding locations of all the slices at different orientations, \textsection 5.4.1) for segmenting the femoral head in 3D US. We found that this approach can be reliably used for segmenting the femoral head and for extracting the \( FHC_{3D} \) (inter-exam variability of \( FHC_{3D} \) reduced by around 65% compared to variability of manual \( FHC_{2D} \), \( p < 0.01 \)) \cite{127}.

- We developed a fully automatic method for extracting 3D dysplasia metrics. To the best of our knowledge, our method currently is the only automatic method for extracting 3D dysplasia metrics. We demonstrated that the 3D dysplasia metrics extracted using our method had a significantly lower inter-exam variability in repeated diagnosis compared to the current standard 2D diagnosis (approximately 70% reduction in variability, \( p < 0.01 \)) \cite{125,127,128}.

6.2 Future Work

In this thesis, we developed automatic methods that markedly improves repeatability of DDH diagnosis. Our contributions opened up new research questions and some potential future works in US-based DDH diagnosis:
**3D US Image Adequacy:** In our study, we found that in approximately 9% of hip examinations, our surgeons were unable to obtain adequate 3D US scans. This points to a clinical feasibility issue in our method: the run-time for extracting dysplasia metrics takes around three minutes, so we currently cannot provide any near real-time feedback to assure surgeons that they have obtained an adequate scan. To address this issue, in the future we plan on using a machine learning approach (similar to our 2D image adequacy work in [chapter 4]) to provide near real-time feedback to users on whether any adequate 3D images have been recorded. Also, Paserin et al. [111] recently proposed a slice-based CNN classifier for identifying adequate 3D volumes. They demonstrated a perfect 100% classification accuracy on 3D US volumes collected from 15 infant hips, and reported a processing speed of around 2 seconds per volume. Considering that it takes around 2.5 seconds for our 3D transducer to acquire one volume, an approximately 2 seconds processing time for identifying adequacy can be considered as near real-time. In future, we may therefore add a CNN-based adequacy classifier to our Sonix Q+ US machine that will provide active feedback to an user on whether any of the collected US volumes were adequate for extracting 3D dysplasia metrics.

**2D US-based DDH Diagnosis:** The impact of improving DDH diagnosis with a 2D US-based system is perhaps considerably more than that with a 3D US-based system, particularly because most of the centers around the world currently have access only to 2D US systems. A limitation of a 2D US-based diagnosis is that dysplasia metric values measured from 2D US images can vary considerably based on orientations and placements of the US transducer during image acquisition. To address this issue, we plan on investigating the reliability of DDH diagnosis using a 2D US based pseudo-3D technique. More specifically, we plan on acquiring 3D US by first measuring the trajectory of our 2D US transducer using a tracking unit as a clinician moves it across the hip joint, and then reconstructing the 3D US based on the acquired 2D US images and the tracked positions of the 2D US images, and finally extract dysplasia metrics from the reconstructed 3D US volume. We anticipate that we will be able to further improve the reliability of our 2D US system by using this pseudo-3D technique.
Dynamic Assessment of DDH: Recently, Paserin et al. [110] developed the first quantifiable 3D dynamic assessment of DDH. Paserin et al. [110] used our $FHC_{3D}$ to get a measure of whether an infant hip is dislocated or not. In future, we may further investigate the application of our proposed 3D dysplasia metrics for use in such a quantifiable 3D dynamic assessment.

Tracing the Evolution of DDH: To understand how DDH evolves in an infant hip, we can use our reliable 3D US-based diagnosis tool to perform a prospective follow-up study on infant hips as those hips develop. We can also do a prospective study to identify reliable thresholds for 3D US-based dysplasia metrics for normal, abnormal and borderline hips. With this prospective study, we hope that we will be able to identify criteria to distinguish between healthy and abnormal hips that we can later provide as guidelines to clinicians using our proposed system.
Bibliography


[160] Y. Tumer, W. T. Ward, and J. Grudziak. Medial open reduction in the
treatment of developmental dislocation of the hip. *Journal of Pediatric


occlusion handling. In *Computer Vision, 2009 IEEE 12th International

[163] T. Woodacre, A. Dhadwal, T. Ball, C. Edwards, and P. Cox. The costs of
late detection of developmental dysplasia of the hip. *Journal of Children’s

solution for automatic fetal neurosonographic diagnostic plane verification
using clinical standard constraints. *Ultrasound in Medicine & Biology*,
2017.

follow-up of congenital subluxation of the hip. *Journal of Pediatric

[166] L. Zhang, S. Chen, C. T. Chin, T. Wang, and S. Li. Intelligent scanning:
Automated standard plane selection and biometric measurement of early

Appendix A

Examples of Automatic Dysplasia Metric Extraction from 2D US Images

The following figures show examples for automatic extractions of 2D US-based dysplasia metrics in a healthy hip, a borderline hip and a dysplastic hip – we found that our approximations of the ilium, labrum, acetabulum and femoral head and our estimation of the 2D US–based dysplasia metrics are realistic.
Figure A.1: Visualization for the extracted dysplasia metrics from a 2D US image in a healthy hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 70^\circ$, $\beta = 29^\circ$ and $FHC = 73.2\%$.

Figure A.2: Visualization for the extracted dysplasia metrics from a 2D US image in a borderline hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 47^\circ$, $\beta = 40^\circ$ and $FHC = 59.8\%$. 
Figure A.3: Visualization for the extracted dysplasia metrics from a 2D US image in a dysplastic hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 38^\circ$, $\beta = 56^\circ$ and $FHC = 41\%$. 
Appendix B

Examples of Automatic Dysplasia Metric Extraction from 3D US Images

The following figures show examples for automatic extractions of 3D US-based dysplasia metrics in a healthy hip, a borderline hip and a dysplastic hip – we found that our approximations of the ilium, acetabulum and femoral head and our estimation of the 3D US-based dysplasia metrics are realistic.
Figure B.1: Visualization for the extracted dysplasia metrics from a 3D US image in a healthy hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 71.8^\circ$ and $FHC = 72.4\%$. The visualization in here was done automatically using 2017a (MATLAB 2017a, the Mathworks Inc., Natick, MA, USA). The 3D plot generated using MATLAB was later manually rotated in 3D to make the planes of the ilium and acetabulum more apparent. In Chapter 5, the visualization was done manually using AMIRA software (TGS, San Diego, USA).
Figure B.2: Visualization for the extracted dysplasia metrics from a 3D US image in a borderline hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 53.1^\circ$ and $FHC = 47.9\%$. The visualization in here was done automatically using 2017a (MATLAB 2017a, the Mathworks Inc., Natick, MA, USA). The 3D plot generated using MATLAB was later manually rotated in 3D to make the planes of the ilium and acetabulum more apparent.
Figure B.3: Visualization for the extracted dysplasia metrics from a 3D US image in a dysplastic hip. In this example, the values of automatically extracted dysplasia metrics are: $\alpha = 35.6^\circ$ and $FHC = 22.3\%$. 