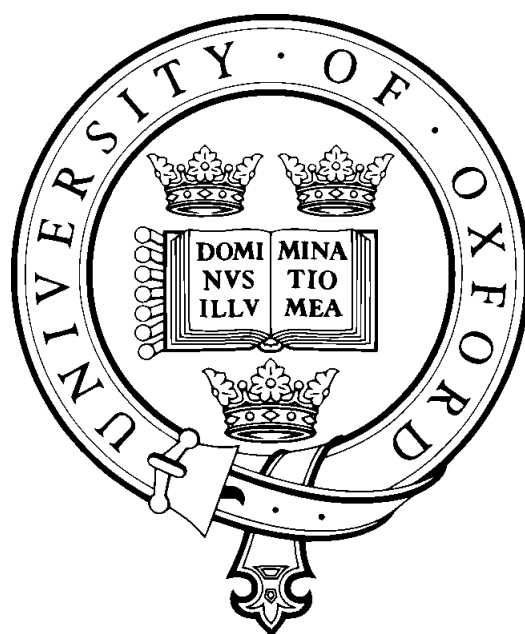


Towards Engineering Ultrasound Image Analysis Solutions For Resource Constrained Environments



Mohammad Ali Maraci

St Hilda's College

University of Oxford

Supervised by

Professor J Alison Noble

Submitted: Michaelmas Term,

January 3, 2016

This thesis is submitted to the Department of Engineering Science,
University of Oxford, in partial fulfilment of the requirements for the
degree of Doctor of Philosophy

Abstract

Obstetric ultrasound has proven an integral part of prenatal care for many applications including detecting pregnancy risk factors and fetal abnormalities. Although ultrasound imaging is widely used in developed countries, its availability in limited-resource settings has been dependent on device portability, associated costs, and trained healthcare workers. In recent years, as ultrasound equipment has become cheaper and portable its availability in limited-resource settings has expanded with promising trends. Yet a major constraint of ultrasound imaging is the steep learning curve to reach scanning proficiency, rendering the process challenging for inexperienced users.

This thesis presents methods towards addressing the fundamental task of empowering novice sonographers to perform targeted diagnostic screening to the same level as experienced ultrasonographers. This is of particular relevance to low-resource settings where pregnant women may not be able to attend specialist obstetric care centres during their pregnancy. We address this task by introducing simplified scanning protocols and using computer vision and machine learning techniques. While aimed at low-income country healthcare needs, the research is also equally applicable to needs of developed world.

Firstly, an original framework for automatic detection of frames of interest in ultrasound videos is presented. Furthermore, the merits of pre-processing images on classification accuracy, on datasets of variable size, is investigated. To this end, a classification framework for 2D ultrasound frames based on the improved Fisher vector encoding of SIFT and LP-SIFT features has been introduced. Additionally a Structured Random Forests edge detection framework is introduced for an accurate structure representation in 2D ultrasound data.

Secondly, a novel framework based on video dynamics and motion patterns is presented for automatic detection of sequences of interest in real world

ultrasound videos. To illustrate clinical applicability, this approach is applied to automatically identify the fetal presentation in the womb, during the third trimester of pregnancy from a predefined freehand ultrasound video sweep. A kernel dynamical textures model is utilised to model the dynamics of video sub-sequences that contain a fetal skull and those sequences that do not. A similarity kernel is then constructed between the model parameters to train a classifier. A classification accuracy of 93% was achieved using this framework.

Thirdly, for the first time, the real world application of automatic analysis of predefined ultrasound videos to detect the fetal presentation and heartbeat is considered. For this application, the frames of interest are initially identified using a frame-by-frame classification framework. Moreover, a linear-chain structured graphical model is used to take into account the temporal information in the 2D predefined sweeps and to regularise the classification results. Consecutive frames of the fetal heart are further investigated for the presence of the heartbeat activity. Classification accuracy achieved using our method was about 88%.

Finally, potential future directions are investigated and initial findings have been reported. These include results on a small pilot study specifically designed to perform a statistical comparison between images obtained from a standard mid-range ultrasound probe and a low-end portable ultrasound probe. Moreover, the efficacy of using a deep convolutional neural network with transfer learning is investigated and initial findings are reported.

Acknowledgements

I am extremely grateful to my supervisor, Professor Alison Noble, whose constant guidance, enthusiasm and patience enabled the completion of this thesis. A big thank you to the BioMedia members for making the lab an excellent environment. Finally, all my appreciation to my wife and family for the many years of unconditional support.

This thesis would have not been possible without the help of our clinical collaborators Dr. Aris Papageorgiou, Dr. Raffaele Napolitano, Dr. Tess Norris, Dr. Sikolia Wanyonyi, Dr. Malid Molloholli, Dr. Christina Aye, and Miss Fenella Roseman, and others at the Nuffield Department of Obstetrics and Gynaecology, John Radcliffe Hospital, who kindly helped during data acquisition. I would also like to thank the Centre for Doctoral Training in Healthcare Innovation and my funding body RCUK Digital Economy Programme grant number EP/G036861/1 for their support.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	4
1.3	Thesis Outline	6
1.4	Notation	6
2	Literature Review	7
2.1	Introduction	7
2.2	Ultrasound Imaging	10
2.3	Ultrasound Imaging For Performing Standard Measurements In Obstetrics	12
2.3.1	Automated Biometry Measurements	14
2.3.2	Current Challenges and Limitations	17
2.4	Ultrasound in Limited-Resource Settings	20
2.4.1	Manual Detection of Fetal Presentation and Heartbeat	23
2.4.2	Emerging Technologies for Limited-Resource Settings	24
2.5	Automated Methods for Fetal Ultrasound Imaging	28
2.5.1	Machine Learning in Medical Imaging	29
2.5.2	Feature Descriptors	30
2.5.3	Learning Mechanisms	34
2.6	Conclusion	39
3	Ultrasound Multi-Class Classification	40
3.1	Introduction	41
3.2	Ultrasound Image Dataset	42
3.3	Ultrasound Structure Representation	43
3.3.1	Contrast Invariant Representation	44
3.3.2	Structured Random Forests (SRF) Edge Maps	49
3.4	Features	54
3.4.1	LP-SIFT Features	55

3.5	Feature Encoding	57
3.5.1	Bag of Visual Words Encoding	57
3.5.2	Fisher Vector Encoding	60
3.6	Validation Experiments	63
3.7	Discussion and Conclusion	71
4	Detecting Sequences of Interest	73
4.1	Introduction	74
4.2	Clinical Task: Breech Presentation	75
4.3	Image Dataset	77
4.3.1	Standard Plane in Ultrasound Obstetrics	77
4.3.2	New Criteria for Characterising Anatomical Objects	78
4.4	The Dynamical System	81
4.4.1	Detecting Sections of Interest	84
4.4.2	Distance Metrics for Comparing LDSs	88
4.5	Experiments and Results	94
4.6	Conclusion	104
5	Detecting Fetal Presentation and Heartbeat	105
5.1	Introduction	105
5.2	The Framework	106
5.3	Experiments and Results	111
5.3.1	Classification of Individual Frames	113
5.3.1.1	Visualising the Results	115
5.3.2	Detecting the Fetal heartbeat	127
5.4	Conclusions	129
6	Discussion and Future Work	131
6.1	Contributions and Limitations	132
6.2	Exploratory Work and Future Directions	134
6.2.1	Using a Portable Probe	135
6.2.2	Deep Learning Approach	140
6.2.3	Real-Time Performance	144
6.2.4	Summary	145
	Bibliography	146

List of Figures

1.1	Three fetal structures in their standard plane and a typical standardised obstetric scan. Image adapted from [143]. Current standardised obstetric ultrasound scanning protocols include multiple search paths across the maternal abdomen such as (b), where a sonographer scans over multiple paths to locate the best visual representation of the structure of interest for further biometry analysis. . . .	2
1.2	Fetal presentation complications. The relative position of the fetus in the womb is referred to as the presentation. Cephalic (longitudinal) presentation is the most common case and refers to a head-first (head-down) position. Breech presentations is the opposite of cephalic and indicates a head-up position as in schematics 1.2a - 1.2c. Transverse presentation is the least common. Images adapted from [58]	3
2.1	World Health Organisation (WHO) map of global neonatal mortality. Darker colours signify larger mortality rates. Dashed lines represent approximate border lines for which there may not yet be full agreement [110].	8
2.2	Fetal presentation schematic. The relative position of the fetus in the womb is referred to as the presentation. Cephalic (longitudinal) presentation is the most common case and refers to a head-first (head-down) position. Breech presentations is the opposite of cephalic and indicates a head-up position. Transverse presentation is the least common and indicates a horizontal position of the fetus in the womb. Image adapted from [143].	10

2.3	Scattering and speckle in ultrasound images. (a) Scattering is caused by the reflection of ultrasonic waves at irregular (non-smooth) surfaces when the scale of the irregularities is the same as the ultrasound wavelength. This results in echoes of similar weak amplitudes in many different directions. (b) As in scattering, each small target re-radiates the ultrasound. Due to multiple closely located scatterers, the ultrasound wave can either reinforce or cancel out in the tissues around the scatterers. If the waves are in phase, where their pressure peaks are in line, they reinforce (constructive interference), but if the waves are out of phase and the pressure peak of one corresponds to the pressure trough of another, then in that direction the waves cancel (destructive interference). Figures adapted from [80].	12
2.4	Example fetal abdomen structures at various gestational ages (17, 21, 28, 33 and 38 weeks). The first column contain highly occluded examples, the middle column contains examples where key structures are partially occluded, and the last column contains high quality images. Image adapted from [165]	13
2.5	Standard fetal growth measurements schematic. Fetal growth estimates are based on measurements of the fetal skull, abdomen and femur, captured in standard plane. The measurements include the skull and abdominal circumference and the femur length. Many methods have been designed to automate the measurement process. For details refer to text in Section 2.3.1. Image adapted from [143].	15
2.6	Illustration of conventional versus a simplified non-conventional ultrasound scan protocols. (a) The sonographer scans over multiple paths to locate the best visual representation of the key structures, where they are saved for further analysis. (b) Sonographer follows a simple scanning protocol for automated analysis to capture some structure of interest	18

2.7	Ultrasound Scan Spectrum: From left to right of the chart, controlled guidance from the sonographer increases to obtain clinically valid measurements. On the y axis, data acquisition changes from being controlled at the bottom to free-hand on the top. The left-most column represents the default visual output from an ultrasound machine. Object Detection and Localisation in the middle column is still based on constrained scans, requiring a sonographer to navigate the probe to the standard plane. The right-most column represents the algorithms that automated fetal growth measurements from 2D frames. They require the anatomical structures to be captured in their standard plane and therefore require the maximum level of sonographer guidance. Our work falls into the top row of the chart with scan protocols that we refer to as <i>predefined free-hand</i>	21
2.8	Intra-class variation between images obtained from predefined free-hand videos is high. (a) Three example frames of the fetal skull (top row) and the fetal abdomen (bottom row) at various time points during a video sweep. (b) Three standard planes example images of the fetal skull (top row) and fetal abdomen (bottom row) according to the ISUOG standard [179].	22
2.9	A fetoscope can be used to detect the fetal heartbeat. (a) Fetoscope and (b) Pinard's stethoscope are used to listen for the fetal heartbeat. It is important that the hands should not touch stethoscope while listening for the heartbeat. Figures adapted from [162].	23
2.10	Leopolds maneuvers to determine the fetal presentation and position. (a) Assessing the height of the fundus (lower area of the baby) to seeing how many finger breadths below the xiphisternum (bottom of the woman's sternum bone) the baby is laying, (b) Assessing the size of baby and feeling for the baby's back and limbs, (c) Pawlik's grip - the lower part of the uterus is grasped to determine the presenting part, (d) Pelvic palpation to determine the position of the baby's head. Image adapted from [124]	25
2.11	Low-level feature examples. From left to right the grey scale image, local energy filter, local feature symmetry, the product of local feature symmetry and local energy, and signed feature symmetry, edge probabilities using structured edge detection [51], HOG and SIFT features. $\sigma = [28, 30, 32]$ for the filters based on the monogenic signal.	32

2.12	Low-level feature examples. Similar to figure 2.11 but for a different image with significantly varied structures. From left to right the grey scale image, local energy filter, local feature symmetry, the product of local feature symmetry and local energy, and signed feature symmetry, edge probabilities using structured edge detection [51], HOG and SIFT features. $\sigma = [28, 30, 32]$ for the filters based on the monogenic signal. .	33
2.13	Schematic of a linear Support Vector Machine algorithm. . . .	35
2.14	Schematic of Adaboost algorithm.	36
2.15	Schematic of Random Forests algorithm.	37
2.16	Schematic of a single neuron on the left and multi-layer neural network architecture on the right.	38
2.17	LeNet-5 architecture schematic. LeNet-5 Convolutional Neural Network architecture [112]	39
3.1	A simple predefined-freehand scan protocol. A straight ultrasound video sweep is obtained from the cervix up to the top of the maternal abdomen. Three sample frames are shown from various points in the ultrasound video sweep illustrating "Other Structures" (left), "Fetal Skull" (centre) and "Fetal Abdomen" (right). These anatomical structures can provide vital information about the fetal growth and well-being.	41
3.2	Dataset preparation step. Original frame (left). Frame is masked to keep the data at the ultrasound's "field of view" only. This is mainly to remove the text that appears on the top left and right hand sides of the image (centre). Finally the frame is downsampled to half it's original size for computational efficiency (right).	44
3.3	Illustrating intensity invariant features. From left to right, the original frame, Local Energy, Feature Symmetry, the product of the two are illustrated for six different ultrasound images. The filter's σ value was set to $[28, 30, 32]$ for this plot.	46
3.4	The effect of varying the filter σ value. The original images are illustrated on the top row. The product of the feature symmetry and local energy filters are illustrated for each images, increasing the filter's sigma value from 5 to 40.	48
3.5	Comparing feature symmetry and signed feature symmetry filters. Original image, feature symmetry response, product of feature symmetry and local energy, and signed feature symmetry as proposed in [21] from left to write respectively with a filter size of $[28, 30, 32]$. . .	50

3.6	Comparing (feature symmetry $\times$ local energy) and structured random forest edge probability maps. A fetal abdomen (a), and resulting feature symmetry map (b) & edge detection using structured random forest (c))	51
3.7	Illustration of the decision tree node splits. (a) Given a set of structured labels such as segments, a splitting function must be determined. Intuitively a good split (b) groups similar segments, whereas a bad split (c) does not. In practice we cluster the structured labels into two classes (d). Given the class labels, a standard splitting criterion, such as Gini impurity, may be used (e). Figure adapted from [51] . . .	54
3.8	An example SIFT detection 3.8a, and descriptor 3.8b. The Detection selects a point on the image and the descriptor contains the gradient information for the specific local region.	56
3.9	The schematic of computing dense SIFT. A dense SIFT descriptor calculated densely on the entire image. Bin size indicates the size of the descriptor, and the sampling step indicates the density in which the descriptor is created on the image.	57
3.10	SIFT interest points on an ultrasound image. SIFT keypoint descriptors at various scales and orientations are detected. Only 20 keypoints have been selected and plotted at random for illustration purposes.	58
3.11	Bag of words construction schematic. 3.11a Low-level features (e.g SIFT) are computed from a large portion of training images to populate the feature space. The SIFT feature computed on the images are denoted with green colour, and the black dot is the same features in some high-dimensional feature space. 3.11b The features are quantised into W_N clusters to construct a visual vocabulary of words. The blue dots denote the visual words (the cluster centres). 3.11c Given a test image low-level features are extracted and the nearest visual word is identified for each feature. 3.11d A histogram of visual words is used to summarise the test image by counting the number of visual words that occur in the image.	59

3.12	Schematic of steps involved in the BoVW framework. Initially SIFT features are calculated from the (local phase processed) images on a dense grid and vectorised into visual words. Once the visual vocabulary is constructed, training and test images can be encoded by keeping a count of the frequency of each visual word through a histogram. If spatial tiling is used, the encoding can be carried out for each tile of the spatial filter.	60
3.13	Fisher Vector assignment strategy. Compared to the Bag of Visual Word, soft assignment strategy is used which helps reduce information loss during the quantisation process.	61
3.14	Encoding method overview. Fisher vector encoding is computed on dense SIFT features followed by the classification using an SVM classifier. Input images can either be the original images(1), feature symmetry maps(2) or the SRF edge maps(3).	62
3.15	LP-SIFT Results. Top nine ranked images based on the classifier scores on test data for Skull vs not Skull (a,b), Abdomen vs not Abdomen (c,d) and Heart vs not Heart (e,f). SIFT features are used in (a) & (c) & (e) and LP-SIFT in (b) & (d) & (f) with a codebook of 50 words.	67
3.16	LP-SIFT Results. Top nine ranked images based on the classifier scores on test data for Skull vs not Skull (a,b), Abdomen vs not Abdomen (c,d) and Heart vs not Heart (e,f). SIFT features are used in (a) & (c) & (e) and LP-SIFT in (b) & (d) & (f) with a codebook of 100 words.	68
3.17	Comparing the mAV achieved using the BoVW and FV.	71
4.1	Obstetric scanning protocols. (a) Conventional standardised ultrasound scanning protocol where a sonographer scans over multiple paths to locate the best visual representation of the structure of interest for further analysis. (b) User follows a simple predefined free-hand scanning protocol, where automated analysis is carried out on this data to find the structure of interest.	74
4.2	Fetus in cephalic position (left) and breech position (right). The red dotted arrow indicates the ultrasound sweep path. In a video obtained from the described predefined free-hand protocol, identifying the fetal skull in the first part of the video is an indication of the cephalic presentation.	76

4.3	Ultrasound probe's position and angle. The appearance of anatomical structures depends on the position and angle of the ultrasound probe. Here three different planes of the fetal skull are illustrated, where different structures are visible on each plane. Image adapted from [1].	79
4.4	Example frames extracted from a predefined-freehand sweep. Nine borderline frames from a sweep are shown here. From top to bottom the 93 rd , 100 th , 149 th and 156 th frames are on the border of being labelled a fetal skull. Similarly, frames 215, 216, 217 and 218 and frames 241, 246, 248 and 252 are on the borderline of being labelled a fetal heart and abdomen receptively.	80
4.5	Accommodating criteria for fetal skull. The only necessary criterion is presence of a semi-elliptical shape with or without the other standard key landmarks. The criteria are set loosely and independent of the presence of key anatomical structures.	81
4.6	Accommodating criteria for fetal abdomen. A semi-elliptical shape is required with at least one of the key landmarks. The criteria are set very loosely and are not dependent on the presence of key anatomical structures.	81
4.7	Fetal heartbeat as a continuous sequence. The motion patterns can play an important role in detection of certain sequences. For example in this sequence it is challenging to find and located the fetal heart on any single frame. However, it can be seen easily through a continuous sequence, through its distinct motion patterns.	83
4.8	The dynamic texture model graphical representation. In this model $\mathbf{x}_t$ and $\mathbf{y}_t$ are the hidden state and observed video frame at time t respectively.	85
4.9	Snapshots of three key structures at two time points on the noodle phantom. Key structures are short sequences from arbitrary locations on the noodle phantom. Image adapted from [108].	94
4.10	Method schematic overview. Given an unseen test video as in step <i>A</i> , it is treated as an LDS and its model parameters are estimated in step <i>B</i> . A distance kernel is computed between the test and training model parameters using an appropriate distance metric in step <i>C</i> . Finally, in step <i>D</i> a pre-trained classifier is used to classify the test video using the distance kernel from step <i>C</i> .	96

4.11	Variation of classification accuracy for the two distance metrics when using feature symmetry. Classification accuracy is shown (y-axis) using a system order of $n=2$ to $n=10$ (x-axis). The plot shows the effect of changing model order when using the pixel intensities versus using the feature symmetry maps of centre-wavelength=5 to 20.	100
4.12	Variation of classification accuracy for the two distance metrics when using signed feature symmetry. Classification accuracy is shown (y-axis) using a system order of $n=2$ to $n=10$ (x-axis). The plot shows the effect of changing model order when using the pixel intensities versus using the signed feature symmetry maps of centre-wavelength=10 to 30.	101
4.13	Feature symmetry (FS) and signed feature symmetry (SFS) example plots with increasing value of the log-Gabor centre frequency parameter σ. SFS value lies in the range -1 to 1, where -1 indicates a highly symmetric trough (black), 0 indicates no symmetry (grey), and 1 indicates a highly symmetric peak (white). FS values do not take into account the sign, so symmetric local structures are in white and unsymmetrical structures are in black.	102
4.14	Precision-Recall Curves. FS with center-wavelength of 10(top left) and 20 (top right) for Model order of $n=2$ to 10 and SFS curves with the same settings at the bottom.	103
5.1	The main steps of the framework. Given a new training video, all the frames are first classified into "skull", "abdomen", "heart" or "other". If a set of consecutive fetal heart frames are detected in step A, they are further analysed in step B to identify whether a heartbeat can be found. In step T ^A , the green colour represents the training dataset of frames corresponding to the fetal abdominal class, yellow indicates the training dataset of fetal hearts, red indicates the dataset of fetal skulls and white indicates the dataset of frames which belong to the "other" class.	107
5.2	Set of profiles and basis functions with $j = 3$, $k = 4$ (only $k \geq 0$ displayed). The saturation and hue represent the complex magnitude and argument respectively. Image adapted from [21]	110

5.3	Gestational age spread of the diverse clinical fetal cohort. Bar chart of the number of subjects with unknown gestational age and gestational age of younger than 26 weeks (on the left) and bar chart of the number of subjects at each gestational age (on the right).	112
5.4	(a) illustrates SIFT descriptors of various sizes including patches of 8×8 , 16×16 , 32×32 , and 64×64 pixels. Moreover SIFT descriptors are only computed within the red box, to stop computing redundant descriptors. (b) and (c) illustrate the spatial subdivisions. L_1 indicates no spatial subdivisions and L_5 indicates a 4×4 subdivision in addition to the L_1 . A small number of randomly selected descriptors are plotted for illustration purpose and not to clutter the image.	114
5.5	Visualising FV responses on the six examples of the <i>fetal skull</i>. Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores. .	117
5.6	Visualising FV responses on the six examples of the <i>fetal abdomen</i>. Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores. Here the frame at the bottom right corner has been classified incorrectly.	118
5.7	Visualising FV responses on the six examples of the <i>fetal heart</i>. Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores. Here the last two frames have been misclassified.	119
5.8	Visualising FV responses on the six examples of the "<i>other fetal structures</i>". Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores.	120

5.9	Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). The example frames illustrate the visual appearance of the video at various time points. As it is illustrated, the misclassified frame (green spike on the top bar) has been smoothed out after applying the CRF model (on the bottom bar).	121
5.10	Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). The six sample frames correspond to various time points on the video. As is it shown, the misclassified frame (second frame on the top bar) has been corrected for after applying the CRF model.	122
5.11	Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). The heart (first yellow block) and abdomen (first green block) are correctly identified, however the second block of frames classified as heart have been incorrectly labelled. The smoothing process does not rectify this error as expected.	123
5.12	Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). As in the other examples, the green spike on the top bar has been removed as desired after applying the CRF smoothing.	124
5.13	Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). As in the other examples, the green spike on the top bar has been removed as desired after applying the CRF smoothing.	125
5.14	Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). As in the other examples, the green spike on the top bar has been removed as desired after applying the CRF smoothing.	126
5.15	The sequence is cropped around the detected fetal heart in 5.15b to investigate whether this improved the classification accuracy	127
5.16	Protocol limitations. As the fetus size gets smaller in proportion of the available space in the womb, the chances of capturing key anatomical structures of interest decreases due to the flexibility of movement of the fetus in the womb.	130

6.1	The two transducers used. The Philips V7-3 transducer is shown on the left and the Interson Seemore USB transducer on the right. Philips image adapted from http://ultrasound-store.com/products/qz-6073-qzew	136
6.2	Images with different quality. Two sample images are shown obtained using the Philips HD9 and the V7 – 3 transducer and Interson SeeMore probe.	137
6.3	a)The Interson Seemore and Philips images are shown on the top and bottom rows respectively with the Ellipse fitting results on the left and BFM on the right - The fitted ellipse has been highlighted for clarity. (b) The edge detection has not worked well due to the bright attenuation marks at the top mainly on participants who are at earlier gestations ages where the skull for example is small.	138
6.4	A schematic of the CNN used in this thesis. There are 4096 neurons in each of the fully connected layers. There are 4 neurons in the output layer. Image from [70]	141
6.5	Trasnfer learning architecture. Top row: The CNN-A is pre-trained on the ImageNet database for classification. The labelled purple rectangles (e.g. Wc1) rep- resent the weights learned by that layer. Bottom row: In the CNN-B, the first n weight layers (in this example, n = 3) are copied from the base network A. The remaining layers are randomly initialised, and then the entire network is trained on ultrasound datasets with the transferred parameters locked (frozen) or unlocked (fine-tuned). Image from [70]	142
6.6	Classification performance of models directly trained on ultrasound images. Image from [70]	143
6.7	Convolutional kernels learned by the first convolution layer: (left) standard model, (right) added dropout and augmentation. Image from [70]	143
6.8	Classification performance of transfer learning models. Image from [70]	144

List of Tables

3.1	Two-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from the Skull and Other Structures classes, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes.	65
3.2	Two-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from the Skull and Abdomen classes, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes. (b)	65
3.3	Two-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from the Heart and Other Structures classes, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes (c).	66
3.4	Four-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from all four classes of Skull, Abdomen, Heart and Other Structures, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes (d).	66
3.5	Classification results on the larger dataset.	70
4.1	Scoring system for fetal ultrasound images as introduced by Salomon et al. [180]. For each of the three fetal structure, presence of each measurement is associated with a positive score.	78
4.2	Classification on the phantom dataset(a) and the raw clinical (b) and pre-processed clinical (c) datasets using 10-fold cross validation as previously reported in [129].	97
4.3	Classification results for the clinical dataset using Martin distance and BC-max singular values kernel metrics.	98

5.1	Video frame classification results (Step A). The effect of increasing the SIFT descriptor size from an 8×8 to a 16×16 patch is shown, where the number of the GMM is set to 80 modes and the PCA is used to reduce feature dimensions to 40.	115
5.2	Video frame classification results (Step A). The number of GMMs are fixed to 80 words to analyse the effect of varying the feature dimensions from 10 to 80 with $L5$ and without tiling $L1$. With the $L5$ tiling, each image is broken down into 4 parts where a set of feature are computed for that tile. The features are the concatenated together.	115
5.3	Classification results for detecting the fetal heartbeat without localising the fetal heart (Step B). The sequence dimensions are $240 \times 320 \times 30$	128
5.4	Classification results for detecting the fetal heartbeat after cropping the frames around the localised fetal heart (Step B). The sequence dimensions are $120 \times 120 \times 30$, cropped around the detected fetal heart.	128
6.1	The results of visibility & sharpness of the ROI in images obtained with the A and B probes, before and after head detection.	139

Chapter 1

Introduction

1.1 Motivation

Obstetric ultrasonography is a fundamental ingredient of current prenatal care, with its applications ranging from accurate dating of a pregnancy to identifying pregnancy related complications and diagnosis of fetal abnormalities [3]. When compared to other medical imaging modalities, ultrasound benefits from being highly portable and economical, facilitates real-time data acquisition and visualisation, and also benefits from an acceptable safety profile for patients and sonographers alike [11, 86, 139]. In recent years, as ultrasound equipment have become portable and more affordable, the availability of this imaging modality in limited-resource settings has increased with promising trends [5, 25]. However, despite its many benefits, a limitation of ultrasound imaging remains the steep learning curve to scanning proficiency, which can make the process highly challenging for inexperienced users.

The fetal anomaly scan is an essential part of ultrasound screening recommended worldwide for detection of fetal abnormalities and early fetal growth restrictions [153, 201]. During an anomaly scan, a skilled sonographer acquires and records a number of 2D images of key fetal structures and measurement planes following standardised clinical protocols that require the presence of certain anatomical structures [179]. The

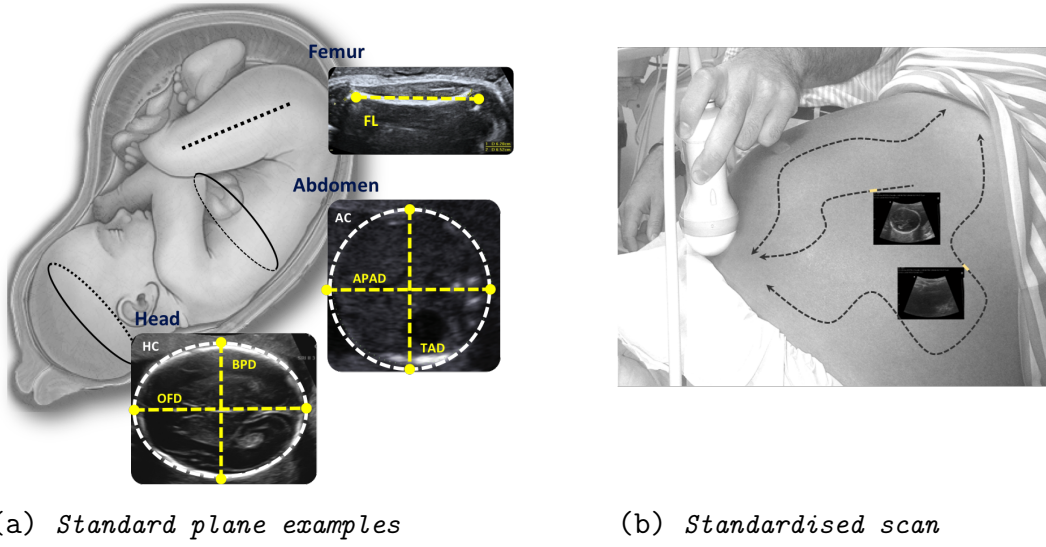


Figure 1.1: *Three fetal structures in their standard plane and a typical standardised obstetric scan. Image adapted from [143]. Current standardised obstetric ultrasound scanning protocols include multiple search paths across the maternal abdomen such as (b), where a sonographer scans over multiple paths to locate the best visual representation of the structure of interest for further biometry analysis.*

goal is to look for the presence of any structural abnormalities and take biometry measurements that are compared against fetal growth charts. Moreover, factors such as fetal presentation are also implicitly identified and noted by the sonographer throughout this process. Figure 1.1a illustrates an example of each of a fetal skull, abdomen and femur standard plane.

Acquiring 2D ultrasound images to comply with standard plane criteria is non-trivial and requires well-trained and skilful sonographers. An example scan path of a conventional ultrasound obstetric procedure is illustrated in Figure 1.1b, where a maternal abdomen is scanned to identify anatomical structures of interest in their standard plane.

Fetal *presentation* and *cardiac activity* are associated with a number of potential pregnancy complications that can be identified in a prenatal obstetric scan. The terms *fetal presentation* and *lie* refer to the relative position of the fetus in the womb, in relation to the maternal abdomen. It is usual for the fetus to be positioned either hor-

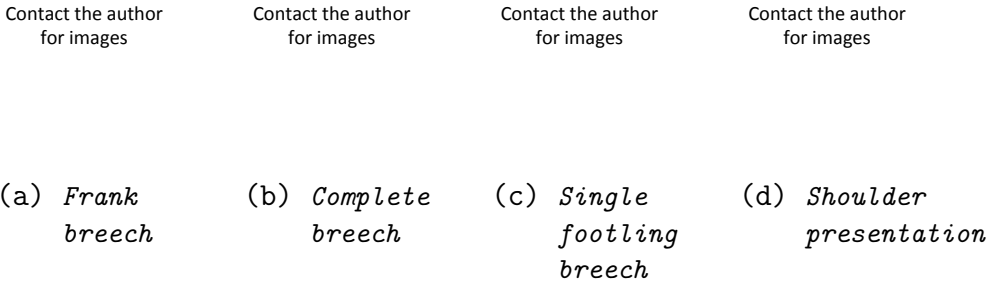


Figure 1.2: *Fetal presentation complications.* The relative position of the fetus in the womb is referred to as the presentation. Cephalic (longitudinal) presentation is the most common case and refers to a head-first (head-down) position. Breech presentations is the opposite of cephalic and indicates a head-up position as in schematics 1.2a - 1.2c. Transverse presentation is the least common. Images adapted from [58]

izontally with respect to the mother’s womb (transverse presentation), or as happens more often to be positioned vertically (cephalic when head is downward or breech when head is upward), as shown in Figure 1.2. More formally, breech presentation [173] is defined as a fetus in a longitudinal lie with the buttocks or feet closest to the cervix and occurs in 3-4% of all deliveries worldwide [87]. The percentage of breech deliveries decreases with advancing gestational age from 22% of births prior to 28 weeks gestation to 7% of births at 32 weeks gestation and 1-3% of births at term [37,48,87,119].

Detecting the fetal cardiac activity and the fetal presentation prior to labour are extremely important. The former confirms fetal viability and the latter allows for effective management of labour. Moreover, previous studies have shown that vaginal birth of a prenatal fetus at a breech position is associated with an increased risk of adverse neonatal outcomes including prenatal mortality and morbidity from difficult delivery, low birth weight from preterm delivery, growth restriction, and even death [119].

1.2 Contributions

This thesis describes a number of methods towards enabling novice sonographers to perform specific and targeted obstetric ultrasound diagnostic screenings. The main contributions of this thesis are as follows:

Automatic Detection Of Frames Of Interest In An Ultrasound Video: A framework for automatic detection of frames of interest in ultrasound videos is proposed in Chapter 2. Parts of this chapter have been presented in the following peer-reviewed conferences:

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Fisher Vector Encoding for Detecting Objects of Interest in Ultrasound Videos. International Symposium on Biomedical Imaging (ISBI), 16-19 April 2015, New York, USA.

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Object Classification in an Ultrasound Video Using LP-SIFT Features. Medical Computer Vision workshop (MCV), MICCAI 2014, Boston, USA.

Automatic Detection Of Sequences Of Interest In An Ultrasound Video: In Chapter 4, the general problem of classifying ultrasound video sequences of interest based on their motion patterns is addressed. Parts of this chapter have been presented in a peer-reviewed conference:

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Searching for Structures of Interest in an Ultrasound Video Sequence. Machine Learning in Medical Imaging workshop (MLMI), MICCAI 2014, Boston, USA.

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Searching for Structures of Interest in an Ultrasound Video Sequence with an Application for detection of Breech. Proceedings of the 24th World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2014, 14-17 September 2014. (Clinical Abstract)

Towards Automatic Detection of High-Risk Pregnancies: In Chapter 5, a framework for automatic detection of the fetal presentation and heartbeat from a predefined-freehand ultrasound video sweep is proposed. Parts of this chapter have been submitted to a peer-reviewed journal and presented at:

Maraci MA, Bridge CP, Napolitano R, Norris T, Aye CYL, Molloholli M, Papageorghiou A, Noble JA. Towards Automating the Six Step Basic Ultrasound Scan. Proceedings of the 25th World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2015, 11-14 October 2015, Montreal, Canada. (Clinical Abstract, **Selected for Oral Communication and Conference Press Release**)

Maraci MA, Bridge CP, Napolitano R, Papageorghiou A, Noble JA. A Framework to Analyse Linear Ultrasound Videos for Detecting Fetal Presentation and Heartbeat. *Submitted to Medical Image Analysis*.

Exploratory Experiments: In Chapter 6, usability of data obtained from a low-cost ultrasound probe has been investigated. Moreover, a collaborative work based on deep convolutional neural networks with transfer learning has also been investigated. Parts of this chapter have been presented in the following peer-reviewed conferences:

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Fetal Head Detection on Images from a Low-Cost Portable USB Ultrasound Device. Conference in Medical Image Understanding and Analysis (MIUA), 17-19 July 2013, Birmingham, UK.

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Automatic Fetal Head Detection on video clips from a Low-Cost Portable USB Ultrasound (US) Device. Proceedings of the 23rd World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2013, 6-9 October 2013. (Clinical Abstract)

Gao Y, **Maraci MA**, Noble JA. Describing Ultrasound Video Content Using Deep Convolutional Neural Networks. International Symposium on Biomedical Imaging (ISBI), 2016, *Accepted paper*.

1.3 Thesis Outline

Chapter 2 describes the clinical application, background knowledge on current approaches to fetal ultrasound examinations and the current challenges and limitations in rolling this technology out to limited-resource settings. This chapter also provides a review of traditional methods, and current state-of-art algorithms relevant to this work.

In **Chapter 3** a general method for multiple object detection in ultrasound images is described. Furthermore, the merits of pre-processing ultrasound images are also evaluated.

Chapter 4 investigates a dynamical model with a suitable metric for classification of ultrasound sequences of interest.

Chapter 5 presents a complete framework with a clinical application for automatic detection of a fetal presentation and heartbeat from a predefined-freehand ultrasound sweep.

Chapter 6 concludes the thesis with initial results on a number of exploratory works that have been carried out, describing potential future work and directions.

1.4 Notation

Throughout this thesis, vectors are denoted using lower case boldface letters and matrices are upper case boldface letters such as $\mathbf{x}$ and $\mathbf{X}$ respectively. Using the current notation by X_{ij} we refer to the entry on the i^{th} row of the j^{th} column in matrix $\mathbf{X}$. Furthermore $\mathbf{X} = [x_1, \dots, x_N]$, $\mathbf{x}_i \in \mathbb{R}^d$ states that matrix $\mathbf{X}$ is composed of N d -dimensional vectors, therefore $\mathbf{X}$ is a $d \times N$ matrix.

Chapter 2

Literature Review

This chapter presents a review of literature relevant to this thesis. The first part provides a brief introduction to ultrasound in medicine, specifically in obstetrics. Section 2.2 describes some of the inherent properties of ultrasound imaging and how they can affect the visual appearance of structures. In Section 2.3, the standard practice of ultrasound-based fetal growth measurement is discussed and some of the main techniques for obtaining automated fetal growth measurements are explained. This is followed by describing the current challenges that are associated with ultrasound imaging in limited-resource settings in Section 2.4. Finally, a discussion of the relevant medical image analysis techniques using machine-learning driven methods is covered in Section 2.5.

2.1 Introduction

Ultrasound is generally the preferred imaging modality for obtaining fetal health and growth measurements, where it accounts for roughly 40% of ultrasound’s total use in medicine [56]. Regular ultrasound examinations provide an effective mechanism for identifying potential pregnancy complications and are therefore an invaluable tool in facilitating pregnancy management. However, recent studies have shown that, annu-

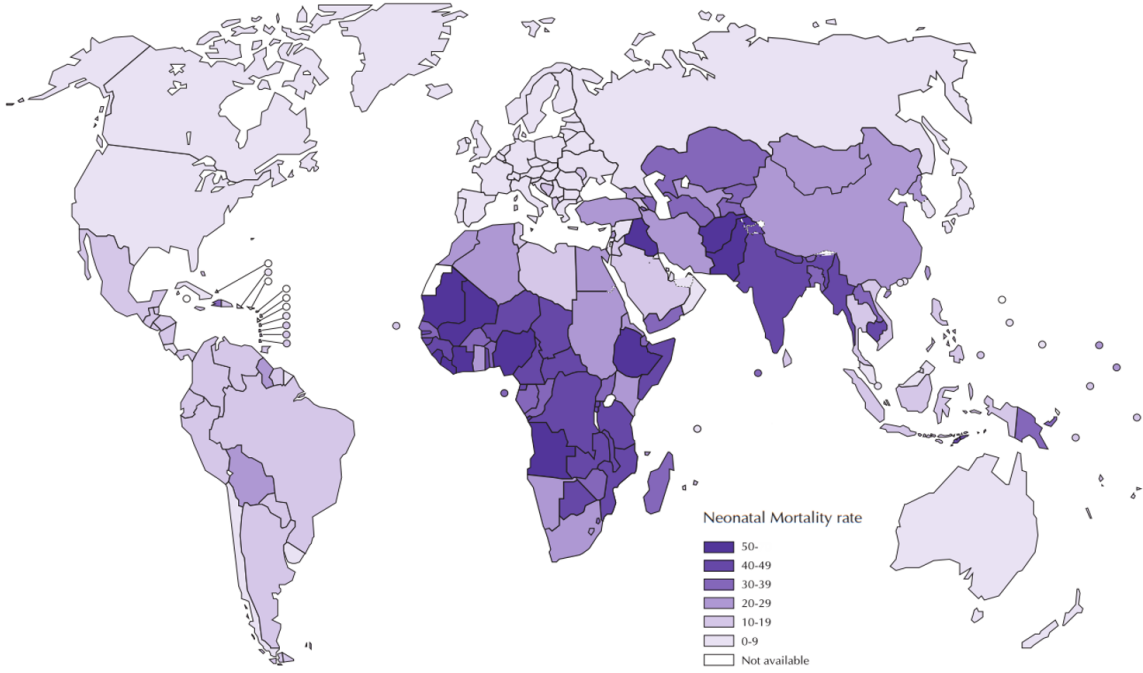


Figure 2.1: *World Health Organisation (WHO) map of global neonatal mortality. Darker colours signify larger mortality rates. Dashed lines represent approximate border lines for which there may not yet be full agreement [110].*

ally, pregnancy related complications account for over half a million deaths in pregnant women [151] and more than three million deaths in neonates, of which over a third could be prevented [110]. In addition, more than 98% of such incidents occur in developing countries, where not all pregnancies are monitored effectively and complications of the simplest kind during pregnancy may result in fatal outcomes. Figure 2.1 illustrates the map of neonatal mortalities around the world, where darker colours represent higher mortality rates.

One of the main challenges with ultrasound imaging is that navigating the probe and performing a diagnostic scan is difficult and requires well-trained and skilled sonographers. Unfortunately in the recent years, there have been numerous reports warning about the lack of well trained sonographers worldwide [41, 79, 144]. For example, in 2003 it was reported that 75% of the medical imaging departments were under-staffed in United States [144] with similar trends in other developed countries too [41, 79].

The challenge is further escalated in developing countries, where the doctor to patient ratio is often low and medical imaging facilities are insufficient and in some instances completely lacking [156], leading to delays in diagnosis and treatment. Therefore incidences of pregnancy related risks are significantly higher in limited-resource settings [49, 67, 88, 136, 158] due to such factors as lack of screening programs, poor nutrition for mothers, and higher likelihood of infections throughout pregnancy.

In order to simplify and promote the use of ultrasound scans throughout pregnancy, Abuhamad et al. [2, 3] has suggested a six step basic fetal ultrasound scan protocol. Their stepwise approach is designed to be simple and is primarily intended for limited-resource settings to identify *high-risk* pregnancies. It includes detection of the fetal presentation, fetal viability (cardiac activity), number of fetuses, placental location, amount of amniotic fluid and finally biometric measurements.

Fetal presentation can be broadly categorised into cephalic, transverse and breech as shown in Figure 2.2. Breech presentation, defined as a fetus in a longitudinal lie with the buttocks or feet closest to the cervix, is associated with an increased risk of pregnancy complication and occurs in 3-4% of all deliveries [87]. The percentage of breech deliveries decreases with advancing gestational age from 22% of births prior to 28 weeks gestation to 7% of births at 32 weeks gestation and 1-3% of births at term [37, 48, 87, 119]. Previous studies have shown that vaginal birth of a prenatal fetus in a breech position is associated with an increased risk of adverse neonatal outcomes and even death [119].

Breech presentation increases the risk of complications at birth as it causes a delay in the birth of the head [38]. If the fetus remains in a breech presentation in the last few weeks prior to labour, a practitioner can decide to turn the fetus around using their hand to a head first position in a process known as the external cephalic version (ECV). There are a number of treatments that may help the success of the ECV. These include using tocolytic drugs (to relax the womb), stimulating the baby with sound

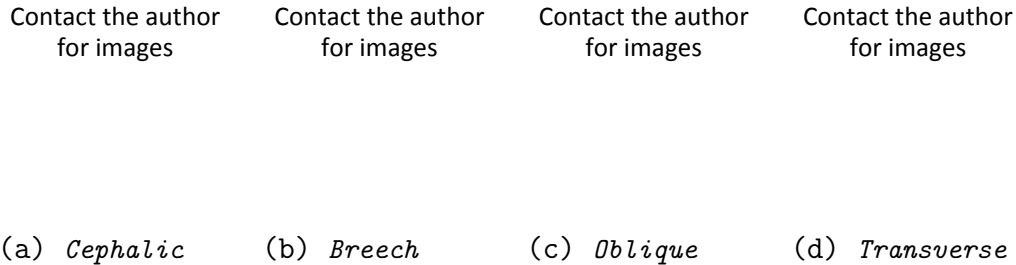


Figure 2.2: Fetal presentation schematic. *The relative position of the fetus in the womb is referred to as the presentation. Cephalic (longitudinal) presentation is the most common case and refers to a head-first (head-down) position. Breech presentations is the opposite of cephalic and indicates a head-up position. Transverse presentation is the least common and indicates a horizontal position of the fetus in the womb. Image adapted from [143].*

through the mother’s abdomen, known as acoustic stimulation, increasing the fluid surrounding the baby (transabdominal amnioinfusion), pain relieving drugs injected into the mother’s lower back to produce regional analgesia or giving the mother opioid drugs to help her to relax [38]. If available, delivery can be performed using a c-section to mitigate the risks associated with breech. Alternatively, vaginal delivery may be performed, in which case prior knowledge of the fetal presentation will be of extreme importance.

2.2 Ultrasound Imaging

Ultrasound is one of the highly utilised imaging modalities in medicine and roughly a fifth of all diagnostic images are obtained through ultrasound scans [80]. Our particular interest is in obstetrics, where ultrasound scans have shown to be an effective method in identifying some of the most prevalent maternal and neonatal risk factors [19, 173]. It is worthwhile mentioning that this imaging modality is mainly utilised by clinicians for a) diagnosis or b) monitoring the success of therapy. In comparison to magnetic resonance imaging (MRI) and X-ray computed tomography (CT), ultrasound images

can be difficult to interpret for some applications. However, characteristics such as lack of radiation, lower associated costs, and imaging soft tissues in real time have made ultrasound the preferred modality in various domains, especially in obstetrics [11, 86, 139, 150].

Despite the benefits of using ultrasound imaging for screening and diagnostic purposes, current approaches to 2D data acquisition and analysis are perceived as key limiting factors for the uptake of this technology by non-experts. Moreover, the inherent properties of ultrasound images can also pose a number of further challenges. For example, image quality, visibility and sharpness are strongly affected by speckle, which causes the characteristic granular appearance in the image. Although this adds the appearance of noise to the image, it may also bear vital and desirable information [149]. In addition, ultrasound waves are poorly transmitted through air and adipose tissue, and are reflected by bone. Aspects such as orientation of the transducer with respect to the imaged object during acquisition, low contrast rates between areas of interest, signal attenuation, and missing information due to signal drop-out and acoustic shadows will affect the overall image quality.

Specifically in fetal ultrasound, the fetal shape and anatomy varies throughout pregnancy. Moreover, the size of the fetus increases as the mother advances in pregnancy and it directly effects the quality of the ultrasound images. The fetal development results in better defined anatomical structures. However, depending on factors such as the mother's body mass index (BMI), it may also lead to an increase in artefacts and shadows during the last few weeks of pregnancy. Figure 2.3 schematically illustrates the relationship between scattering and speckle patterns in ultrasound images. Moreover, Figure 2.4 illustrates fetal abdominal images of poor, fair and good image qualities as determined through a visual assessment in [165] at five different gestational ages. It is important to note that although the anatomical structures are better defined and benefit from better acoustic contrast in later gestational ages, the possibility

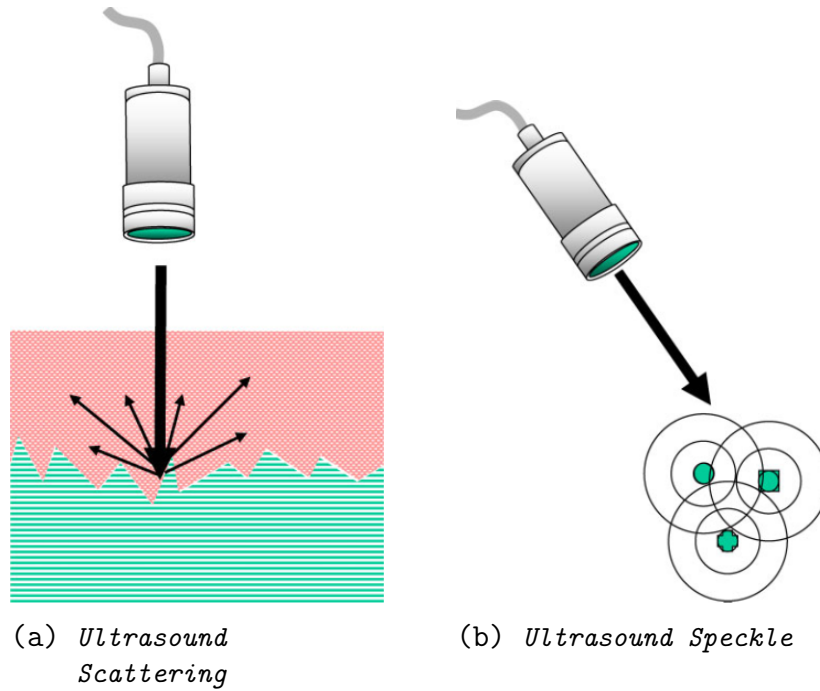


Figure 2.3: Scattering and speckle in ultrasound images. (a) Scattering is caused by the reflection of ultrasonic waves at irregular (non-smooth) surfaces when the scale of the irregularities is the same as the ultrasound wavelength. This results in echoes of similar weak amplitudes in many different directions. (b) As in scattering, each small target re-radiates the ultrasound. Due to multiple closely located scatterers, the ultrasound wave can either reinforce or cancel out in the tissues around the scatterers. If the waves are in phase, where their pressure peaks are in line, they reinforce (constructive interference), but if the waves are out of phase and the pressure peak of one corresponds to the pressure trough of another, then in that direction the waves cancel (destructive interference). Figures adapted from [80].

of occluded structures and artefacts also increases with advancing pregnancy.

2.3 Ultrasound Imaging For Performing Standard Measurements In Obstetrics

Fetal biometry measurements are carried out at several intervals during pregnancy at 6-13 weeks of gestation, 13-25 weeks of gestation and 26-42 weeks of gestation. During a typical fetal growth monitoring scan at 6 to 13 weeks of gestation, the fetus is still very small and fetal age is estimated by measuring the Crown-Rump Length (CRL). The measurement is the longest longitudinal dimension of the fetal pole, from the

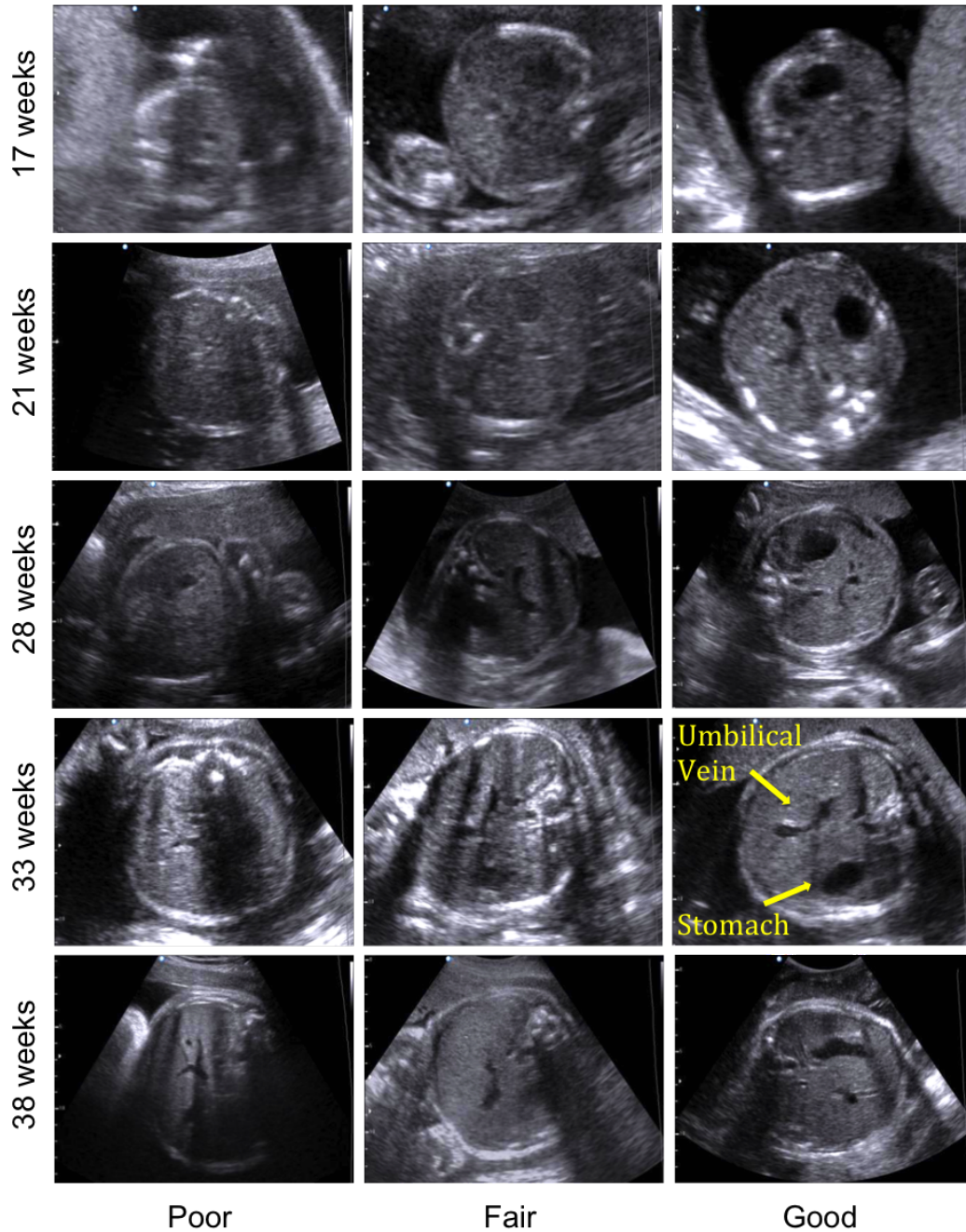


Figure 2.4: *Example fetal abdomen structures at various gestational ages (17, 21, 28, 33 and 38 weeks). The first column contain highly occluded examples, the middle column contains examples where key structures are partially occluded, and the last column contains high quality images. Image adapted from [165]*

top of the head to the bottom of the rump [75]. In the United Kingdom, standard ultrasound fetal biometry measurements are typically taken at the second and third trimester from the skull, abdomen and femur.

From the fetal skull, a number of standard measurements are taken including but not limiting to the Biparietal Diameter (BPD) measured from the outer border of the parietal bones at the widest part of the skull, the Occipito-Frontal Diameter (OFD) measured across the longest part of the skull from the outer border of the occipital and frontal edges and the Head Circumference (HC). Similarly, standard measurements taken from the fetal abdomen include the Anterio-Posterior Abdominal Diameter (APAD) measured from the outer borders of the body outline covering the spine to the anterior abdominal wall, Transverse Abdominal Diameter (TAD) measured from the outer borders of the body outline across the abdomen at the widest point perpendicular to the APAD and Abdominal Circumference (AC) as illustrated in Figure 2.5.

2.3.1 Automated Biometry Measurements

Many automated image analysis methods have been developed in recent years to enhance clinical reporting and eliminate operator variability. However, a large portion of these methods have been designed for single 2D fetal ultrasound images acquired at 18 to 22 weeks of gestation, and mainly focus on standard fetal biometry and growth measurements [153,201]. Fetal biometry measurements are commonly performed manually by well-trained sonographers. One of the main drawbacks of this approach is the high intra-variability between the operator's measurements, which has been a key driving factor for developing automated methods, specifically for fetal biometry analysis and measurements. Earlier developments in automated analysis were mostly based on morphological features. For example fetal skull and femur segmentation was carried out by deploying edge detection, edge linking and Hough transform [81,121,135,199,224]. Al-

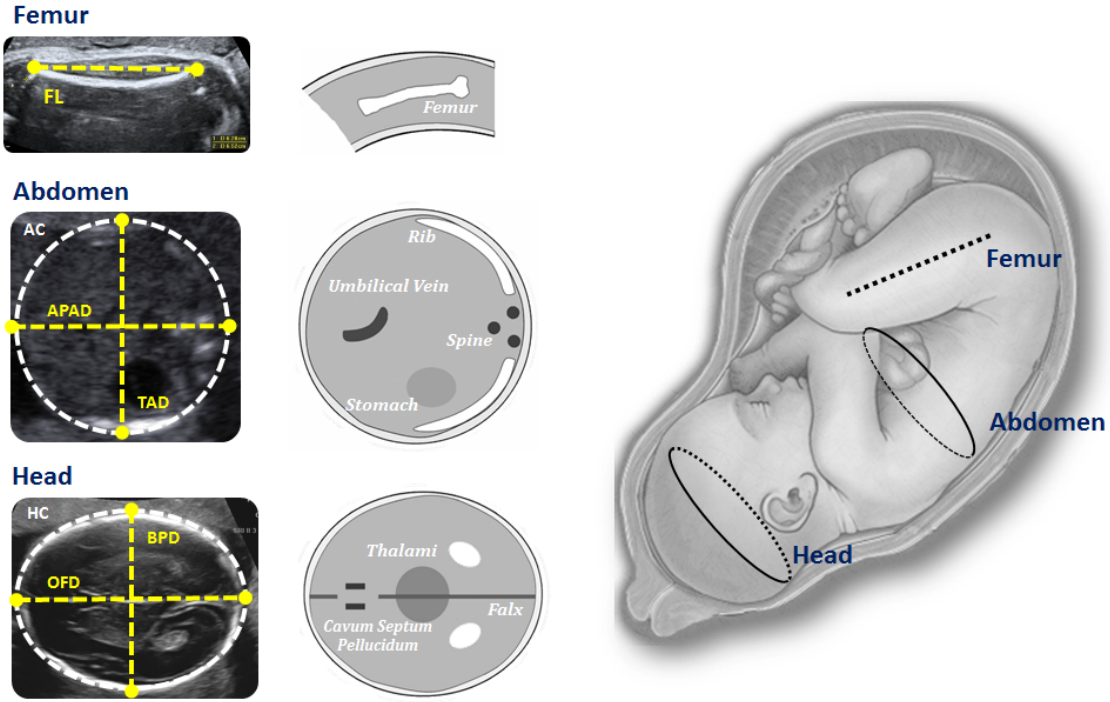


Figure 2.5: Standard fetal growth measurements schematic. Fetal growth estimates are based on measurements of the fetal skull, abdomen and femur, captured in standard plane. The measurements include the skull and abdominal circumference and the femur length. Many methods have been designed to automate the measurement process. For details refer to text in Section 2.3.1. Image adapted from [143].

though the results illustrated high agreement with expert measurements, such methods are heavily based on threshold values selected during the preprocessing stage, which of course can vary greatly between images. Chalana et al. [30] and Pathak et al. [159] used a different approach, where active contour was used to develop a model for fetal head and abdomen segmentation. However, as a limitation, in instances where a local minima was found instead of the global minima, their model required manual correction. Later Jardim et al. [91,92] proposed a parametric deformable shape model for femur and skull segmentation. This model was based on the Rayleigh distribution and was used to segment the pixels based on their texture. However, a key limitation of this model was the need for an initial guess from the user, thus no guarantee of reaching the optimal solution if the initial guess from the user was placed outside a

region of interest due to missing anatomical cues.

For segmentation of the fetal head, abdomen and femur, a gradient vector field method was introduced by Yu et al. [221–223] to fit an ellipse to the edge of the fetal anatomical structures, where the ellipse fitting algorithm was initialised by a Hough transform. However, this method depended on a good initialisation step and was too sensitive to edges in other structures. Detection and segmentation of the fetal head, abdomen and femur was also pursued and developed by Carneiro et al. [29]. They employed a discriminative constrained probabilistic boosting tree classifier for the detection and measurement of head, femur and abdominal structures. In their framework the probabilistic boosting tree classifier was trained on a database of annotated key structures, where the nodes of the binary tree were strong classifiers trained using AdaBoost. Rahmatullah et al. [166, 168] used Adaboost for anatomical object detection in 2D fetal abdominal ultrasound images. Their framework was designed to detect the presence of the correct abdominal landmarks for a standard plane. Sun [196] used a graph-based approach for automatic detection of the fetal skull, where initially the shortest circular path was detected. An ellipse was fitted to the shape for finding the skull boundary. Ponomarev et al. [163] used a multilevel thresholding approach combined with edge detection and shape-based recognition for segmentation of the fetal skull.

While much of the work has been focused on single 2D ultrasound frames, analysis of single 3D ultrasound volumes has recently received some attention too. Automatic extraction of the frontal surface of the fetus from 3D fetal ultrasound was carried out in [147] based on texture classification. In [6] the fetus was segmented from a 3D volume, where amniotic fluid was modelled with a Rayleigh distribution and the fetus with an exponential distribution model. Furthermore, Carneiro et al. [28] extended their framework to 3D volumes of the fetal head for finding specific anatomical structures. Yaqub et al [217] developed a promising approach based on Random Forests to perform

fetal femur segmentation from 3D ultrasound volumes. This framework was later extended to automatically detect local brain structures in 3D fetal ultrasound images [218]. Namburete et al. [142] also utilised Regression Forests to estimate the gestational age of a fetus from sonographic signatures within the brain region. In the latter case, the accuracy of the method in the third trimester was shown to be higher than the current clinical biometry standard.

2.3.2 Current Challenges and Limitations

It is important to note, though not often discussed, that for many of the aforementioned applications in standard 2D and 3D fetal sonography screening, a sonographer follows a standardised clinical protocol, which defines criteria for the plane definition - see for instance the ISUOG guidelines for standard plane criteria [179]. Standardised 2D planes of acquisition undergo specific quality control to ensure they meet a set of predefined criteria. Moreover, sonographers need to be specifically trained to be able to meet these standards, as training programmes have previously shown to improve measurement variability [181] and image quality [215]. We refer to this standardised protocol as a **constrained scan**¹ since all images should have a similar appearance and contain certain anatomical structures, i.e. their appearance is deliberately constrained. These constraints can sometimes assist automated image analysis - for instance in abdominal circumference (AC) measurement, clear visualisation of the stomach bubble, umbilical vein and often the spine is expected - but importantly reduce the degrees of variability with respect to the appearance of a general ultrasound scan of the fetus. Constrained scans are widely used in clinical practice, and simplify the image analysis challenge. However they have a key limitation. Acquisition of constrained scans requires a skilled sonographer. For wider adoption of clinical ultrasound in medicine and

¹In the clinical setting this is referred to as a *standardised scan* but we chose not to use it here as we discuss a more general definition of standardised scans, which accounts for the protocol complexity.

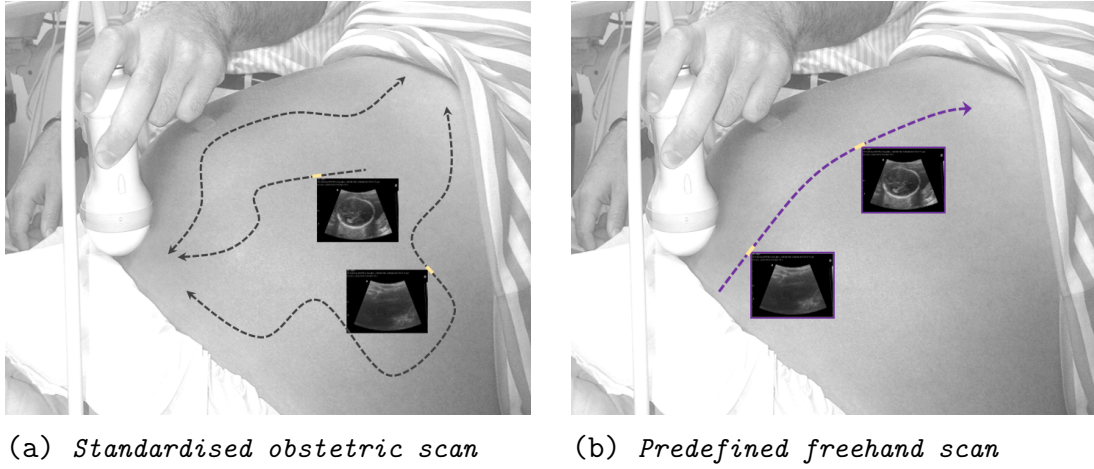


Figure 2.6: *Illustration of conventional versus a simplified non-conventional ultrasound scan protocols. (a) The sonographer scans over multiple paths to locate the best visual representation of the key structures, where they are saved for further analysis. (b) Sonographer follows a simple scanning protocol for automated analysis to capture some structure of interest*

for uptake of ultrasound in limited-resource settings, the need to acquire constrained scans has to be relaxed in favour of much simpler scanning protocols that non-experts can readily adopt.

Encouraging results from observational studies demonstrated that trained and standardised healthcare workers in developing countries could perform as well as qualified sonographers in terms of measurements reproducibility [175]. An automatic video acquisition analysis could potentially help in training, standardisation and quality control in basic obstetric ultrasound for evaluating the fetal presentation and viability. The simplest scanning protocol to learn would be a linear ultrasound video sweep as illustrated in Fig. 2.6b. In our work, we propose the use of this type of scan and name it a **predefined free-hand** acquisition protocol. A novice sonographer can readily be trained to acquire data of this type. It is the analysis of data of this kind that we consider in this thesis. The question is then what useful diagnostic information can be automatically extracted and analysed from such videos.

Given that one of the main aims of this thesis is to take a step towards developing solutions for limited-resource settings and the encouraging results of studies such as

[175] as discussed above, utilising simplified and targeted ultrasound scans can be extremely useful as it would empower less experienced users, midwives, and nurses to carry out certain obstetric screenings. However, it is important to note that such simplified ultrasound scans will contain limited information and cannot be generalised to all types of conditions. Moreover, it is worth noting that key anatomical structures may be missed from such simple sweeps and would therefore require the operator to repeat the acquisition until the necessary structures are detected. Thus, the analysis of data of this kind would necessitate taking such limitations into account. In this thesis, it is the analysis of this data and the image interpretation (classification) that is focused on and the clinical task of positioning the ultrasound probe is an area that requires to be researched further.

The literature suggests a balance between the extent of automation in fetal biometry measurements versus the flexibility of the scanning protocol, such that currently, the acquisition of automatic fetal biometry measurements require a rigid scanning protocol. This thesis however, aims to promote more flexible scanning protocols for achieving relevant and reliable clinical diagnostic outcomes. Accordingly, to place our work in perspective, Figure 2.7 schematically summarises how some of the current state-of-the-art literature in fetal ultrasound image analysis fall between the *skill needed for acquisition and type of image interpretation and analysis (none, detection & localisation, quantification)*. As can be seen, most image analysis literature are in the lower third of this graph (data acquired by a skilled sonographer). We have included the assisted free-hand works of [23, 94, 95], which use controlled mechanical movement of the probe or subject for elastography on the middle row. These methods generate visualisation of ultrasound information and require a small amount of user input to guide probe placement.

Free-hand video sweeps are not yet widely used in ultrasound medical diagnostics and only a few groups [36, 84, 105, 107, 108], including ours, have studied the potential

benefits of this approach, which include enabling novice users during data acquisition and interpretation. Recently Abuhamad et al. [2] has suggested a six step basic fetal ultrasound scan protocol for training non-experts in the developing world to assess a women in pregnancy. Their stepwise approach is designed to be simple and is primarily intended for low-resource settings to identify the *high-risk* pregnancies and includes detection of the fetal presentation and heart beat. We illustrate that automated analysis can be added to this protocol for effective implementation in limited-resource settings. Moreover, an example frame from a fetal abdomen, heart and skull are shown in their standard plane and also how they typically look in a free-hand sweep in Figure 2.8.

2.4 Ultrasound in Limited-Resource Settings

It has previously been shown that ultrasound imaging can indeed play a significant role for diagnosis and management of patients in developing countries [17, 97, 99, 102, 125, 184, 191, 194, 195, 202]. However it is generally accepted that rural and remote regions (particularly in developing countries) suffer from limited diagnostic imaging facilities [156]. The lack of medical experts and advanced medical devices are amongst the contributing factors as to why the latest healthcare products and services may not be accessible in such deprived regions. In addition, it has been suggested by Shah et al. [183] that health care providers in the developing countries have identified lack of training as a primary barrier to implementation of ultrasound in their practice. In this section, we explain the traditional (manual) methods for identifying the fetal presentation and heartbeat that are still used in parts of the world where ultrasound scanning is not yet readily available. Furthermore we review some of the studies designed for promoting the use of ultrasound scans in low-income and limited-resource settings.

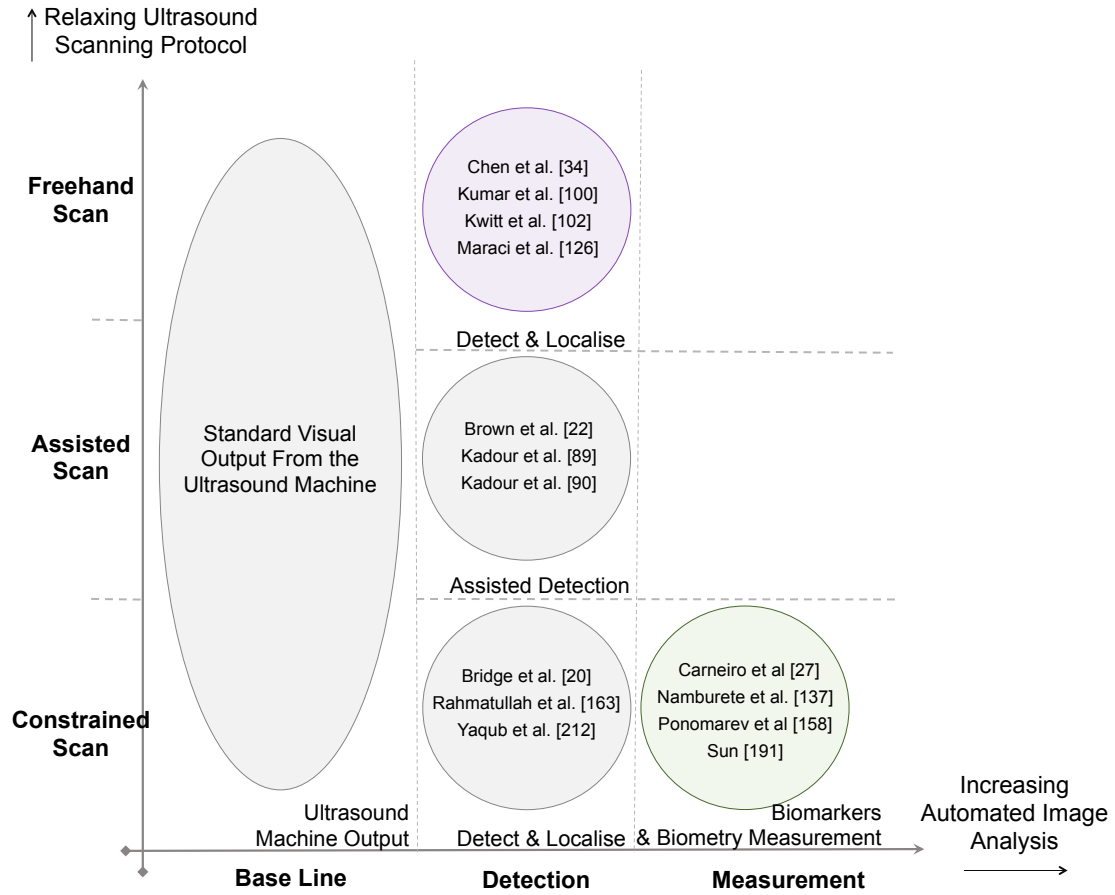
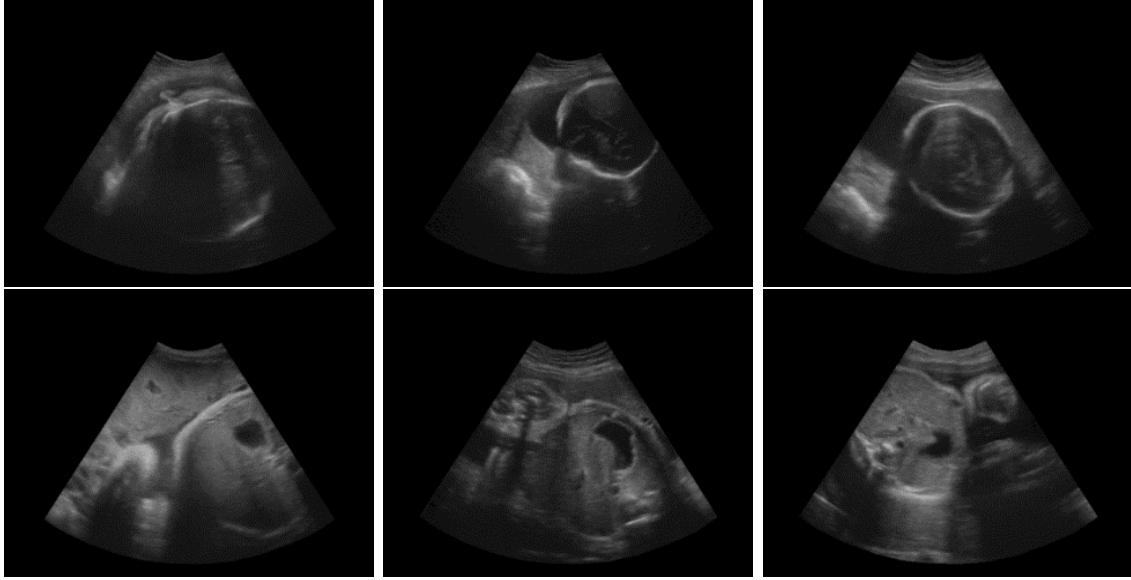
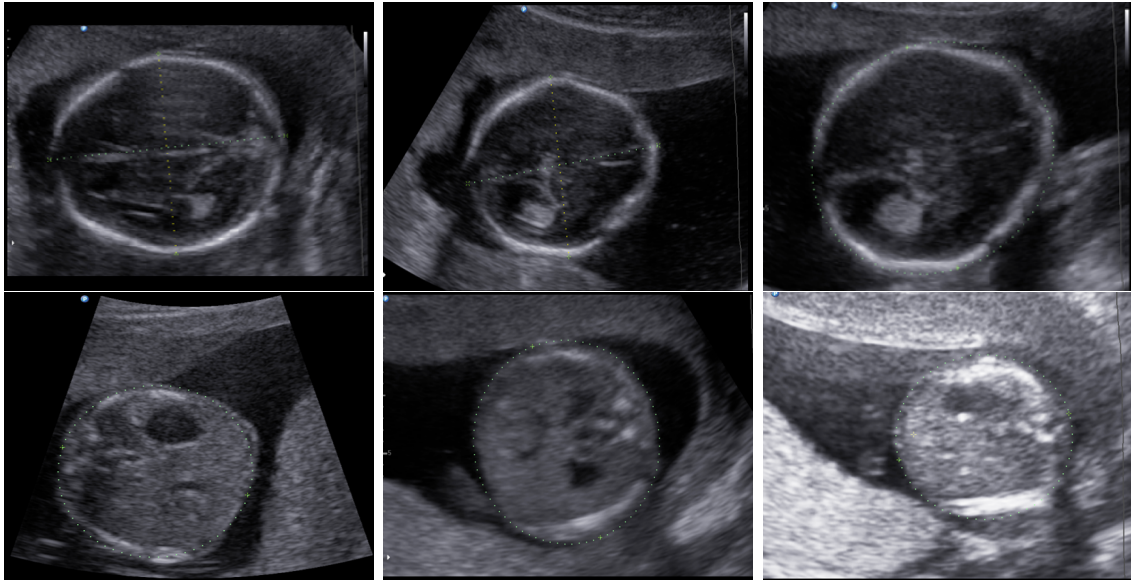


Figure 2.7: Ultrasound Scan Spectrum: From left to right of the chart, controlled guidance from the sonographer increases to obtain clinically valid measurements. On the y axis, data acquisition changes from being controlled at the bottom to free-hand on the top. The left-most column represents the default visual output from an ultrasound machine. Object Detection and Localisation in the middle column is still based on constrained scans, requiring a sonographer to navigate the probe to the standard plane. The right-most column represents the algorithms that automated fetal growth measurements from 2D frames. They require the anatomical structures to be captured in their standard plane and therefore require the maximum level of sonographer guidance. Our work falls into the top row of the chart with scan protocols that we refer to as predefined free-hand.



(a) *Example 2D ultrasound images from a predefined free-hand scan.*



(b) *Example 2D ultrasound images in their standard plane.*

Figure 2.8: *Intra-class variation between images obtained from predefined free-hand videos is high. (a) Three example frames of the fetal skull (top row) and the fetal abdomen (bottom row) at various time points during a video sweep. (b) Three standard planes example images of the fetal skull (top row) and fetal abdomen (bottom row) according to the ISUOG standard [179].*

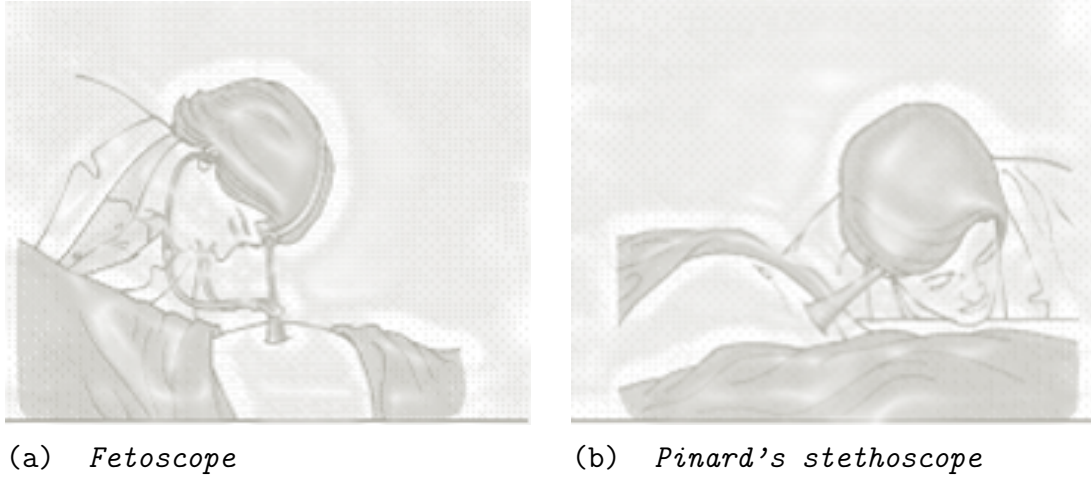


Figure 2.9: *A fetoscope can be used to detect the fetal heartbeat. (a) Fetoscope and (b) Pinard's stethoscope are used to listen for the fetal heartbeat. It is important that the hands should not touch stethoscope while listening for the heartbeat. Figures adapted from [162].*

2.4.1 Manual Detection of Fetal Presentation and Heartbeat

Fetal Heartbeat (Viability)

Identifying the fetal heartbeat throughout pregnancy indicates fetal viability and is extremely important. Furthermore, particular heartbeat patterns during labour are often a significant indicator of fetal well-being. Traditionally, in the absence of an ultrasound device capable of measuring the fetal heartbeat or visualising the fetal cardiac activity, the fetal fetoscope [68, 145] has been used. A fetoscope is similar to a stethoscope and functions by amplifying the sound of the fetal heartbeat. Figure 2.9 illustrates a fetoscope and the original Pinard's stethoscope. Fetoscopes provide a non-invasive form of monitoring and have no adverse effects on the baby. This type of monitoring requires a direct and one-to-one care and is still used in some rural parts of the world. A potential drawback of this approach is the possibility of mistaking the mother's heartbeat for that of the fetus, or in the case of multiple gestations, distinguishing them accurately. Moreover, some studies such as Byaruhanga et al. [26] have shown that more fetal heart rate abnormalities can be detected when ultrasound technology is used compared to using the Pinard's stethoscope.

Fetal Presentation

Knowing the position of the fetus in the uterus is essential for labour management. In the absence of 2D ultrasound scans, this can be carried out manually with four manoeuvres, known as Leopold's manoeuvres, named after the gynaecologist Christian Gerhard Leopold [60]. The first manoeuvre determines which fetal pole occupies the uterine fundus, and whether the fetus is in a breech position. The second manoeuvre is designed to identify on which side the fetal back or smaller parts (feet and hands) are located. As the smaller parts are less firm than the back, the movement can often be apparent. In the third manoeuvre an attempt is made to determine which fetal part is at the lower maternal abdomen, and should validate the findings from the first step. The last manoeuvre is to locate the fetus face. This requires the operator to be facing the pregnant woman's feet. The manoeuvre is illustrated in Figure 2.10. Moreover, in a report by the National Collaborating Centre for Women's and Children's Health in the UK, it was stated that there is a lack of good-quality evidence on the diagnostic value of such abdominal palpation strategy [65].

2.4.2 Emerging Technologies for Limited-Resource Settings

Although lack of access to medical experts and technology remains a challenge, new avenues are opening up in this field with the advancements in the medical imaging technology. Recent studies show that due to the introduction of portable and low cost ultrasound devices, an increase in the use of this technology can be expected in rural areas [189]. In particular, hand-carried portable ultrasound devices represent a valuable alternative to the standard ultrasound machines [45, 98, 100, 116, 122] even though there exist many related challenges that cannot be overlooked such as image quality and user training [5, 25, 102].

Comparative studies across the developed and developing countries have concluded that the overall level and quality of training received by professionals is instrumental to

Contact the author for Images

Contact the author for Images

(a) *Assess fundus height*

(b) *Assess fetus size*

Contact the author for Images

Contact the author for Images

(c) *Pawlik's grip*

(d) *Pelvic palpation*

Figure 2.10: Leopolds maneuvers to determine the fetal presentation and position. (a) Assessing the height of the fundus (lower area of the baby) to seeing how many finger breadths below the xiphisternum (bottom of the woman's sternum bone) the baby is laying, (b) Assessing the size of baby and feeling for the baby's back and limbs, (c) Pawlik's grip - the lower part of the uterus is grasped to determine the presenting part, (d) Pelvic palpation to determine the position of the baby's head. Image adapted from [124]

their ability to successfully use an ultrasound device [4,10,18,22,61,73,85,113,203,209]. In an attempt to empower midwives to perform diagnostic tasks in rural Uganda, Brunette et al. [25] designed and built a low cost portable ultrasound system using off-the-shelf hardware and custom software. In order to allow midwives to perform simple diagnostic tasks with minimal training, a simplified user interface was designed and implemented by removing the unnecessary options for performing a basic scan. This enabled the detection of breech presentation, multiple gestation and placenta previa. Furthermore, Anderson et al. [5] expanded the framework by improving the built-in help and tutorial software, to further facilitate the training of the midwives throughout a basic ultrasound exam.

Another approach to training skilled sonographers both in the developed and developing countries is through software simulations. Although phantoms exist and are often used for educational purposes, they are usually expensive and are only available for a limited number of anatomical parts. To overcome this, Nicolau et al. [148] constructed a low-cost simulator to allow ultrasound image practice and also probe manipulation through a CT constructed model. They illustrated that the movement of a probe phantom could be tracked and registered to a CT model of the anatomy. However they concluded that simulating realistic ultrasound data needed further work and enhancements. In another study conducted in collaboration with a university hospital in Ghana, the authors designed a low cost system for training ultrasound imaging techniques using components from the Nintendo Wii game console [7]. In their study, they intended to further enhance sonographer skills through simulated ultrasound training.

In another approach, the Ifakara Health Institute in Tanzania [137] conducted a study, where a number of non-expert primary healthcare workers including midwives, nurses and medical attendants were trained for 10 days to use a handheld ultrasound device. They were then asked to perform basic obstetric ultrasound scanning at rural health centres and an independent radiologist was asked to evaluate the accuracy

of their diagnosis. The user's diagnostic accuracy was found to be above 80% leading to the conclusion that non-experts could be trained to perform basic ultrasound screenings. In rural Uganda, willingness of locals to attend the antenatal clinics was evaluated and it was shown that implementing ultrasound in a primary care facility led to a 72% increase in total antenatal care visits [176]. In rural Zambia, the effect of equipping midwives with portable ultrasound was evaluated and it was observed that this helped 17% of obstetric cases to change and optimise the clinical management of the patients, particularly in cases with malpresentation and multiple gestations [97].

Yet another approach to tackle the lack of experts in limited-resource settings is to use teleradiology as proposed for example by Kuntalp et al. [106] and DeStigter et al. [50]. Their proposed care model was built to enable minimally-trained primary healthcare workers to capture and transmit ultrasound free-hand sweeps of key anatomical regions of interest. The data was then uploaded to a central server to be accessed by experts either locally or remotely. These expert radiologists then reviewed the images for key indications and transmitted their findings via text message back to the rural areas via the cellular phone network. As reported by Pool et al. [164], the local operators underwent on-site training to learn how to acquire a video or cine sweep through a target organ or body region using only external anatomical landmarks, with no specific internal imaging anatomy or pathology training. Video-telephone technology was used to allow remote communication and further training of users. This approach focused on training local primary healthcare workers with simple acquisition protocols but still required expert radiologists on one end of the system to interpret the data.

2.5 Automated Methods for Fetal Ultrasound Imaging

Many automated methods suggested in previous studies aim to reproduce the standard fetal clinical measurements, to directly analyse and assess fetal growth throughout pregnancy at different gestational ages. Their aim has been to reduce inter and intra observer variabilities, and more importantly establish a standardised and reproducible measurement procedure. The standard approach for many automated algorithms for detection, localisation and segmentation of fetal anatomical structures is applying machine learning techniques. Machine learning is arguably very well-suited to recognise sonographic patterns in ultrasound images, and can form the basis of image-based decision-making. As previously discussed, research has looked into automating biometry measurements. However this process is dependant on skilled sonographers for data acquisition. One approach is to simplify the data acquisition protocols by way of free-hand video sweeps, and develop solutions to automatically interpret this data through machine learning methods. Several methods have been recently proposed for automatic detection and localisation of anatomical fetal structures from ultrasound videos. Linear Dynamical Systems (LDS) were used to localise structures of interest in an ultrasound video obtained from a phantom by Kwitt et al. [107]. In our own work, a method for clinical ultrasound video sequences was developed [129], where the original video is broken into smaller sequences of shorter length such that all sub-sequences have the same length. The pixel motion patterns (dynamics) of the sequences are then learned using a linear dynamical system. Identification and classification of the sequences of interest are then based on the similarities between the estimated LDS model parameters. Due to the standardised freehand acquisition protocol, the probe motion patterns are a constant factor throughout the video for all the video sweeps in our dataset and do not need to be taken into account separately. In an attempt to au-

tomate the detection of the fetal abdominal standard plane, Kumar and Shriram [105] applied a method based on the spatial configuration of key anatomical landmarks in a video to identify the fetal abdominal best plane. Chen et al. [36] utilised a convolutional neural network (CNN) for standard plane localisation of the skull and abdomen from an ultrasound video, although the details of acquisition were not stipulated. In previous work we have used the bag of visual words approach with feature symmetry filters [130] as well as Improved Fisher Vector (IFV) encoding [131] with a Support Vector Machine (SVM) to identify frames of interest in an ultrasound video sweep. In this section we review some of the most widely used machine learning techniques that are utilised throughout this thesis.

2.5.1 Machine Learning in Medical Imaging

Machine learning can be broadly defined as the ability to learn interesting patterns and information from data and has played an important role in the field of medical image analysis in recent years. It can be divided into mainly supervised (when groundtruth labels are available) versus unsupervised learning, and discriminative (when the model maps the input to the output) versus generative (where the full representation of the input data is modelled) [146]. Machine learning algorithms used for object detection in medical imaging are mainly supervised. They can be interpreted as systems or frameworks that learn from a clinician's prior knowledge about diagnostic analysis of a given data [197]. Examples of applications which use such algorithms include computer-aided diagnosis [142], image segmentation [114], image registration [43] and anatomical object detection and localisation [168, 197, 217, 218]. Learning-based approaches to automatic object detection in ultrasound images can work well, as high-dimensional input data can be efficiently mapped to compact representations of image patterns representing various anatomical structures. This requires finding good discriminating (high-dimensional) features. In what follows, a variety of feature descriptors used in

machine learning applications are discussed.

2.5.2 Feature Descriptors

An important step in constructing many robust machine learning classifiers is to select appropriate feature descriptors to allow for effective discrimination between different classes. Features that are based on the visual appearance of objects such as optic flow vectors [12, 15, 24, 89, 123, 198], texture and raw pixel intensities [54, 78, 83, 111, 126] or edges [27, 63, 155, 186, 187] are powerful attributes for object detection. However, in ultrasound images speckle patterns, low contrast, and poor signal-to-noise ratio can make it difficult to use such features solely on their own, especially if the training data sample is not large enough to capture such variabilities in data. To overcome this, intensity invariant methods and filters such as local feature symmetry [103] and signed feature symmetry [21] have been previously used to enhance features of ultrasound images [21, 62, 76, 77, 140, 168, 226]. Figures 2.11 and 2.12 illustrate two example ultrasound images of different fetal structures, which have been processed using local energy, feature symmetry, product of local energy and feature symmetry, signed feature symmetry, structured edge probabilities [51], histogram of oriented gradients (HOG) features [44], and scale invariant feature transform (SIFT) [120] features.

Gradient based features are another set of popular feature descriptors within the computer vision community, which have been widely used for human detection [44, 177]. A subset of methods within the family of gradient based features use contour information to represent an object of interest. This was demonstrated in [72], where a contour based hierarchical chamfer matching detector was used for pedestrian detection. These features have also been adapted in the medical imaging community where boundary fragments were used in [192] for representation and segmentation of the left ventricle. However, a high number of false positives can be expected in cluttered regions due to the weak notion of shapes. Moreover relatively large training sets are required in or-

der to learn shape variations. An alternative approach is to use statistical information obtained from gradient features, an example being SIFT, introduced by Lowe [120], which has become one of the well known and widely used approaches of this type. To compute SIFT features, a histogram of local oriented gradients is calculated around a point of interest, which can then be used to describe objects of interest in an image. The HOG [44] is yet another powerful feature descriptor which is based on the occurrences of gradient orientation in localised portions of an image.

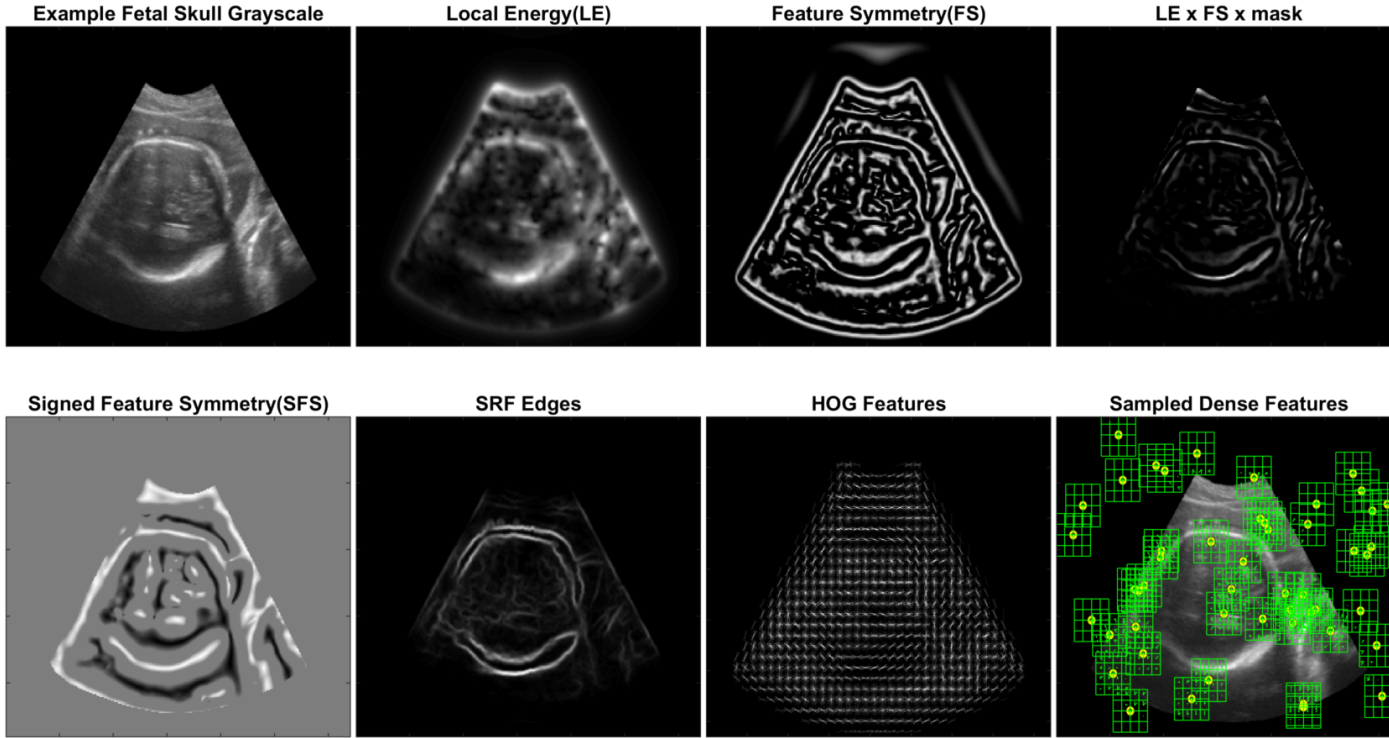


Figure 2.11: Low-level feature examples. From left to right the grey scale image, local energy filter, local feature symmetry, the product of local feature symmetry and local energy, and signed feature symmetry, edge probabilities using structured edge detection [51], HOG and SIFT features. $\sigma = [28, 30, 32]$ for the filters based on the monogenic signal.

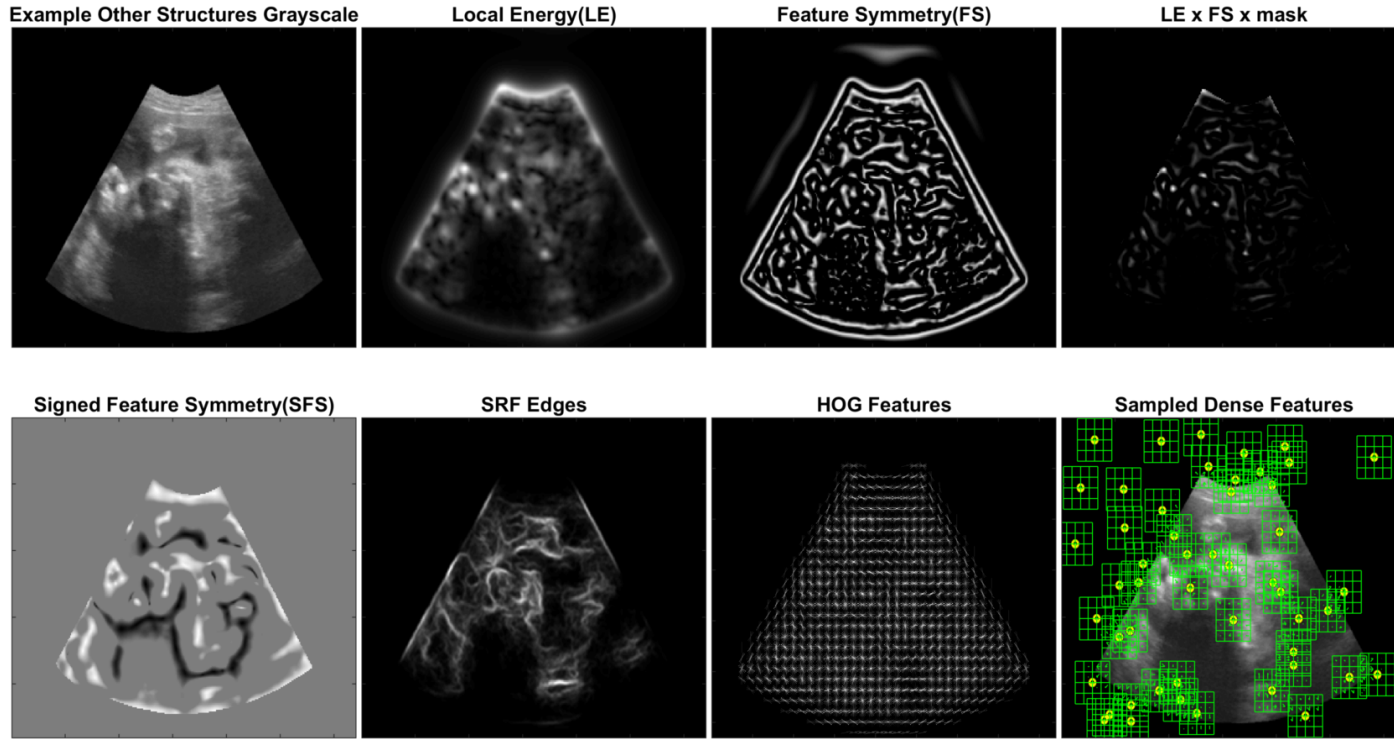


Figure 2.12: Low-level feature examples. Similar to figure 2.11 but for a different image with significantly varied structures. From left to right the grey scale image, local energy filter, local feature symmetry, the product of local feature symmetry and local energy, and signed feature symmetry, edge probabilities using structured edge detection [51], HOG and SIFT features. $\sigma = [28, 30, 32]$ for the filters based on the monogenic signal.

2.5.3 Learning Mechanisms

Support Vector Machine (SVM) A popular and now classic learning mechanism for object detection is using a Support Vector Machine (SVM) [39, 204]. SVMs are a type of supervised, discriminative maximum margin classifier. Given some training data for two different classes with their target labels, the algorithm tries to find a decision boundary or a hyperplane such that the margin between the two classes are maximised in the feature space as illustrated in 2.13.

During the training phase, the decision boundary is set based on only a subset of the training examples, known as the support vectors, such that the complexity of the system is dependent on the number of support vectors used. One of the advantages of using SVMs for solving classification problems is that they allow for an efficient use of kernels. By this we mean that in many real-world problems, it is unlikely that the data points from different classes are separable in the input space. SVMs allow an efficient use of kernels to transform the data into a higher dimensional feature space, where data points can be separated effectively. While linear kernels are computationally efficient [93], non-linear kernels have shown to produce higher classification accuracy [225].

In our work (Chapter 3), we make use of the homogeneous kernels, which have shown to significantly improve classification results with only slight increase in the computational cost [206]. Specifically, this learning mechanism has been used as part of a multi-class classification to categorise the frames into four classes of skull, abdomen, heart and other, making use of the linear kernel, the χ^2 kernel and the Hellinger's kernel. SVM's have previously been used in the medical imaging domain including in various applications such as [59, 129–131].

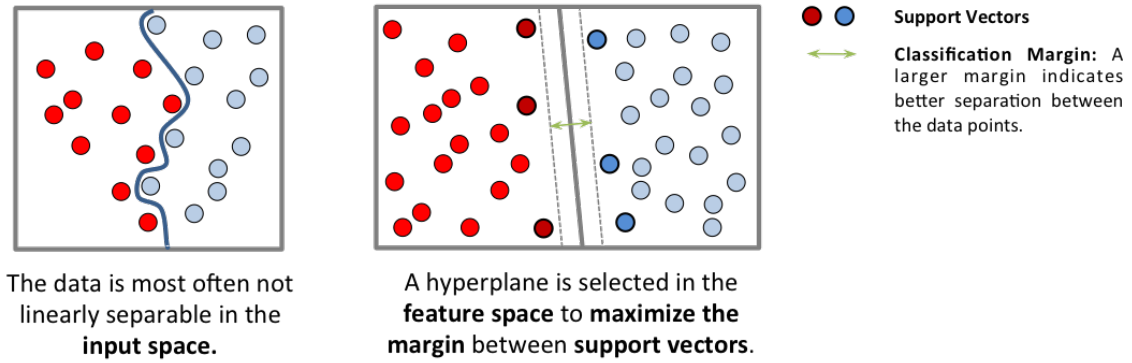


Figure 2.13: *Schematic of a linear Support Vector Machine algorithm.*

Adaptive Boosting (Adaboost) Adaptive Boosting also known as Adaboost, is a well known machine learning algorithm that works by combining a weighted sum of multiple weak classifiers to form a strong classifier [69, 213]. This algorithm benefits from an iterative framework such that at each step, the training examples that have been misclassified in the previous step are given a higher weight as show in Figure 2.14. Boosting could lead to accurate results even if the weak classifiers have a performance that is only marginally better than random, hence the term 'weak learners'. This method benefits from having no parameters that require tuning other than the number of iterations. One of the earlier applications of Adaboost is the work of Tieu and Viola [200] which used boosting to select a small set of discriminatory features from a larger pool for object tracking in natural scene. This inspired one of the earlier face detectors by Viola and Jones [213]. Adaboost has been successfully used in a number of fetal biometry applications including [141, 168]. In this thesis however we have decided not to use this learning framework, as kernels integrate naturally in an SVM framework to produce satisfactory yet computationally efficient results.

Random Decision Forests Random Forests [20] is a discriminative learner that is composed of a collection of binary decision trees constructed by randomly selecting the best features at each node, based on some criteria, to partition the data into the possible class labels. During this step, the best feature is generally selected such that

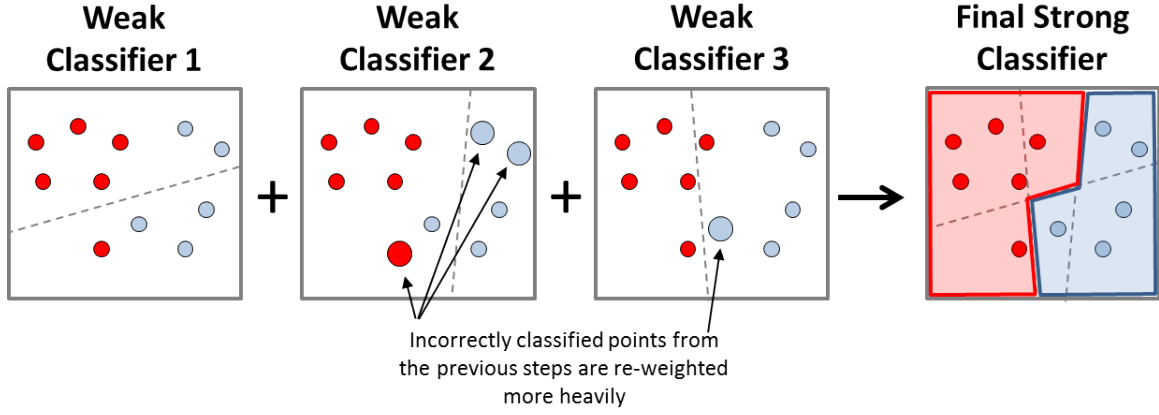


Figure 2.14: *Schematic of Adaboost algorithm.*

the information gain is maximized. At the testing stage, the test data points traverse the tree, where the data is split at each node based on the pre-calculated threshold, until a leaf node is reached. Each tree will then produce a probability for the class labels, and the final prediction is achieved through a combination of all the tree outputs as shown in Figure 2.15.

Random Forests provide a strong learning mechanism and have been successfully used in the medical imaging domain for many applications including automatic delineation of myocardium in 3D echocardiography [114], and 2D + t [219] ultrasound images. Furthermore, to perform fetal femur segmentation from 3D ultrasound volumes, Yaqub et al [217] successfully used Random Forests. This framework was later extended to automatically detect local brain structures in 3D fetal ultrasound images [218], where the accuracy of the algorithm varied from 75% to 95% for localising various fetal brain structures. Namburete et al. [142] used Regression Forests to estimate the gestational age of a fetus from sonographic signatures in the brain, where the accuracy of the method in the third trimester was shown to be higher than the current clinical standard, reducing the root mean squared error from 7.01 from manual detection to 6.1. In this thesis, we have utilised a Structured Random Forest (SRF), which is an extension of standard Random Forests. The SRF is used and discussed further in Chapter 4 in order to learn edge probabilities for anatomical structures of

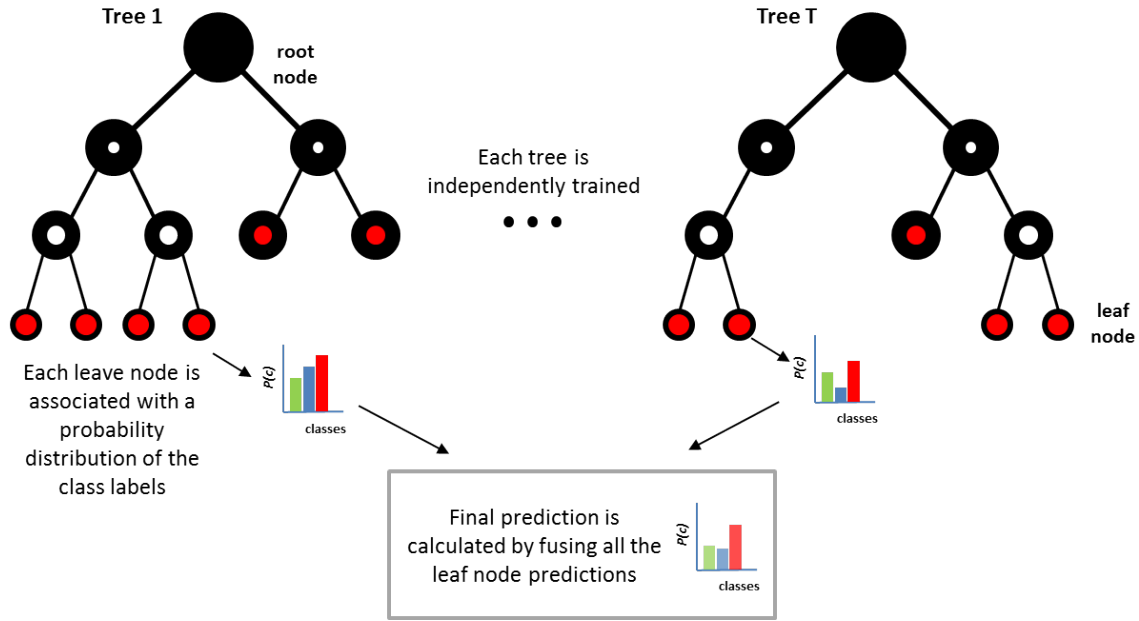


Figure 2.15: *Schematic of Random Forests algorithm.*

interest in ultrasound images.

Deep Learning and CNNs Early application of machine learning originated from the development of the artificial neural networks (ANNs), inspired by studies on the human brain neurons [13, 138]. ANNs are generally a compilation of inter-connected *neurons*, which exchange messages with each other. Each connection in the network has an associated weight that is iteratively tuned or *learned* based on experience. For any given application, a neural network is defined by a set of input neurons (i.e. the input data such as image patches), where their output becomes the input to the next layer. As an example, an ANN for handwriting recognition is defined by a set of input neurons that can be activated by the pixels of an input image. The activation of these neurons (the outputs) are then passed on to other neurons (as their input), once weighted and transformed by a predefined activation function (such as the sigmoid function). This process is repeated until an output neuron is activated. This process determines which character was read. Figure 2.16 illustrated a multi layer ANN.

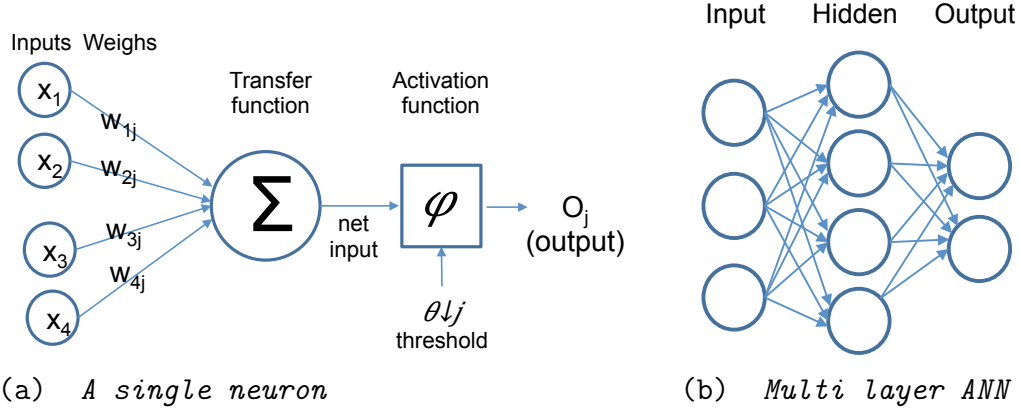


Figure 2.16: *Schematic of a single neuron on the left and multi-layer neural network architecture on the right.*

Convolutional Neural Networks (CNN), a variant of multi-layered neural networks, was inspired by biological mechanisms discovered by Hubel and Wiesel's early work on the cat's visual cortex [90]. It is now known that the visual cortex contains a complex arrangement of cells sensitive to small sub-regions of the visual field, called a receptive field. These sub-regions are tiled covering the entire visual field. Furthermore, the cells act as local filters over the input space (for example the image) and can effectively exploit local spatial correlations in images. A feature map is obtained by applying a function across sub-regions of the entire image repeatedly. This equates to convolving the input image with a linear filter, adding a bias term followed by applying a non-linear function. Figure 2.17 illustrates the structure of LeNet-5 [112], one of the earlier and well known CNNs in the literature.

There has been a rapidly growing interest in utilising CNNs in the medical imaging domain with highly promising results. For instance, Wu et al. [216] applied a deep learning framework for classification of focal liver lesions with contrast-enhanced ultrasound. A CNN learning framework was utilised in [36] for standard plane localisation in fetal ultrasound, where a transfer learning mechanism was used to fine-tune a pre-trained network on medical ultrasound data. An advantage of such framework is that instead of using low-level hand-made features such as SIFT and HOG, low-level

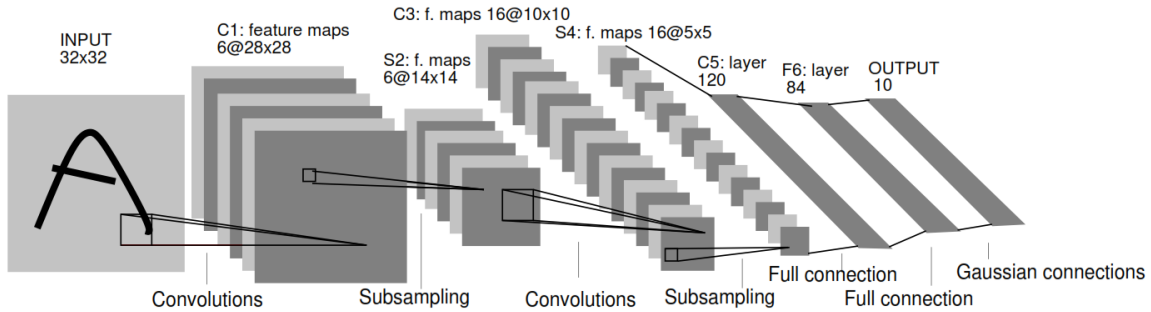


Figure 2.17: *LeNet-5 architecture schematic.* *LeNet-5 Convolutional Neural Network architecture [112]*

features can be learned from the data. This however requires having a large dataset with tens of thousands of images and is further investigated as part of collaborative work and we report our initial results in Chapter 6.

2.6 Conclusion

Ultrasound is the imaging modality of choice in prenatal care. However, ultrasound signal properties and its interactions with tissues can cause characteristics such as attenuation and acoustic shadows. These are unique properties of ultrasound that are not exhibited in other imaging modalities. Despite the body of work on object detection and localisation, little work has been carried out in relation to ultrasound videos. Thus, it is necessary to design and implement ultrasound specific tools to address these challenges in order to take one more step in facilitating the use of ultrasound for pregnancy management in rural and limited-resourced regions. This thesis is set to do this. We demonstrate how this imaging modality can be utilised by non-experts to acquire data and make important diagnostic decisions.

Chapter 3

Multi Object Detection from Ultrasound Videos

In this chapter we develop and compare two original methods to automatically identify the frames of interest in an ultrasound video sweep. As pre-processing ultrasound images prior to any analysis is a common theme in many previous studies, we investigate the merits of pre-processing images on classification accuracy on two datasets of variable sizes. The data acquisition step is defined by a simple standardised clinical protocol as illustrated in Figure 3.1 and is explained in more detail later in Chapter 4. In this chapter, the framework and the process for automatic detection of anatomical structures of interest from each video frame are discussed. The technical contributions of this chapter are as follows; firstly to introduce LP-SIFT (Local Phase based SIFT) features, secondly to illustrate the efficacy of preprocessing ultrasound images before classification, and finally a classification framework for 2D ultrasound frames based on the improved Fisher Vector encoding. An early version of the work presented in this chapter has been reported in:

Maraci MA, Napolitano R, Papageorgiou A, Noble JA. Fisher Vector Encoding for Detecting Objects of Interest in Ultrasound Videos. International Symposium on Biomedical Imaging (ISBI), 16-19 April 2015, New York, USA.

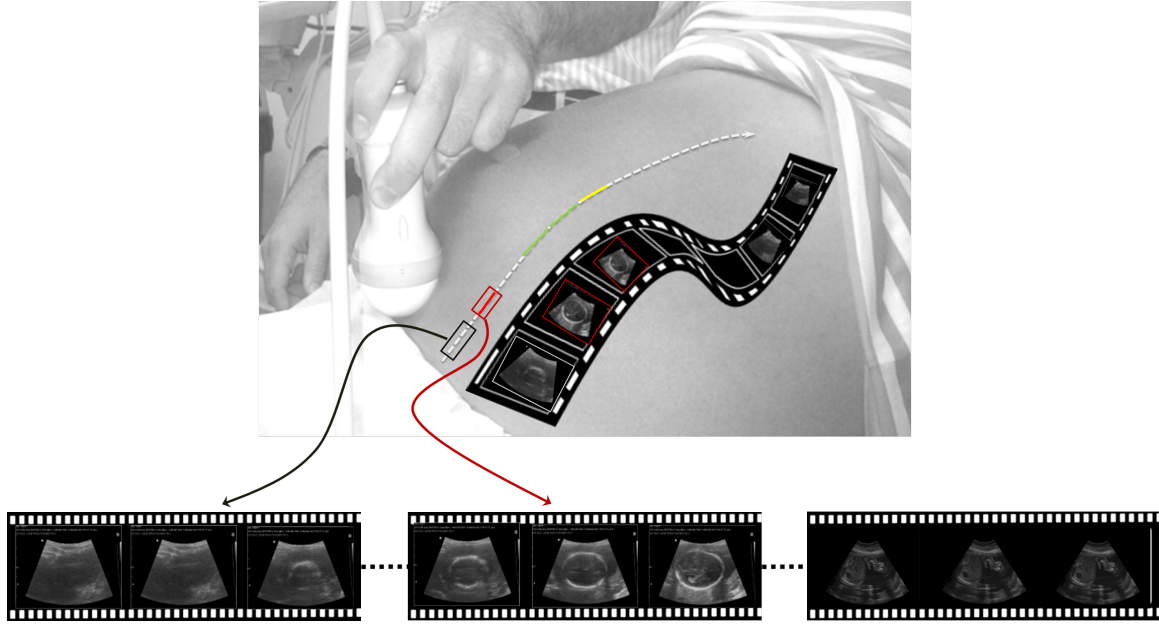


Figure 3.1: *A simple predefined-freehand scan protocol. A straight ultrasound video sweep is obtained from the cervix up to the top of the maternal abdomen. Three sample frames are shown from various points in the ultrasound video sweep illustrating "Other Structures" (left), "Fetal Skull" (centre) and "Fetal Abdomen" (right). These anatomical structures can provide vital information about the fetal growth and well-being.*

Maraci MA, Napolitano R, Papageorgiou A, Noble JA. Object Classification in an Ultrasound Video Using LP-SIFT Features. Medical Computer Vision workshop (MCV), MICCAI 2014, Boston, USA.

3.1 Introduction

To empower relatively untrained sonographers to perform diagnostic ultrasound screening, two main prerequisites need to be fulfilled: Firstly, specific scanning protocols have to be designed for novice users such that they can be readily learned, yet they must contain sufficient information for automated analysis. Secondly, a framework has to be developed for automated analysis of such data, to aid and inform the sonographers of clinical conditions. The latter is considered in this chapter. Moreover, this is of particular relevance to limited-resource settings, where pregnant women may not be

able to attend specialist centres for obstetric care regularly or even at all throughout the course of their pregnancy. Thus, empowering non-experts in performing diagnostic ultrasound scans can potentially have a significantly positive impact.

In this chapter, a general framework is proposed for automatic detection of a number of key fetal anatomical structures including the fetal skull, heart, and abdomen from a continuous 2D ultrasound video sweep. Furthermore, we report on two well-known feature encoding frameworks; the Bag of Visual Words (BoVW) and the improved Fisher Vector (IFV) encoding models. From here on, improved Fisher Vector will be denoted as FV. The low-level features encoded with these models are the Scale-invariant feature transform (SIFT) [120] and the LP-SIFT (dense SIFT applied to contrast-invariant images) features [130]. We demonstrate that pre-processing and filtering 2D ultrasound images prior to data analysis can be beneficial on smaller datasets, where data variability may be difficult to capture. For such cases we propose learning relevant edge probability maps as suggested by Dollar et al. [51] through a structured random forest, as opposed to the previously used feature symmetry map [169]. In what follows, the image dataset and acquisition protocol are explained and discussed.

3.2 Ultrasound Image Dataset

Clinical 2D fetal ultrasound video sweeps were acquired specifically for this study using a Philips HD9 ultrasound machine with a V7-3 transducer. Data acquisition was carried out by five qualified sonographers who were asked to follow a simple standardised scanning protocol. Prior to data acquisition, each sonographer was informed of the objectives of the study and was asked to place the ultrasound probe at the bottom of the participant’s abdomen and acquire a linear ultrasound sweep by moving the probe to the top of the abdomen. This simple training procedure lasted approximately 10 minutes. The participants included in this study were healthy pregnant volun-

teers at 28⁺ weeks of gestation, who were taking part in the INTERGROWTH-21ST project [157, 210, 212]. Data acquisition was covered by appropriate ethics approval.

INTERGROWTH-21ST Study

The International Fetal and Newborn Growth Consortium for the 21 Century (INTERGROWTH-21ST) was designed to develop new standards for defining normal fetal growth, preterm growth and newborn nutritional status. It was conducted in collaboration with eight different sites with geographically diverse populations around the world. The study only included participants with a low risk impairment of fetal growth or risk factors for IUGR (alcohol or smoking) and clinical history (diabetes or chronic illness). Full study demographics can be found from Villar et al. [212]. Moreover, ultrasound video sweeps for this thesis were acquired from some of the participants in the Oxford site of the INTERGROWTH-21ST project. These participants were selected at random and the only selection criteria was a gestational age of 28 weeks or over.

Data Preparation Step

Before carrying out any analysis on the data, a number of standard steps were undertaken to prepare the dataset. Initially, for computational efficiency, all video frames were downsampled to half their original size (from 460×640 to 230×320) using a Gaussian pyramid reduction. Furthermore, each individual video frame was masked based on the ultrasound *field of view*, to mask out and discard any text or comments on the top right and left hand corners of the data as illustrated in Figure 3.2.

3.3 Ultrasound Structure Representation

In this section two popular structural representation techniques for ultrasound images are discussed. Firstly the feature symmetry maps and secondly structured edge

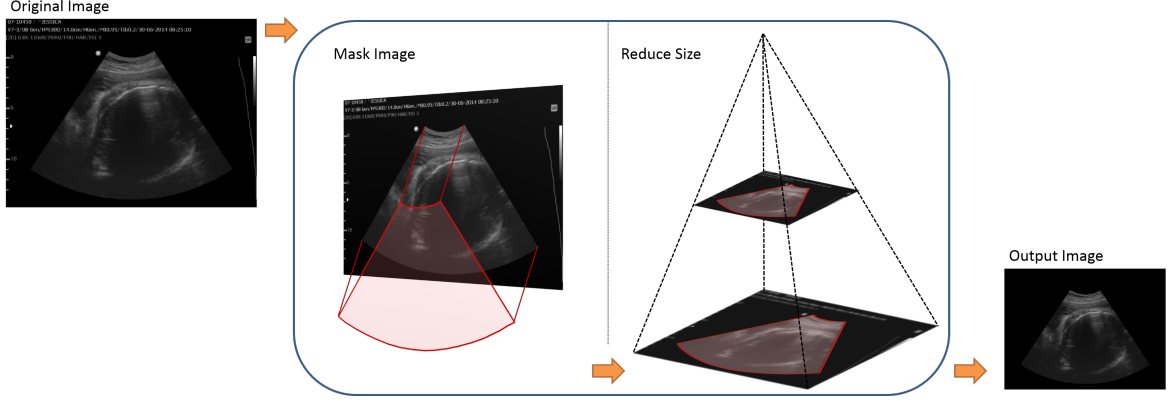


Figure 3.2: *Dataset preparation step.* Original frame (left). Frame is masked to keep the data at the ultrasound’s “field of view” only. This is mainly to remove the text that appears on the top left and right hand sides of the image (centre). Finally the frame is downsampled to half it’s original size for computational efficiency (right).

probability maps based on a structured random forest learning framework.

3.3.1 Contrast Invariant Representation

Pre-processing can be an important step in extracting rich features from ultrasound data, as simple edge detection methods can generally be too sensitive to speckle patterns and shadows in ultrasound images. Local phase based filters as introduced by Koveti [103] have been used for computing contrast invariant features and have shown to be a popular choice for ultrasound images [21, 129, 130, 167, 170]. We investigate the efficacy of this family of filters in this chapter. Moreover, this family of filters are required as a first step in constructing the LP-SIFT features as proposed in [130].

Local phase based filters are constructed using the monogenic signal [62], which provides a degree of robustness to contrast variation and speckle distortion in ultrasound images. Furthermore, it can be used to encode local structural information in an image and has therefore been a potentially suitable and natural first step for ultrasound images as suggested in [167, 170]. The monogenic signal is composed of an even-symmetric bandpass image filter (log-Gabor filter in this case) denoted as $even(x, y)$, and two odd-symmetric filters formed from its Riesz transform denoted as

$odd(x, y)$. Specifically to construct the even (symmetric) component of the monogenic signal, the 2D image $I(x, y)$ is initially convolved with an even isotropic band-pass filter $b(x, y)$,

$$even(x, y) = I_b(x, y) = b(x, y) * I(x, y). \quad (3.1)$$

One of the commonly used such filter types is the isotropic log-Gabor filter $G^{lg}(\omega) = \exp(-\frac{\log^2(\omega/k)}{2\log^2(\sigma_\omega)})$, where k is the centre frequency of the filter and $0 < \sigma_\omega < 1$ is related to the spread of the frequency spectrum in the logarithmic domain. Furthermore to construct the two odd (anti-symmetric) components, the band-passed images are convolved with the Riesz filter to produce the following,

$$\begin{aligned} odd^{(1)}(x, y) &= h^{(1)}(x, y) * I_b(x, y), \\ odd^{(2)}(x, y) &= h^{(2)}(x, y) * I_b(x, y), \end{aligned} \quad (3.2)$$

where the spatial representations of these 2D odd filters are: $h^{(1)}(x, y) = -x/(2\pi(x^2 + y^2)^{\frac{3}{2}})$ and $h^{(2)}(x, y) = -y/(2\pi(x^2 + y^2)^{\frac{3}{2}})$. Therefore the monogenic signal $I_m(x, y)$ of image $I(x, y)$ can be expressed as,

$$I_m(x, y) = \begin{bmatrix} I_b(x, y), & h^{(1)}(x, y) * I_b(x, y), & h^{(2)}(x, y) * I_b(x, y) \end{bmatrix}. \quad (3.3)$$

As fetal ultrasound images typically exhibit non-uniform intensities and shadows (highly dependent on the presence of fatty tissues/bones) which can affect the visual appearance of the same anatomical structures in different scans (for example the fetal skull, heart or abdomen), it is also possible to define a multi-scale feature symmetry map such that

$$FS(x, y) = \sum_s \frac{[|even_s(x, y)| - |odd_s(x, y)| - T_s]}{\sqrt{even_s(x, y)^2 + odd_s(x, y)^2 + \epsilon}}, \quad (3.4)$$

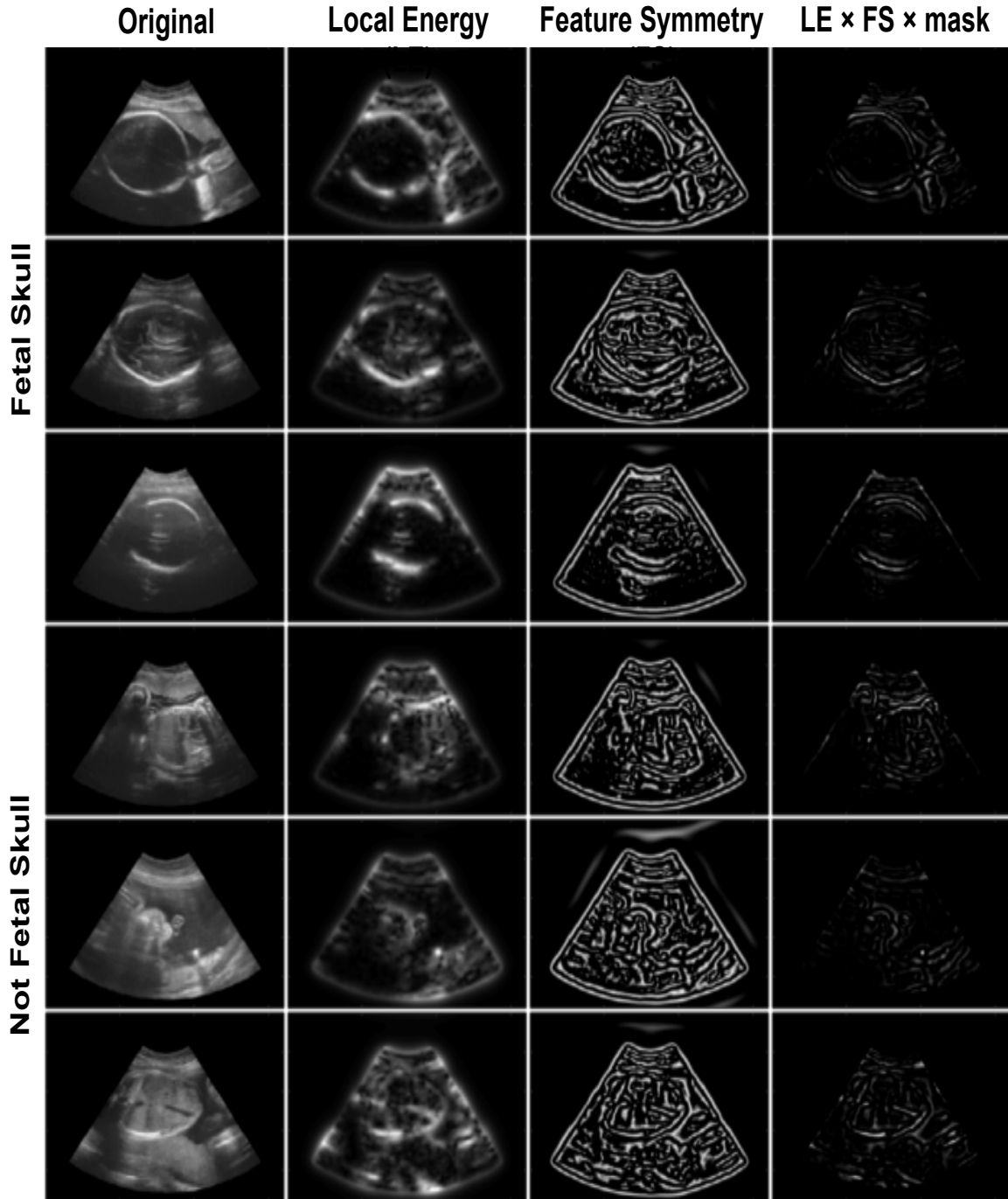


Figure 3.3: *Illustrating intensity invariant features.* From left to right, the original frame, Local Energy, Feature Symmetry, the product of the two are illustrated for six different ultrasound images. The filter's σ value was set to $[28, 30, 32]$ for this plot.

where s is the scale of the band-pass filter, ϵ is a small positive constant to avoid division by zero, $[\cdot]$ denotes replacing negative values with zeros, and T_s is the noise compensation threshold. This threshold can either be calculated empirically or through the following formulation $T_s = \exp\left(\text{mean}(\log(\sqrt{\text{even}_s(x, y)^2 + \text{odd}_s(x, y)^2}))\right)$ as used in previous work [169] to present a scale specific noise compensation term corresponding to the maximum response generated from noise only in the signal. Furthermore the *odd* component above is constructed from the two odd components of the filter such that: $\text{odd} = \sqrt{\text{odd}^{(1)2} + \text{odd}^{(2)2}}$. Figure 3.3 illustrates the corresponding feature symmetry maps for three examples of the fetal skull versus three other examples of various other fetal structures.

Apart from the feature symmetry maps, we also construct and utilise the corresponding local energy (also known as the local amplitude) map for each image. The local energy corresponds to structures which exhibit high amplitude and is formulated as:

$$A_s(x, y) = \sqrt{\text{even}_s(x, y)^2 + \text{odd}_s(x, y)^2}, \quad (3.5)$$

where s is the scale of the band-pass filter. The combination of the feature symmetry and local energy therefore produces a feature symmetry map of locations with high amplitude. This can be a good strategy to discriminate between the soft tissues and hard structures such as the fetal skull. Furthermore, the effect of varying the bandpass scale parameter s from $\sigma = 5$ to $\sigma = 40$ for the product of feature symmetry and local energy is illustrated in Figure 3.4.

A potentially richer representation is obtained by using *signed* feature symmetry [21]. This measure retains the sign of the even filter response, $I_b(x, y)$, in order to discriminate between symmetric structures of high intensity (image *peaks* such as bone and echogenic boundaries) and low intensity (image *troughs* such as the stomach, heart or blood vessels). We can calculate this measure as follows:

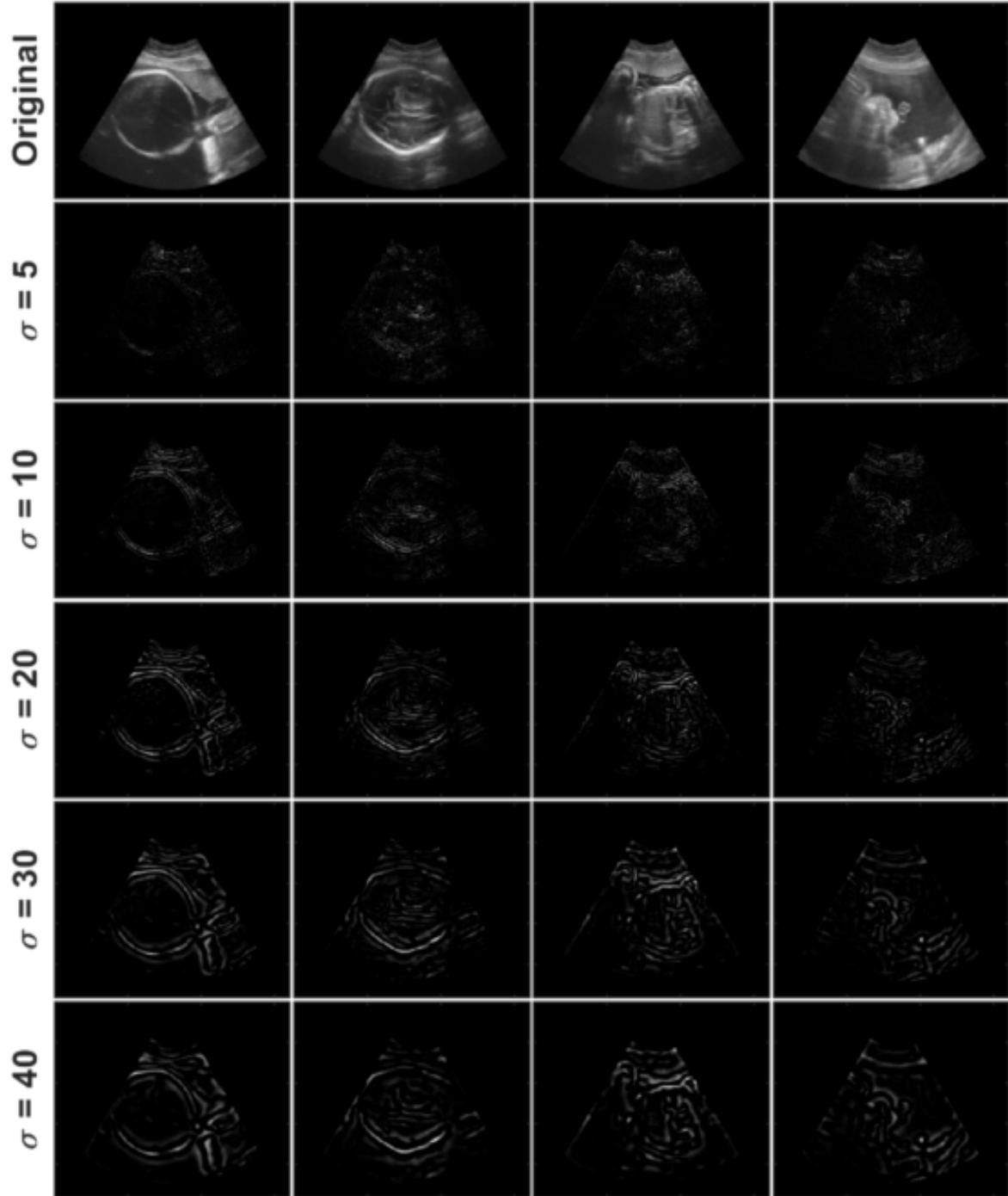


Figure 3.4: *The effect of varying the filter σ value. The original images are illustrated on the top row. The product of the feature symmetry and local energy filters are illustrated for each images, increasing the filter's sigma value from 5 to 40.*

$$SFS(x, y) = \sum_s \frac{[|even_s(x, y)| - |odd_s(x, y)| - T_s]}{\sqrt{even_s(x, y)^2 + odd_s(x, y)^2 + \epsilon}} \cdot \text{sgn}(even_s(x, y)) \quad (3.6)$$

where $\text{sgn}(x)$ is the *signum* function.

The resulting signed feature value lies in the range -1 to 1, where -1 indicates a highly symmetric trough, 0 indicates no symmetry, and 1 indicates a highly symmetric peak. In our experiments we investigate what advantage, if any, this representation has over the standard feature symmetry measure. Figure 3.5 demonstrates the result of the SFS filter on the same set of images for comparison purposes.

3.3.2 Structured Random Forests (SRF) Edge Maps

The local phase feature symmetry maps are fast to compute and highlight key geometric structures but may not capture finer image details. Moreover given a 2D image, edges in local patches are interdependent and contain well-known patterns such as straight lines, parallel lines and Y or T junctions [115, 174]. For problems exhibiting such characteristics, a family of learning approaches known as *structured learning* [101] have been used in the past for example in semantic image labelling [152] and edge detection [51]. Recently structured random forests (SRF) edge detection has been suggested for ultrasound feature enhancement [52] and here we investigate its use for a more accurate structure representation. Figure 3.6 illustrates an example of four different anatomical structures in original gray scale, processed using the FS and LE filters and processed using the SRF edge detector.

Random Decision Forests

SRFs build on the concept of random decision forests (RDF) [42] where a decision tree $f_t(x)$ classifies a sample $x \in X$ through a recursive process of branching left or right until a leaf node is reached. In particular a binary split function is associated with

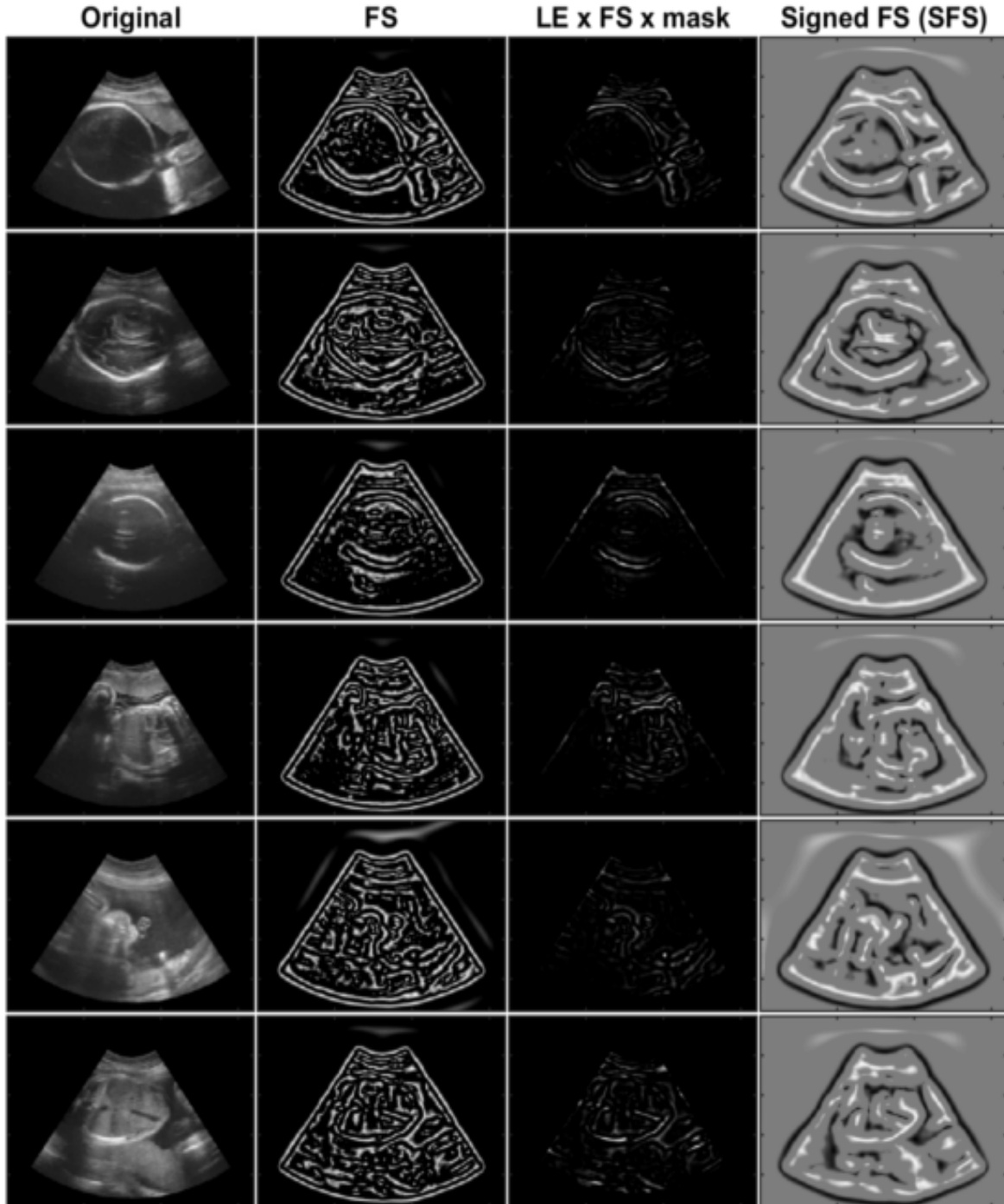


Figure 3.5: *Comparing feature symmetry and signed feature symmetry filters. Original image, feature symmetry response, product of feature symmetry and local energy, and signed feature symmetry as proposed in [21] from left to right respectively with a filter size of [28, 30, 32].*

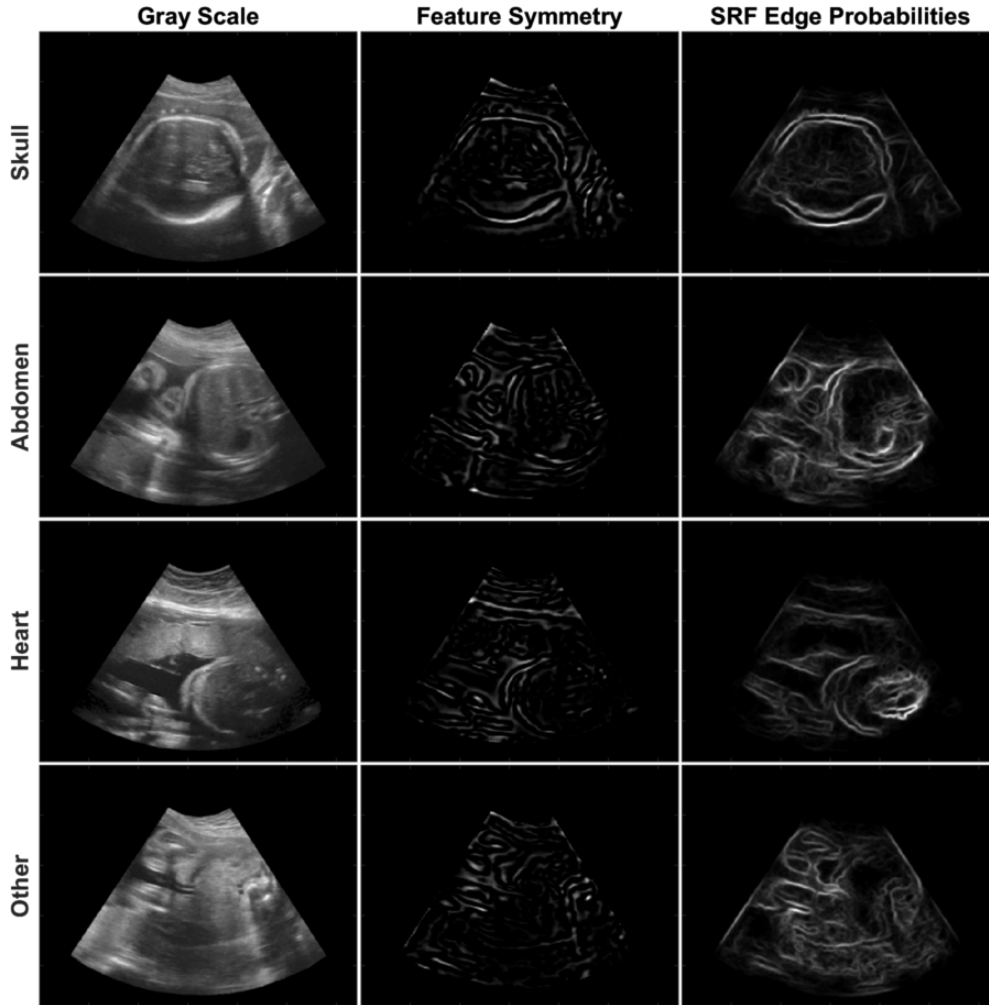


Figure 3.6: *Comparing (feature symmetry $\times$ local energy) and structured random forest edge probability maps. A fetal abdomen (a), and resulting feature symmetry map (b) & edge detection using structured random forest (c)*

each node j of the tree as:

$$h(x, \theta_j) \in \{0, 1\}$$

where θ_j is the function parameter. Here if $h(x, \theta_j) = 0$, the sample x is sent to the left child of node j and if $h(x, \theta_j) = 1$ node j send x right, where this process terminates when the leafnode is reached. Thus the output of the tree for an input x is the prediction stored at the leaf reached for the input x . This can be a target label or a distribution over all the labels Y . Once an input data x traverses a tree, the leaf node that is reached depends only on the input data. This allows for arbitrary

complex outputs, including structured outputs, to be stored at each leaf node [51, 101]. Moreover a decision forest is an ensemble of T independent trees f_t . Given an input sample x , an ensemble model is used to combine the prediction $f_t(x)$ from all the trees into a single output.

Each tree is trained independently to find the optimal parameter θ_j of the split function $h(x, \theta_j)$ for any given node j and training data set $S_j \subset X \times Y$. In order to do this, an information gain criteria needs to be defined as follows;

$$I_j = I(S_j, S_j^L, S_j^R),$$

where $S_j^L = (x, y) \in S_j | h(x, \theta_j) = 0$ and $S_j^R = S_j \setminus S_j^L$, i.e. all the elements of S_j that are not in S_j^L . Here the splitting parameter θ_j is chosen such that the information gain I_j is maximized. This process is repeated at each level of the tree to train each node with the S_j^L data for the left node and S_j^R training the right node until either the tree's maximum depth is reached or the training set size or information gain falls below a predetermined threshold. For a multi-class classification the following standard definition of information gain can be used:

$$I_j = H(S_j) - \sum_{k \in \{L, R\}} \frac{|S_j^k|}{|S_j|} H(S_j^k),$$

where $H(S)$ can denote the Shannon entropy or the Gini impurity.

Structured Random Forests

In the SRFs, x represents an image patch and y encodes the corresponding annotation such as the segmentation mask. Two immediate complications can arise with training random forests with structured labels. Firstly structured output spaces are likely to be very high dimensional, thus scoring candidate splits in this space can be computationally expensive and secondly calculating information gain in such a space is not

well defined. We use the implementation of [51] where all the structured labels $y \in Y$ at a given node are mapped to a discrete set of labels $c \in C$, $C = \{1, \dots, L\}$, where similar structured labels are assigned to the same discrete label c . Given this discrete space, information gain can efficiently be calculated over C . Furthermore to simplify the calculation of information gain, Y is mapped to an intermediate space Z where Euclidean distance can easily be measured.

This mapping is defined in the form

$$\Pi : Y \rightarrow Z$$

such that Euclidean distance is used in Z to approximate dissimilarity of $y \in Y$. This mapping can be further explained using an example. Given a set of 16×16 segmentation masks as the edge detection labels $y \in Y$, the mapping for each mask $z = \Pi(y)$ is defined as a long binary vector that encodes whether every pair of pixel in the patch y belong to the same segment. Given that Z can be high-dimensional, its dimensionality is reduced by firstly sampling only m dimensions of the original vector Z and then applying PCA to further reduce the dimensions. During training, a distinct mapping is randomly selected and applied to training labels Y_j at each node j . Here m has been set to 256.

To compute the information gain given the mapping as described above, the structured labels $y \in Y$ are mapped into a discrete set of labels $c \in C$ where $C = \{1, \dots, k\}$, such that similar structures are given the same label as shown in Figure 3.7. For the ensemble model, a set of n labels $y_1, \dots, y_n$ can be combined into a single prediction. During the training phase, this amounts to associating each label with the corresponding nodes and during testing this amounts to merging the predictions. For this the label whose z_k minimizes the sum of distances to all other z_i is selected. In this thesis, the input to the SRF are 32×32 images patches. For obtaining the discrete label set C given Z , PCA quantization with $k = 2$ was used where the quantization has been

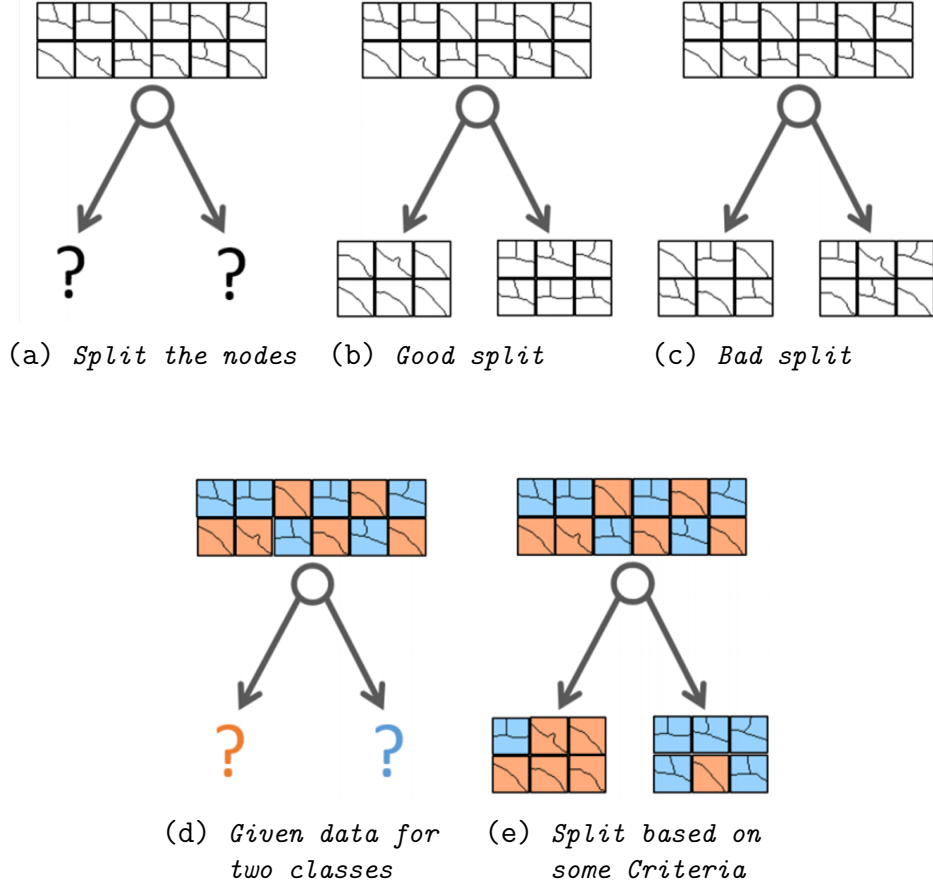


Figure 3.7: Illustration of the decision tree node splits. (a) Given a set of structured labels such as segments, a splitting function must be determined. Intuitively a good split (b) groups similar segments, whereas a bad split (c) does not. In practice we cluster the structured labels into two classes (d). Given the class labels, a standard splitting criterion, such as Gini impurity, may be used (e). Figure adapted from [51]

based on the top $\log_2(k)$ PCA dimensions.

3.4 Features

The features used in this chapter are constructed from a) the raw pixel intensity values, b) feature symmetry filtered representation of the image, and c) the structured edge probability map of the original image. The Low-level features extracted from the aforementioned images are dense SFIT features, which are explained in more detail

below.

3.4.1 LP-SIFT Features

LP-SIFT features are constructed by first applying the feature symmetry and local energy to the original image, followed by constructing dense SIFT features. The SIFT features are an important compartment of the framework and are explained in this section.

SIFT Keypoint Detector

Gradient based features such as SIFT have shown to provide a simple and powerful set of descriptors for a variety of applications. SIFT transforms an image into a large collection of local feature vectors invariant to rotation, scaling or translation of the image. A SIFT key-point - or interest-point - detector is a circular image region with an orientation as illustrated in Figure 3.8a. It is described by a geometric frame of four parameters including a centre coordinates x and y , the radius of the region referred to as the *scale*, and its *orientation*. Originally SIFT was designed to detect image structures which resemble *blobs*, and by searching for blobs at multiple scales and positions, it showed to be invariant to translation, rotations, and re-scaling of an image. Searching for keypoints at multiples scales is achieved through a *Gaussian scale space*, which in practice equates to multiple versions of the image smoothed at various scales or more simply reducing the image resolutions multiple times.

SIFT Descriptor

A SIFT descriptor is a 3D spatial histogram of local oriented gradients characterising the appearance of a SIFT, either around an identified interest points as in Figure 3.8b or on a dense grid as in Figure 3.9. The gradient at each pixel is regarded as a sample of a three-dimensional feature vector, constructed from the pixel location and the gradient orientation. The gradient norm is used to weight these samples and

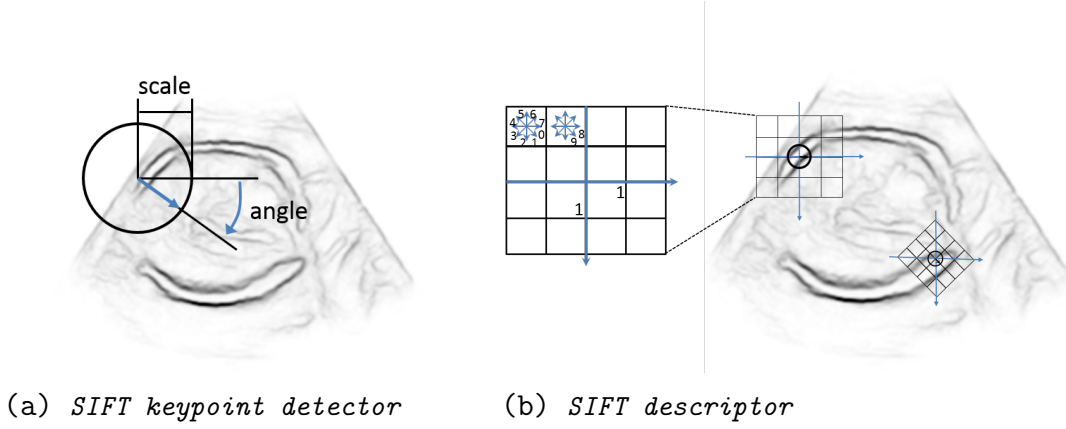


Figure 3.8: *An example SIFT detection 3.8a, and descriptor 3.8b. The Detection selects a point on the image and the descriptor contains the gradient information for the specific local region.*

accumulate them in a 3D histogram to form the SFIT descriptor. Finally gradients further away from the keypoint center are given less importance through an additional Gaussian weighting function. The conventional SIFT descriptor is a 3D histogram consisting of $8 \times 4 \times 4 = 128$ bins as illustrated in Figure 3.8b. Moreover, Figure 3.10 illustrates 20 random SIFT descriptors at various scales and orientations which have been identified as interest points on a 2D ultrasound image of the fetal skull.

A drawback of using keypoint detectors to identify interesting blobs is that they can be susceptible to imaging noise, speckle patterns, signal attenuation and shadows in ultrasound images as signal attenuation could have the same blob effect in an ultrasound image. Dense feature extraction has become an essential part of many state-of-art image classification methods, where low-level features such as SIFT are computed on a dense grid (e.g. every n^{th} pixel) rather than sparsely using an interest point detector on an image [96]. Given the characteristics of medical ultrasound images, where the level of shadows, speckle patterns and attenuation vary between images and also depend on the anatomical object being scanned (e.g. fetal heart, abdomen), dense feature extraction is utilised here.

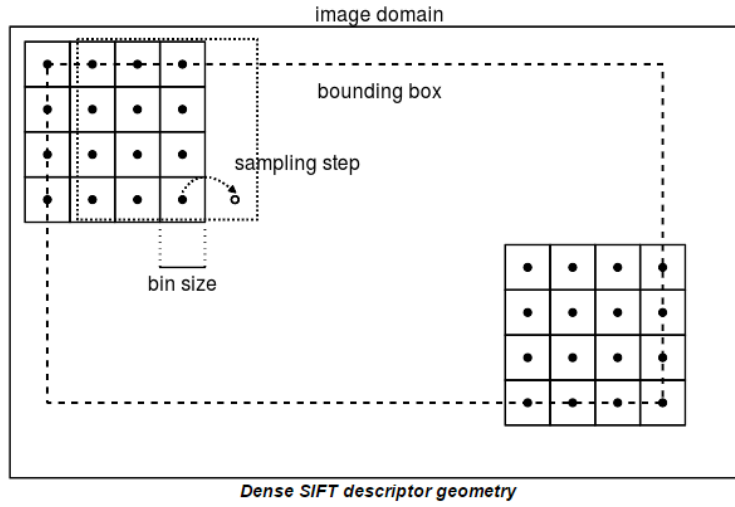


Figure 3.9: *The schematic of computing dense SIFT. A dense SIFT descriptor calculated densely on the entire image. Bin size indicates the size of the descriptor, and the sampling step indicates the density in which the descriptor is created on the image.*

LP-SIFT Finally to construct the LP-SIFT features, the gray scale ultrasound images are initially filtered as described in Section 3.3.1 to obtain enhanced images, followed by computing the SIFT feature densely on the filtered images as described in this section. Encoding these features is covered in the next section.

3.5 Feature Encoding

3.5.1 Bag of Visual Words Encoding

The concept of visual vocabulary has been inspired from the text retrieval community [9] and enables indexing for local image features. Text documents contain a distribution of words and can therefore be efficiently summarised by their word count, also known as bag of words. In order to be able to apply this methodology to the high-dimensional image domain, a quantisation step needs to be performed on the feature space of local image descriptors. [190].

Therefore the standard framework for constructing a bag of visual words consists of

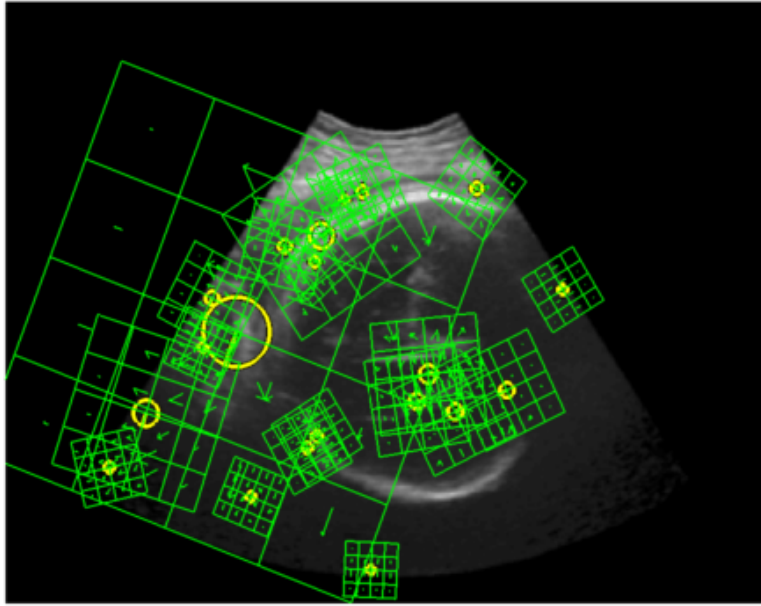


Figure 3.10: *SIFT interest points on an ultrasound image.* SIFT keypoint descriptors at various scales and orientations are detected. Only 20 keypoints have been selected and plotted at random for illustration purposes.

initially collecting large samples of low-level features such as dense SIFT from training images, which in practice can be a portion of the overall training dataset. Secondly learn a *visual vocabulary* from the extracted features. The *visual words* of the vocabulary are constructed by quantising the feature space (e.g. using k-means), where the cluster centres are the visual words and the number of clusters is defined by the user. Once the vocabulary of visual words is constructed, novel images can be described in terms of the frequencies of visual words of the learned vocabulary. This is schematically illustrated in Figure. 3.11.

The steps for constructing the BoVW in our framework are illustrated schematically in Figure 3.12. Once an ultrasound video is acquired following a simple and repeatable protocol, such as in Figure 3.1, each video frame can be processed using any of the pre-processing steps described before in Section 3.3 to enhance acoustically significant structures in images. This step proves to be vital during the feature extraction process when limited amounts of data are available. Dense SIFT features are then extracted

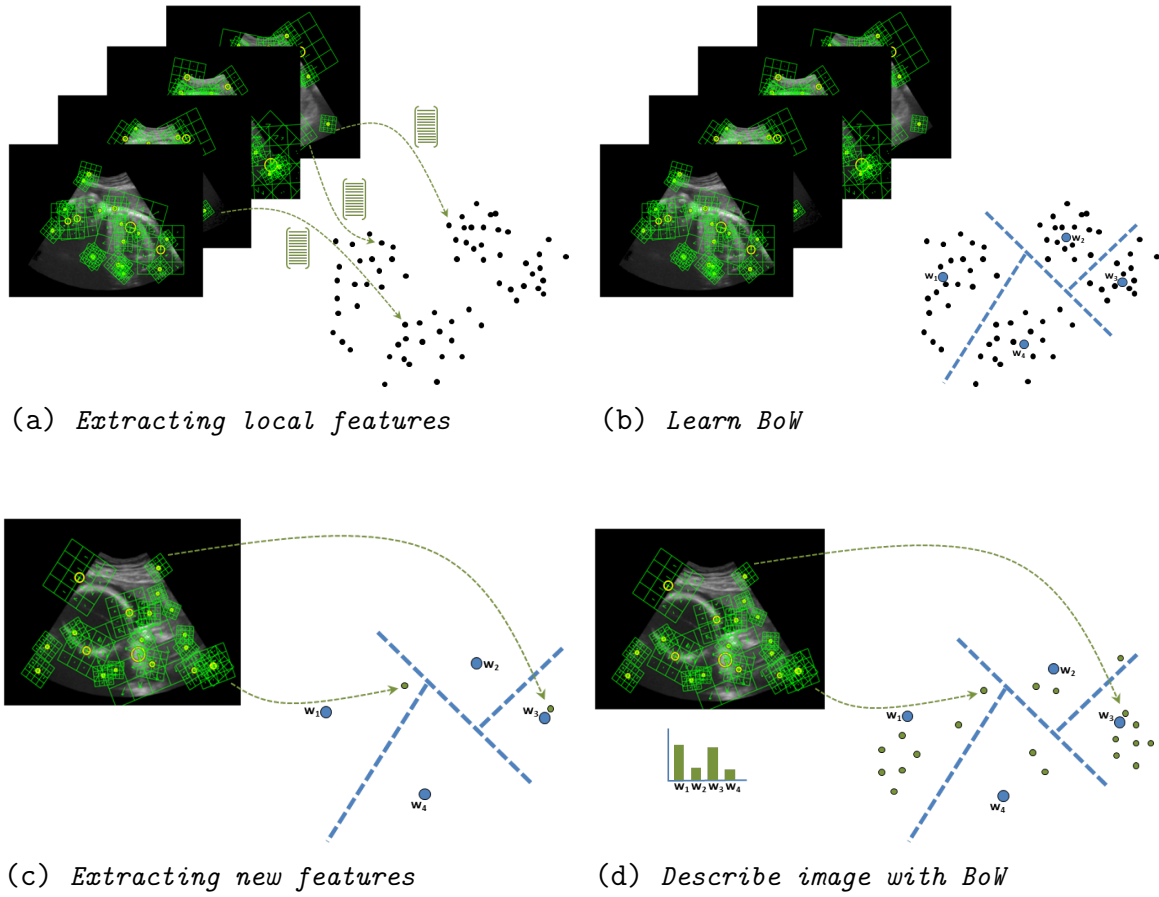


Figure 3.11: Bag of words construction schematic. 3.11a Low-level features (e.g. SIFT) are computed from a large portion of training images to populate the feature space. The SIFT feature computed on the images are denoted with green colour, and the black dot is the same features in some high-dimensional feature space. 3.11b The features are quantised into W_N clusters to construct a visual vocabulary of words. The blue dots denote the visual words (the cluster centres). 3.11c Given a test image low-level features are extracted and the nearest visual word is identified for each feature. 3.11d A histogram of visual words is used to summarise the test image by counting the number of visual words that occur in the image.

from the filtered images. Each frame (2D image) is then encoded as a spatial histogram using the vocabulary of visual words learnt during the training phase. For classification of the data, an SVM classifier with a homogeneous χ^2 kernel map [207] is then trained and evaluated on unseen data.

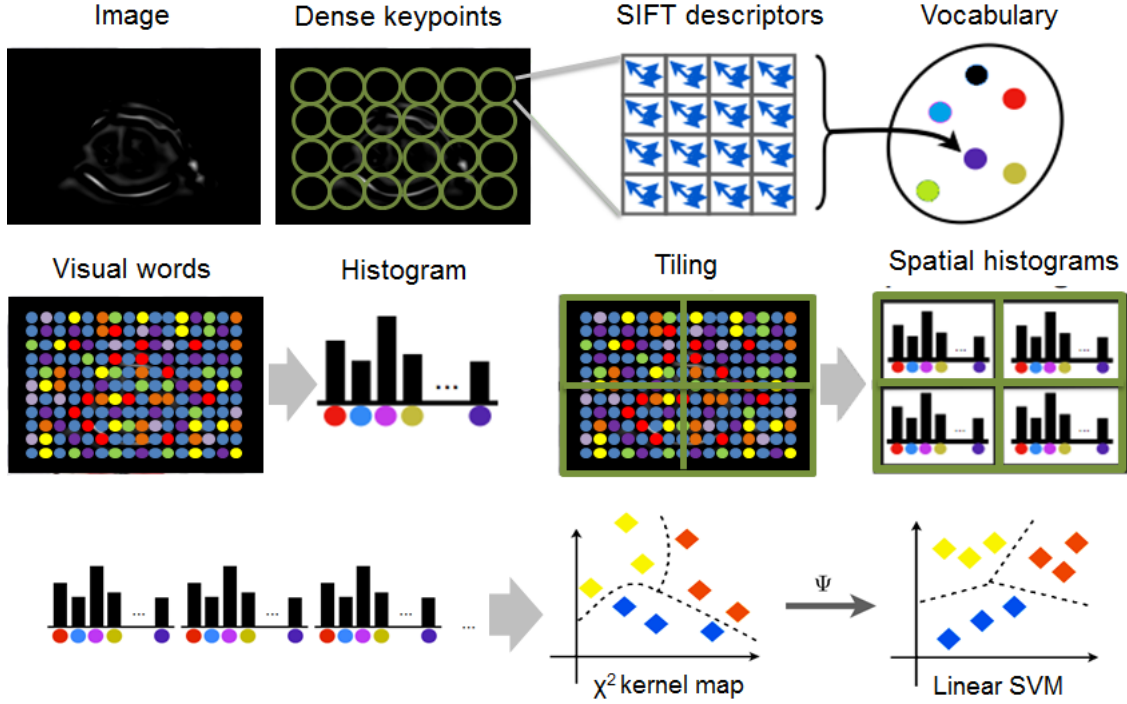


Figure 3.12: *Schematic of steps involved in the BoVW framework. Initially SIFT features are calculated from the (local phase processed) images on a dense grid and vectorised into visual words. Once the visual vocabulary is constructed, training and test images can be encoded by keeping a count of the frequency of each visual word through a histogram. If spatial tiling is used, the encoding can be carried out for each tile of the spatial filter.*

3.5.2 Fisher Vector Encoding

The Fisher Vector (FV) [160, 161] encoding aims to reduce the loss of information caused by the vector quantization step in BoVW. Figure 3.13 illustrates the additional information that is used in FV encoding compared to the BoVW. Moreover, Figure 3.14 illustrates schematically the FV encoding process used here.

The FV encoding approach works by aggregating a large set of low-level feature vectors, into a high-dimensional space. A common approach, which is also utilised here is to initially reduce the SIFT feature dimension using PCA and then square root the features vectors (also known as RootSIFT), as it has shown to improve classification results [8]. A parametric generative model such as the Gaussian Mixture Model (GMM) is then fitted to the features. and the derivatives of the log-likelihood of the

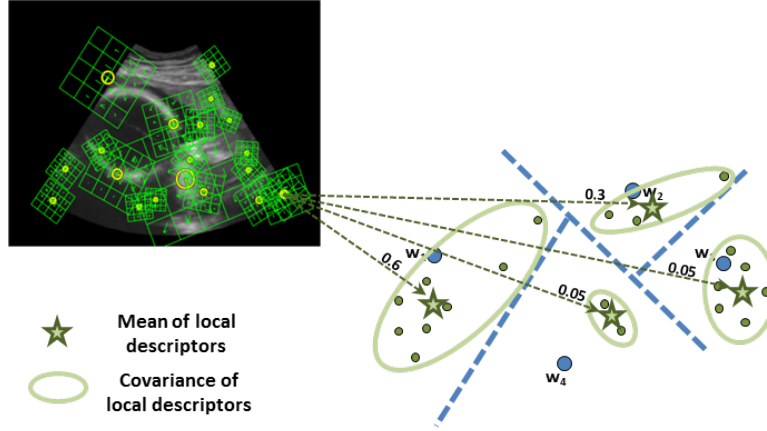


Figure 3.13: Fisher Vector assignment strategy. Compared to the Bag of Visual Word, soft assignment strategy is used which helps reduce information loss during the quantisation process.

model with respect to its parameters are encoded. First and second order differences between the dense features and each of the GMM centres can then be estimated.

More formally, given $I = (x_1, \dots, x_N)$, a set of D dimensional SIFT feature vectors extracted from an image, and $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, \dots, K)$ the parameters of a Gaussian Mixture Model fitting the distribution of the descriptors, (where K is the number of multivariate Gaussian distributions, μ_k , Σ_k and π_k are the mean, variance and the prior probability of each Gaussian distribution k), the GMM associates each vector x_i to a mode k in the mixture with a strength given by the posterior probability such that,

$$q_{ik} = \frac{\exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right]}{\sum_{t=1}^K \exp\left[-\frac{1}{2}(x_i - \mu_t)^T \Sigma_k^{-1}(x_i - \mu_t)\right]}. \quad (3.7)$$

For each mode k the mean and covariance deviations vectors are defined such that,

$$u_{jk} = \frac{1}{N\sqrt{\pi k}} \sum_{i=1}^N q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}}, \quad (3.8)$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi k}} \sum_{i=1}^N q_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right], \quad (3.9)$$

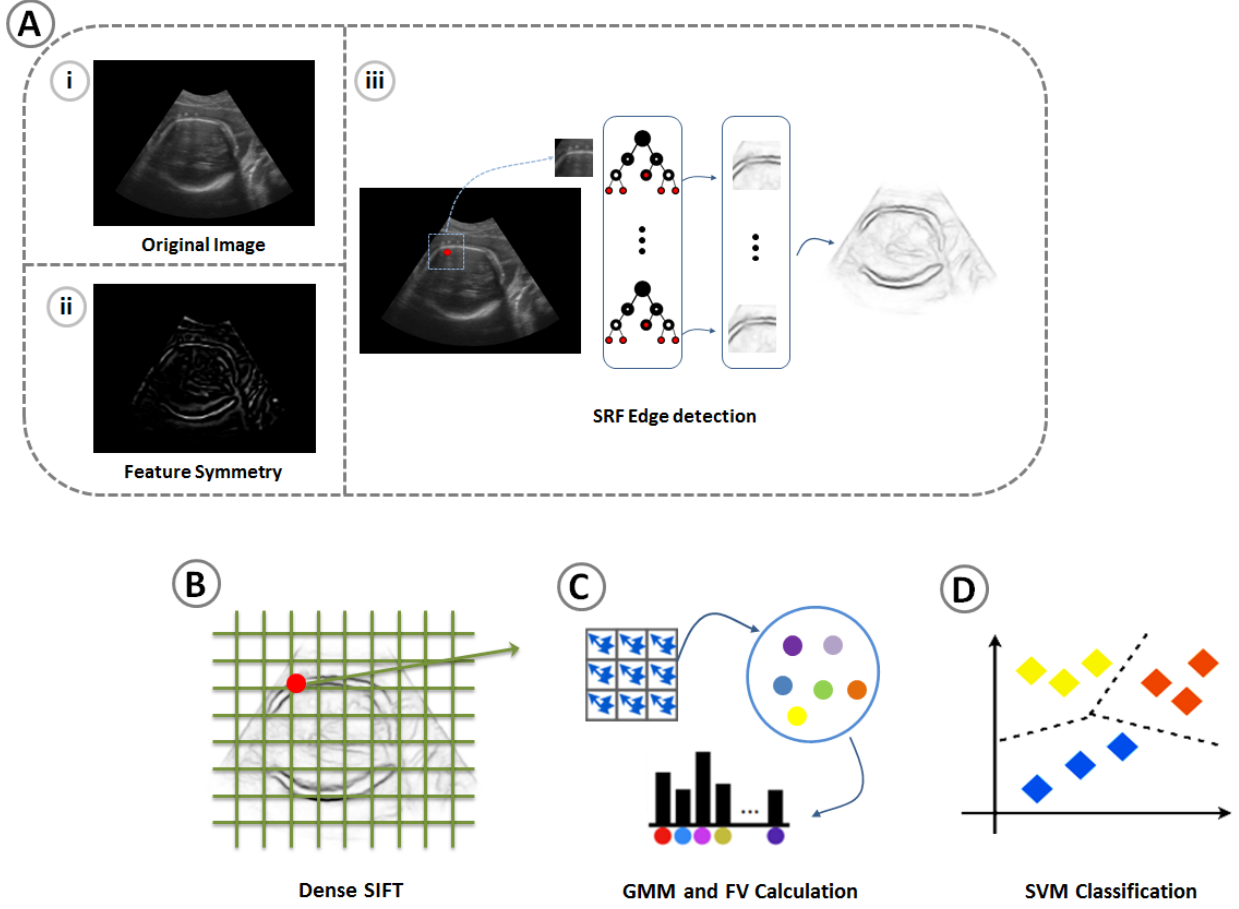


Figure 3.14: Encoding method overview. Fisher vector encoding is computed on dense SIFT features followed by the classification using an SVM classifier. Input images can either be the original images(1), feature symmetry maps(2) or the SRF edge maps(3).

where $j = 1, \dots, D$ and represents the vector dimensions. The Fisher vector Φ of image I is then constructed by stacking the vectors u_k and v_k for each of the K modes in the Gaussian mixtures,

$$\Phi(I) = [u_1^\top, v_1^\top, \dots, u_K^\top, v_K^\top]^\top \quad (3.10)$$

Moreover, as suggested by Perronnin et al. [161] non-linear additive kernels are applied to each dimension of the fisher vector $\Phi(I)$ and the fisher vector is l_2 normalized to improve classification accuracy. The VLFeat toolbox [205] was used for computing the features and evaluating accuracy in the experiments reported below.

3.6 Validation Experiments

The experiments are designed to validate the proposed frameworks explained earlier in this chapter and are evaluated on two datasets of various sizes. Both datasets contain four classes of video frames, namely the *abdomen*, *heart*, *skull* and *other structures*. "Other structures" class includes any video frame that did not fall into the other three categories. The smaller dataset consists of a total of 60 ultrasound videos, obtained from 30 participants (two videos from each participant), whereas the larger dataset consists of 86 clinical ultrasound videos. This dataset included the first 60 videos and in addition an additional 26 videos obtained from 26 different participants. Moreover, the merits of preprocessing ultrasound images prior to classification can be assessed on the two datasets. It is worth noting that the filtering algorithms may have introduced artefactual edges in some instances, which may effect the classification results. However, there is no clear evidence from the results that it led to the performance of the classification algorithms being compromised.

Experiments on the Smaller Dataset:

The experiments were designed to allow a systematic comparison between dense SIFT features computed on the gray scale pixel intensities and the LP-SIFT features. The BoVW encoding strategy was applied and classification performance was analysed. Three binary classification experiments were designed. The first experiment was carried out to measure how well the frames that exhibit the fetal skull were separated from the ones that did not (*Fetal Skull versus All Other Structures*), where four fifths of the data which constitutes to 1510 images are randomly selected for training and the remaining one fifth which constitutes to 378 images is randomly selected for testing. For the second experiment the same procedure was repeated but for the fetal heart class (*Fetal Heart versus All Other Structures*), where 148 images are randomly selected for training and 38 for testing. The third experiment was to compare classifi-

cation accuracy between the frames that exhibit the fetal skull and those that exhibit the abdominal structures (*Fetal Skull versus Fetal Abdomen*), where again 148 images are randomly selected for training and 38 for testing. Given that the fetal skull and abdominal structures demonstrate similar oval geometric characteristics, this experiment was designed to investigate how well the two structures can be distinguished with the two features. Finally, the fourth experiment evaluated the accuracy for a four class classification problem using the proposed framework, where 296 images are randomly selected for training and 76 for testing. All experiments were performed using a 5-fold cross-validation where the data has been trained on 4/5 of the data and accuracy is evaluated on 1/5 of the data. The results are summarised in Table 3.3.

Two-Class Classification: Results from this experiment are summarised in Table 3.1 for fetal skull verses other structures, Table 3.2 for fetal skull versus abdominal structures and Table 3.3 for fetal heart versus other structures. Generally a higher classification accuracy is achieved when using the proposed LP-SIFT features and a higher performance is obtained as the number of codewords increase on the smaller dataset.

For the fetal skull versus other structures, the highest mean accuracy obtained using the LP-SIFT features is 95.87% (200 codewords) compared to 86.51% (300 codewords) using the SIFT features. Similarly for the fetal skull versus the abdomen, the highest mean accuracy was obtained using the LP-SIFT features of 97.90% (100 codewords). For the SIFT features it was 90.53% using 50 codewords. The fetal heart is the exception here where the best result is achieved using the SIFT features and 200 codewords with an accuracy of 92.36% compared to 90.00% with 300 codewords for the LP-SIFT features.

Four-Class Classification: Table 3.4 summarises classification results for the four class classification of fetal skull, abdomen, heart and other structure classes. As can

Table 3.1: *Two-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from the Skull and Other Structures classes, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes.*

Codewords	Features	min/max (%)	Mean (%)
		Ave. Accuracy	Ave. Accuracy
30	LP-SIFT	91.0/95.0	93.10
	SIFT	75.9/80.4	78.15
50	LP-SIFT	93.9/94.2	93.87
	SIFT	81.0/84.9	83.01
100	LP-SIFT	95.0/96.3	95.82
	SIFT	82.8/86.5	84.97
200	LP-SIFT	95.2/97.4	95.87
	SIFT	83.1/86.0	84.60
300	LP-SIFT	94.2/97.4	95.61
	SIFT	85.0/88.6	86.51

Table 3.2: *Two-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from the Skull and Abdomen classes, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes. (b)*

Codewords	Features	min/max (%)	Mean (%)
		Ave. Accuracy	Ave. Accuracy
30	LP-SIFT	92.1/100.0	94.98
	SIFT	73.7/86.6	77.89
50	LP-SIFT	92.1/100.0	96.84
	SIFT	89.5/92.1	90.53
100	LP-SIFT	94.7/100.0	97.90
	SIFT	79.0/97.4	89.47
200	LP-SIFT	94.7/100.0	97.37
	SIFT	84.2/92.1	86.84
300	LP-SIFT	94.7/100.0	96.84
	SIFT	84.2/92.1	88.42

be seen the highest mean classification has been achieved using the LP-SIFT features (78.95%) using 200 codewords. For the SIFT features, the best results have been achieved using 100 codewords with a mean accuracy of 66.05%.

Table 3.3: Two-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from the Heart and Other Structures classes, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes (c).

Codewords	Features	min/max (%)	Mean (%)
		Ave. Accuracy	Ave. Accuracy
30	LP-SIFT	85.3/92.7	89.61
	SIFT	88.2/92.7	90.87
50	LP-SIFT	88.2/92.7	89.71
	SIFT	72.1/92.7	87.35
100	LP-SIFT	85.3/92.7	89.70
	SIFT	85.3/92.7	89.41
200	LP-SIFT	83.8/92.7	89.41
	SIFT	91.2/92.7	92.36
300	LP-SIFT	83.8/92.7	90.00
	SIFT	88.2/92.7	91.77

Table 3.4: Four-class classification results on the smaller dataset. Comparing LP-SIFT and dense SIFT features extracted from all four classes of Skull, Abdomen, Heart and Other Structures, encoded via the BoVW framework. Minimum, maximum and mean accuracy are reported over a 5 fold cross validation for codebooks of various sizes (d).

Codewords	Features	min/max (%)	Mean (%)
		Ave. Accuracy	Ave. Accuracy
30	LP-SIFT	54.0/86.8	72.63
	SIFT	54.0/73.7	64.74
50	LP-SIFT	64.5/80.3	70.51
	SIFT	58.0/65.8	62.10
100	LP-SIFT	71.1/81.6	77.37
	SIFT	51.3/75.0	66.05
200	LP-SIFT	72.4/85.5	78.95
	SIFT	56.6/71.1	65.00
300	LP-SIFT	65.9/81.6	73.10
	SIFT	51.3/79.0	64.74

Classifier Performance In order to visually assess the classification performance for the different classes on the test data, the classifier scores for each image have been ranked in descending order, where the classifier confidence is highest for the first item in this list and lowest for the last image. Figures 3.15 and 3.16 illustrate the top nine images with the highest scores in each class for classification results using a codebook of 50 and 100 visual words respectively, where figures (a) and (b) correspond to a binary

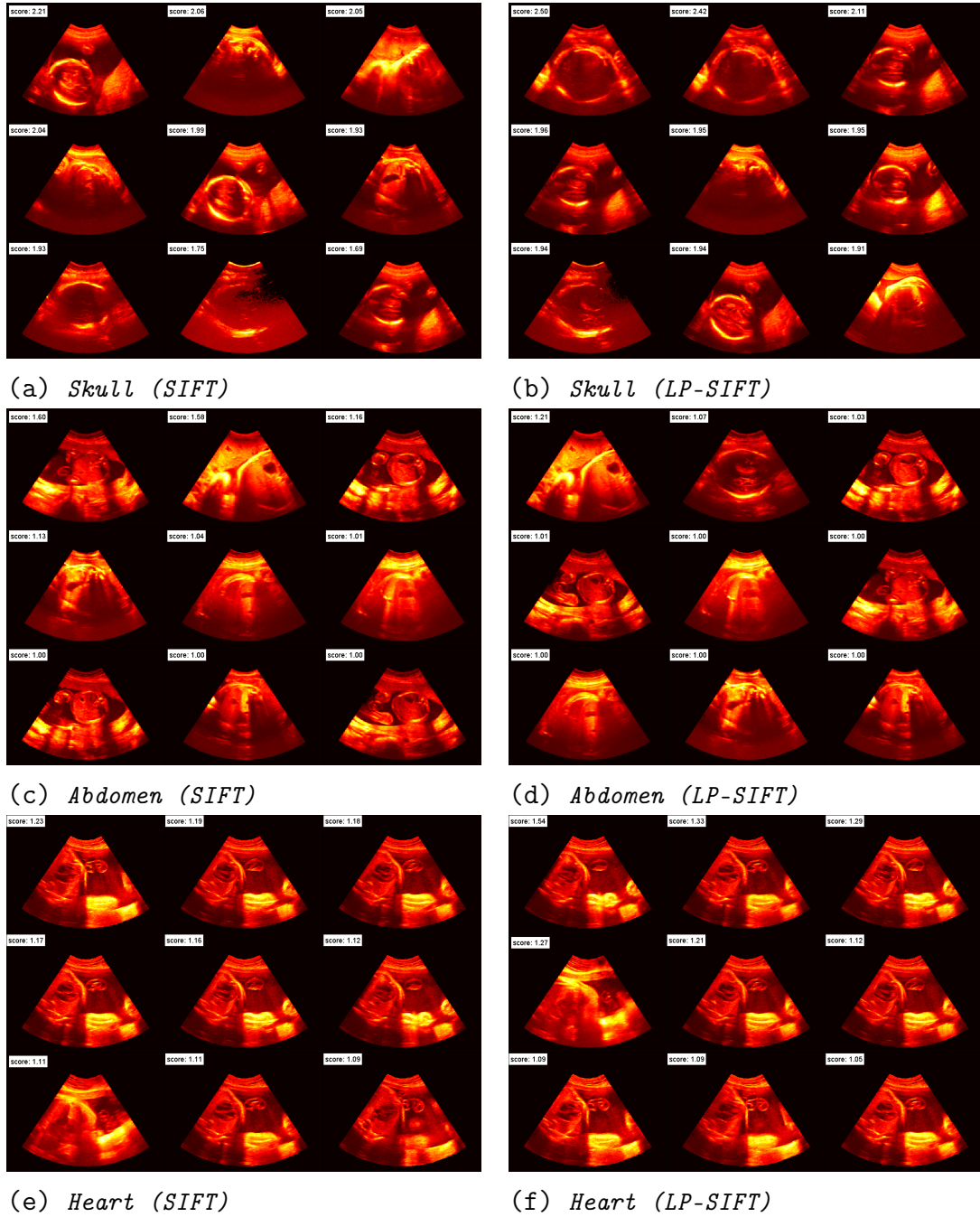


Figure 3.15: LP-SIFT Results. Top nine ranked images based on the classifier scores on test data for Skull vs not Skull (a,b), Abdomen vs not Abdomen (c,d) and Heart vs not Heart (e,f). SIFT features are used in (a) & (c) & (e) and LP-SIFT in (b) & (d) & (f) with a codebook of 50 words.

classification for fetal skull versus other structures, (c) and (d) for the fetal abdominal structures versus other structures and (e) and (f) show classification results for fetal heart versus other structures, using the two features discussed in this paper.

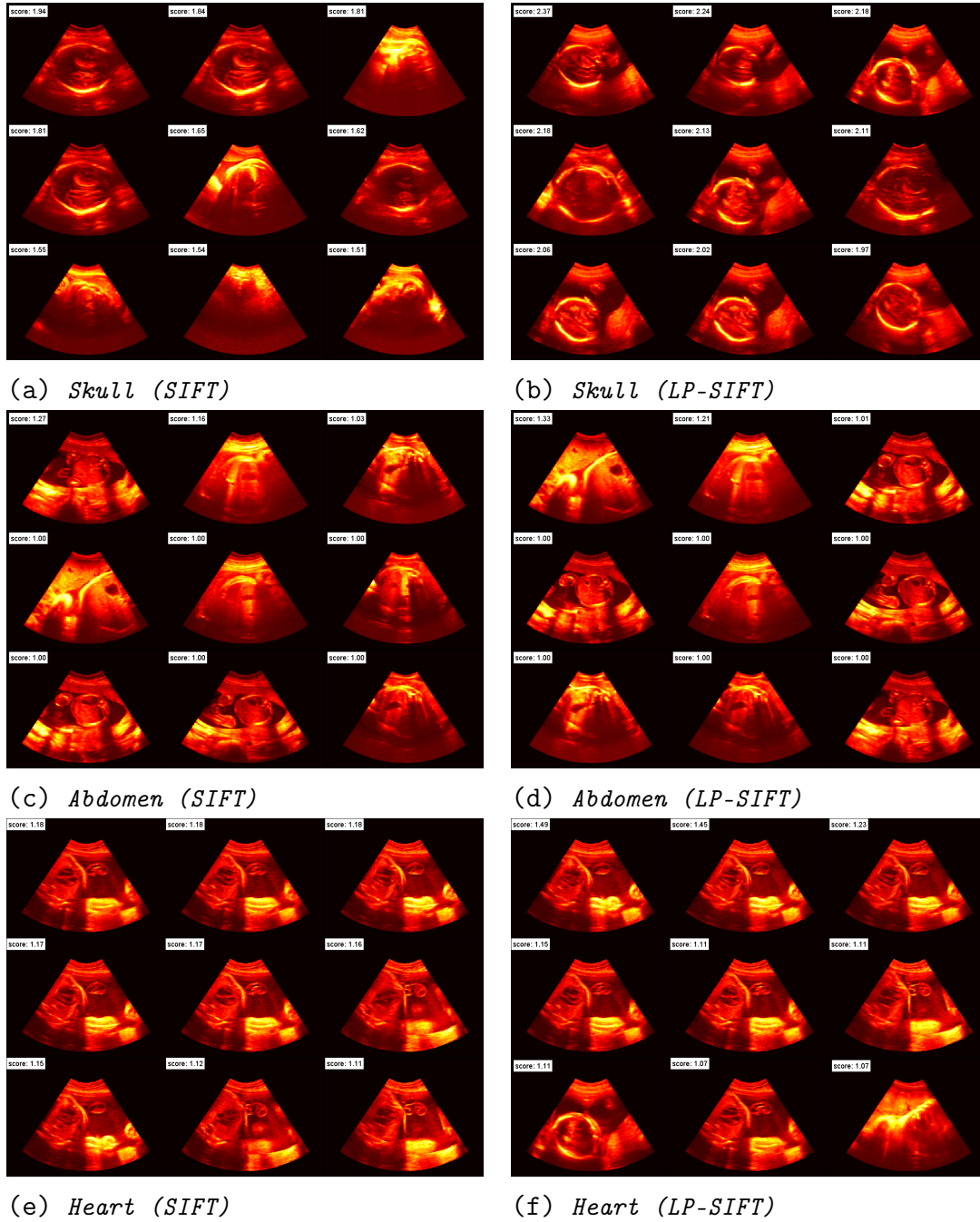


Figure 3.16: LP-SIFT Results. Top nine ranked images based on the classifier scores on test data for Skull vs not Skull (a,b), Abdomen vs not Abdomen (c,d) and Heart vs not Heart (e,f). SIFT features are used in (a) & (c) & (e) and LP-SIFT in (b) & (d) & (f) with a codebook of 100 words.

It is interesting to note that using a larger codebook of visual words using the LP-SIFT features has allowed for a diverse set of images to be selected. For example the top nine fetal skulls selected using LP-SIFT with 50 codewords are all of similar

structures as shown in Figure 3.15 whereas the same setting with 100 visual words has allowed for a wide variety of head structures to be selected as show in Figure 3.16. Furthermore although generally increasing the number of visual words has resulted in a higher classification accuracy, the performance gain was found to be only marginal after 100 words. It is also important to note that classification results for the fetal heart are better using the SIFT features. This may be due to the fact that fetal heart is relatively small compared to the skull and abdomen and the pre-processing filters might have removed necessary data.

Experiments on the Larger Dataset:

Three experiments were designed to evaluate the classification accuracy of the proposed framework. The low-level feature computed in each of these experiments were: a) LP-SIFT feature maps, b) SRF edge maps, and c) dense SIFT features without any pre-processing. In each experiment, the model was trained on 850 randomly selected images from each class (3400 frames over all classes) and tested on 350 randomly selected unseen images from each class (1400 frames over all classes). The experiment was repeated five times, each time with a different set of images selected for training and testing. For each experiment the accuracy and mean average precision (mAP) were calculated and their average over the five repetitions is reported. Precision is the positive predictive value and is defined as the total number true positives divided by the sum of all true positives and false positives detected by the classifier. Moreover, recall describes the true positive rate and is defined as the total number of true positives divided by total number of true positives and false negatives. Average precision can be described as follows: given a ranked set of results for a query, precision is computed at every correctly returned result (image) and then the average is computed. This is done over the results for all the four classes to compute mAP.

Table 3.5: *Classification results on the larger dataset.***(a)** *Results using the dataset used in [130]*

	min/max	Mean
	Accuracy (%)	Accuracy (%)
BoVW _{FS}	71.1/76.3	73.95
BoVW _{SRF}	75.0/86.8	80.79
BoVW _O	54.0/76.3	67.37

(b) *Classification using BoVW & FV encoding.*

	min/max	Mean	min/max	Mean
	Accuracy (%)	Accuracy (%)	mAP	mAP (%)
BoVW _{FS}	78.4/80.9	80.03	83.9/86.2	85.16
FV _{FS}	97.7/98.7	98.10	99.3/99.8	99.49
BoVW _{SRF}	83.6/84.7	84.07	88.1/89.8	89.12
FV _{SRF}	97.7/98.1	97.88	99.5/99.6	99.52
BoVW _O	86.0/88.8	87.12	90.8/92.1	91.50
FV _O	98.5/99.4	98.90	99.8/99.9	99.85

FV versus BoVW encoding The results are summarised in Table 3.5b, where bold font face indicates highest accuracy achieved. The mAP plot for one of the repetitions is illustrated in Figure 3.17.

Initially we have followed [130] and used the same dataset to perform a four class classification on feature symmetry (FS) maps. The mean classification accuracy and the mean mAP achieved are 80.03% and 85.16% respectively using a BoVW approach and 98.10% and 99.49% using the FV. As an alternative to FS, we used Structured Random Forests (SRF) as outlined in Section 3.3.2. For this experiment, the mean classification accuracy and the mAP achieved are 84.07% and 89.12% respectively using a BoVW approach and 97.88% and 99.52% using the FV. Finally classification was performed and evaluated on the original grey scale frames. The mean classification accuracy and the mAP achieved are 87.12% and 91.50% respectively using a BoVW encoding and 98.90% and 99.85% using the FV encoding framework.

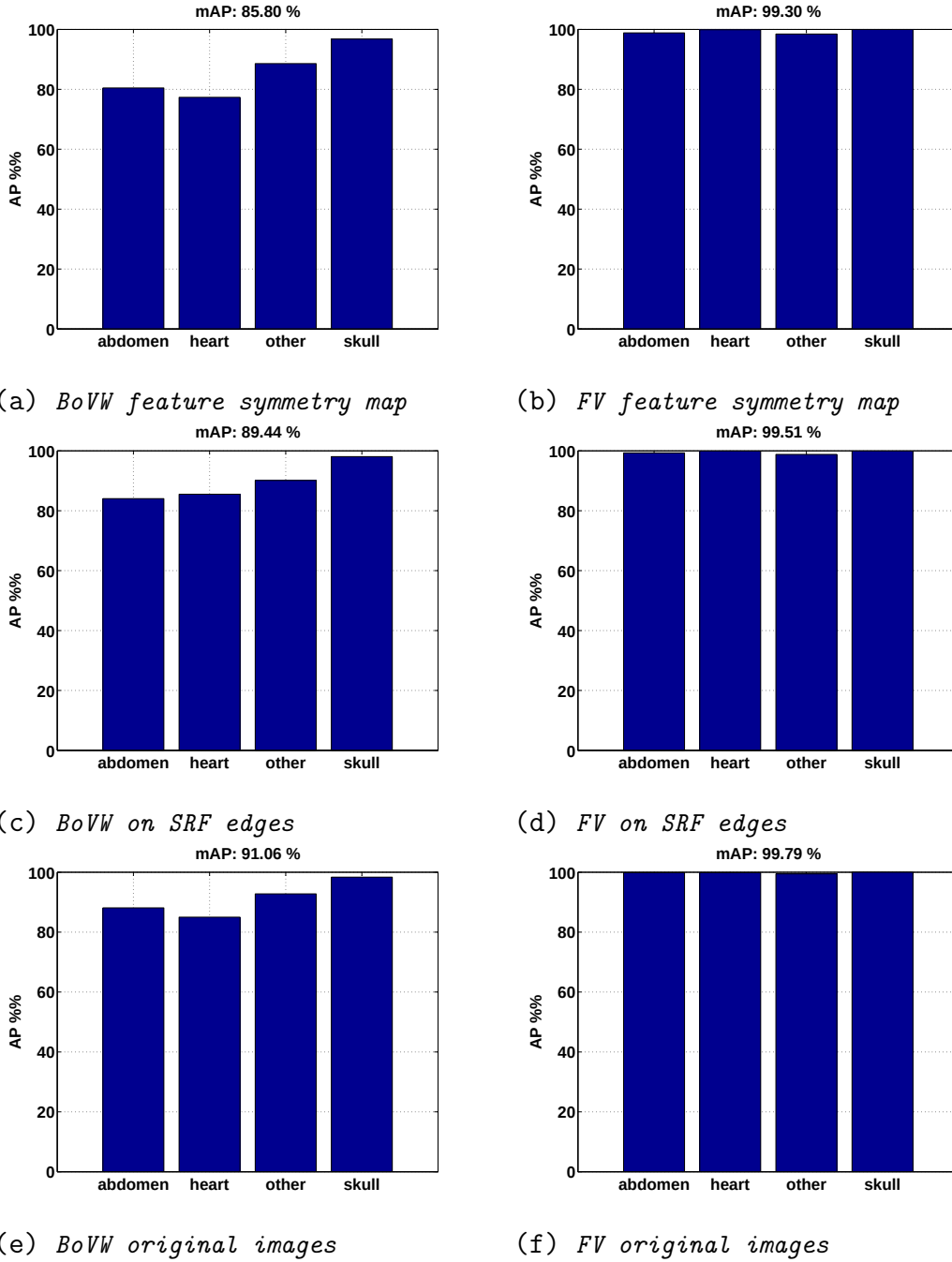


Figure 3.17: Comparing the mAP achieved using the BoVW and FV.

3.7 Discussion and Conclusion

Ultrasound image acquisition and interpretation can be a challenging task for non-experts. We have demonstrated that standardised, and possibly non-conventional scanning protocols can be coupled with image analysis methods to extract useful in-

formation from ultrasound videos, which can be used for making simple diagnostic decision. The results suggest that the Fisher vector encoding of dense RootSIFT features illustrates highest classification accuracy. Moreover, preprocessing can only add value when the training datasets are small and the variation of the data may not be captured. Finally, learning the edge representations using structured random forests as a preprocessing step, can improve the classification performance compared to the widely used feature symmetry filters in 2D ultrasound data. In this section mean average accuracy is reported for each experiment. This indicates the average accuracy for the five-fold cross validation, averaged for all the classes involved. Although showing the mean classification accuracy for each class could have been reported, given that the conducted experiments involve the comparison of the skull class versus the other, skull class versus abdomen, and also a four class classification, the individual class accuracies have been omitted for clarity. Moreover, statistical significance is less often reported in such settings. Such information could be useful for a more clinical study, and the statistical significance could then indeed be used to compare the results with the ones in this thesis to identify whether the gain or loss in accuracy has been statistically significant.

Chapter 4

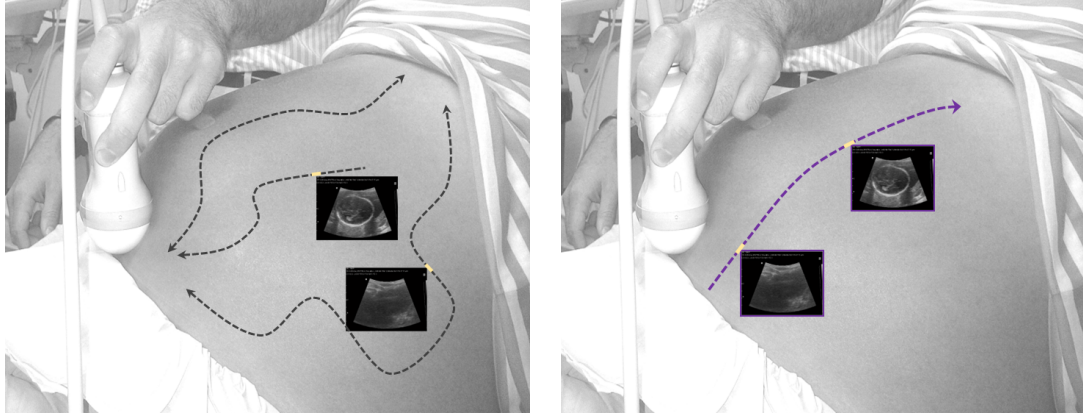
A Dynamical Model for Detecting Sequences of Interest in an Ultrasound Video

In this chapter we address the general problem of identifying video segments that contain structures of interest based on their motion patterns. Furthermore, we demonstrate that this approach can be used in conjunction with predefined free-hand sweeps described in Chapter 2 (see Figure 4.1b), to automatically identify the fetal presentation during the third trimester of pregnancy.

An early version of the work presented in this chapter has been reported in:

Maraci MA, Napolitano R, Papageorgiou A, Noble JA. Searching for Structures of Interest in an Ultrasound Video Sequence. Machine Learning in Medical Imaging workshop (MLMI), MICCAI 2014, Boston, USA.

Maraci MA, Napolitano R, Papageorgiou A, Noble JA. Searching for Structures of Interest in an Ultrasound Video Sequence with an Application for detection of Breech. Proceedings of the 24th World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2014, 14-17 September 2014. (Clinical Abstract)



(a) *Conventional standardised scan* (b) *Predefined free-hand scan*

Figure 4.1: Obstetric scanning protocols. (a) *Conventional standardised ultrasound scanning protocol where a sonographer scans over multiple paths to locate the best visual representation of the structure of interest for further analysis.* (b) *User follows a simple predefined free-hand scanning protocol, where automated analysis is carried out on this data to find the structure of interest.*

4.1 Introduction

Encouraging results from observational studies have demonstrated that trained and standardised healthcare workers in developing countries can perform as well as qualified sonographers in terms of measurements reproducibility [175]. Similarly, we reason that novice sonographers can also be trained to perform *specific* and *targeted* obstetric ultrasound diagnostic screening tasks through simple predefined scanning protocols that are focused on specific conditions. This is of particular relevance to limited-resource settings, where pregnant women may not be able to attend specialist centres for obstetric care on regular basis, during the course of their pregnancy.

In this chapter, we focus on automatic detection of the fetal presentation as the clinical application of interest. In order to fulfil this, a simple free-hand clinical ultrasound scanning protocol is defined¹ instead of the conventional obstetric scans² that are currently used in practice. The two types of scans are schematically illustrated in Figure 4.1. The predefined free-hand scans produce a 2D ultrasound video sequence,

¹We refer to this type of protocol as *predefined free-hand*.

²In the clinical setting this is referred to as a *standardised scan*.

where sub-sequences of interest within the original videos (the sequences containing a fetal skull in this case) are identified through their motion patterns and video dynamics. Moreover, although the proposed clinical application in this chapter is a special case of general ultrasound use, we believe that many clinical tasks that make use of ultrasound, of which fetal biometry analysis is just one, could be performed using simplified scanning protocols that can be readily learnt by novice sonographers.

It is worth noting that simplified ultrasound scans inherently contain limited amounts of information and may not be applied to all types of conditions. Moreover, it is important to bear in mind that for any given task the necessary structures may be missed out from a simplified scan, thus the operator would have to repeat the scan multiple times. Analysis of data of this kind would require taking such limitations into account.

In order to analyse the evolution of dynamic scenes, Linear Dynamical Systems (LDSs) provide a powerful and general framework for modelling temporal evolution [54, 55, 64, 178, 208]. With such models, the task can be broken down into three main parts. Firstly, extracting features that represent the dynamic context of a sequence of frames. These are assumed to be the output of an LDS. Secondly, setting up a recognition pipeline to identify and learn the LDS parameters for the corresponding video sequences. Lastly, defining a metric between the system parameters to allow an effective comparison between the models for classification. These steps are covered in the following sections of this chapter. In what follows, the clinical application and the image dataset are explained before discussing the method and results.

4.2 Clinical Task: Breech Presentation

As discussed earlier in Chapter 1, *fetal presentation* refers to the position of the fetus relative to the maternal pelvis, in obstetrics. A fetus is said to be *breech* if its head is farthest from the maternal pelvis, as illustrated in Figure 4.2. Knowing the

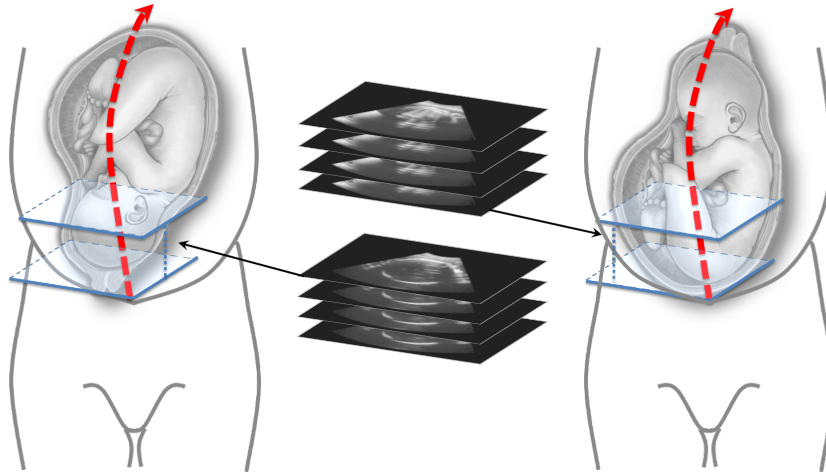


Figure 4.2: *Fetus in cephalic position (left) and breech position (right). The red dotted arrow indicates the ultrasound sweep path. In a video obtained from the described predefined free-hand protocol, identifying the fetal skull in the first part of the video is an indication of the cephalic presentation.*

fetal presentation prior to delivery can be extremely important for effective management and planning of labour. Previous studies have shown vaginal birth of a fetus in a breech position is associated with an increased risk of adverse neonatal outcomes including morbidity from difficult delivery, low birth weight from preterm delivery, growth restriction and even death [119]. Therefore, detection of the fetal head and its relationship with the uterine major axis can be of great importance and is indeed essential to reducing the likelihood of such outcomes. Identifying the relative position of the fetal skull in an ultrasound video that has been acquired following a predefined free-hand scan, such as Figure 4.1b, could be sufficient to identify whether a fetus is in a breech position or not, during the third trimester of pregnancy. In this chapter, we propose and describe a method to identify ultrasound image sub-sequences that correspond to a fetal skull, based on their motion patterns as the basis of a clinical viable protocol for determining fetal presentation.

4.3 Image Dataset

Clinical fetal ultrasound video sweeps were acquired specifically for this study. The ultrasound data sweeps are a stack of 2D frames in time, which constitute to a $(2D+t)$ dataset. Data acquisition was carried out by four qualified sonographers who were trained for 10 minutes and asked to follow a simple standardised predefined-scanning protocol as illustrated in Figure 4.1b. It is important to note eventhough the data acquisition is carreid out by expert sonographers, the scanning protocol is extremely simple and novice sonographers are expected to be able to perform in a similar manner. The participants were healthy volunteers at 28^+ weeks of gestation, that participated in the INTERGROWTH-21st project [157,210]. The clinical application discussed in this chapter requires an in-depth explanation of the current practices and their limitation, which are covered in this section.

4.3.1 Standard Plane in Ultrasound Obstetrics

The current standard practice of fetal ultrasound data analysis is based on acquisition of key anatomical structures, satisfying a specific set of criteria. These criteria require the 2D ultrasound images to be acquired in a particular ultrasound plane, subject to the presence of key landmarks, and are referred to as the *standard plane* [179,180]. As an example, Table 4.1 describes the Salomon criteria [180] for the fetal skull, fetal abdomen and fetal femur. This can be used to calculate a simple score to quantify how well a 2D ultrasound image complies with the *standard plane* criteria. However it is worth noting that given the single predefined free-hand acquisition protocol and the relatively small size of the fetal heart structures, the probability of capturing key structures in their standard plane can be *low*. Therefore other suitable metrics may need to be defined for characterising acceptable forms of anatomical structures of interest.

Table 4.1: *Scoring system for fetal ultrasound images as introduced by Salomon et al. [180]. For each of the three fetal structure, presence of each measurement is associated with a positive score.*

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

4.3.2 New Criteria for Characterising Anatomical Objects

Given that video data are not necessarily acquired by experienced sonographers, the visual appearance of anatomical structures may vary remarkably in different scans, due to the inherent flexibility of the predefined free-hand scanning protocol. The position and angle in which the ultrasound transducer is held directly affects the visual appearance of anatomical structures. This is illustrated in Figure 4.3, where holding the transducer at various angles can effect the appearance of fetal brain structures. Moreover, Figure 4.4 shows the potential ambiguity in a typical predefined free-hand sweep. As can be seen, the inter and intra-variation in appearance of the fetal skull, abdomen and heart can be significantly high.

Therefore, it becomes particularly important to define a set of standardised metrics in order to be able to formally characterise anatomical structures of interest. Although similar metrics already exist for various anatomical structures, they are mainly defined

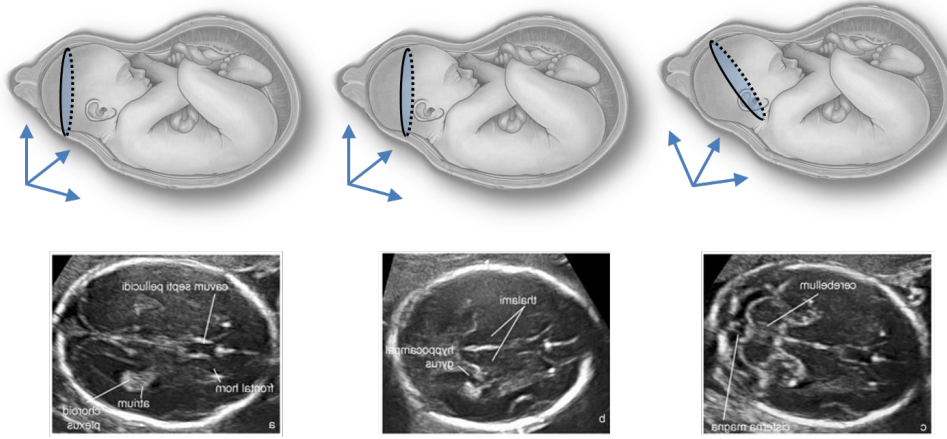


Figure 4.3: *Ultrasound probe's position and angle.* The appearance of anatomical structures depends on the position and angle of the ultrasound probe. Here three different planes of the fetal skull are illustrated, where different structures are visible on each plane. Image adapted from [1].

for structures that have been captured in their standard plane and not for structures that are out of their standard plane.

This facilitates reproducibility of data by different sonographers and also provides the means to quantitatively evaluate the learning algorithm's performance. Considering the clinical application of interest, it is of great importance that the fetal skull and abdomen are captured as part of the acquired ultrasound video in order to detect the fetal presentation. Capturing these anatomical structures in their standardised planes is however not required, and is indeed highly unlikely. Therefore, in collaboration with experienced sonographers, new criteria have been defined for characterising and indeed recognising the fetal skull and abdomen in a predefined free-hand ultrasound sweep. The criteria are schematically illustrated in Figures 4.5 and 4.6. Given that the structures of interest can now be formally characterised, a method to automatically identify video sequences that contain these structures is introduced in the next section.

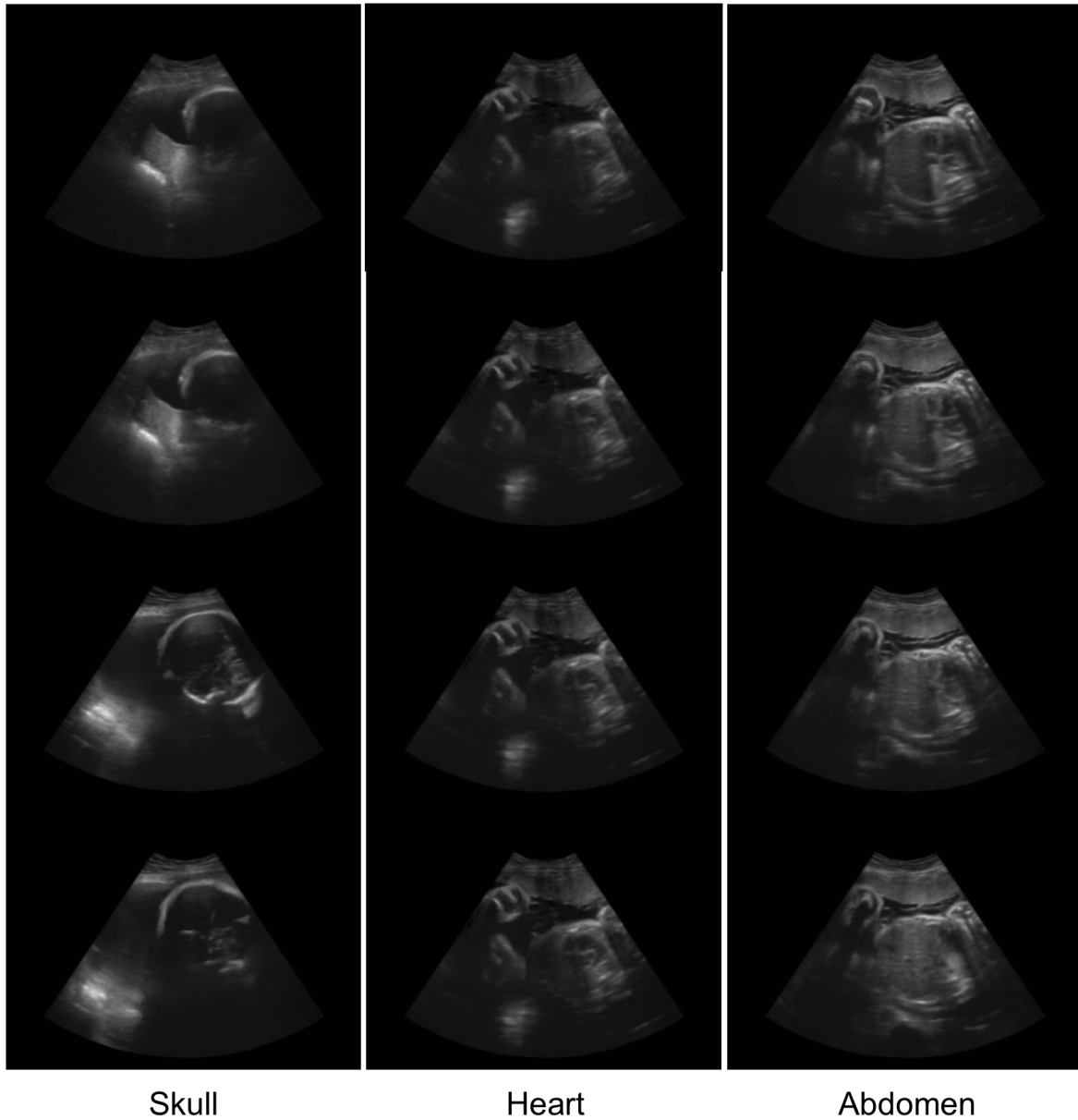


Figure 4.4: *Example frames extracted from a predefined-freehand sweep. Nine borderline frames from a sweep are shown here. From top to bottom the 93rd, 100th, 149th and 156th frames are on the border of being labelled a fetal skull. Similarly, frames 215, 216, 217 and 218 and frames 241, 246, 248 and 252 are on the borderline of being labelled a fetal heart and abdomen receptively.*

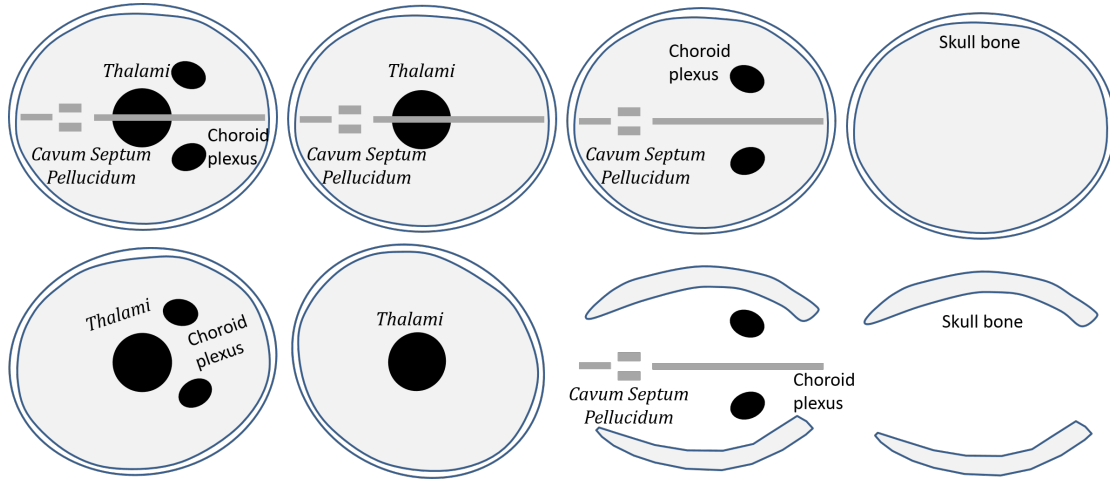


Figure 4.5: *Accommodating criteria for fetal skull.* The only necessary criterion is presence of a semi-elliptical shape with or without the other standard key landmarks. The criteria are set loosely and independent of the presence of key anatomical structures.



Figure 4.6: *Accommodating criteria for fetal abdomen.* A semi-elliptical shape is required with at least one of the key landmarks. The criteria are set very loosely and are not dependent on the presence of key anatomical structures.

4.4 The Dynamical System

Modelling motion in sequences with stochastic motion fields can be challenging. In the biological visual perception, such motion patterns are generally seen as *holistic*

phenomena. For example, it can be much simpler to visually identify a fetal heartbeat from a consecutive set of frames - as a running video - rather than on disjoint 2D images. Figure 4.7 illustrates a set of frames that contain a fetal heart. As can be seen, it is challenging to identify the fetal heart when the frames are looked at individually. In our dataset, given that the standardised freehand scanning protocol is followed during data acquisition, the probe motion is assumed to be a constant factor throughout the dataset

Motion in a video sequence can be modelled using dynamical systems such as the dynamic texture (DT) model introduced by Doretto et al. [54]. An advantage of using this family of methods is that dynamics of the entire sequence are taken into account. This is achieved by considering the motion patterns as samples from a stochastic process, defined over some space and time [54, 55, 64, 178, 208]. The DT model is a linear dynamical system that consists of an observed variable $\mathbf{y}_t$ that determines the appearance of each video frame, and a hidden variable $\mathbf{x}_t$ determining the dynamics in the video over time. In such a framework, both the observed and hidden variables are representative of the entire image, thus allowing characterisation of motion for the entire sequence.

The problem of modelling dynamic textures to capture the temporal behaviour of moving scenes has been successfully addressed within the computer vision community [54, 117, 208]. In addition, various extensions have been suggested in the literature to capture complex non-linear or sudden motion patterns in a video including switching linear dynamical system models [154], codebook elements in a bag of dynamical systems [40, 171], and kernel functions and mixture models [32, 33]. An advantage of utilising kernel functions such as the kernel dynamic texture (KDT) approach, is that as long as a suitable kernel is defined, features other than pixel values can be used. The work of Chaudhry et al. [34] is an example of this, where the dynamics of histogram features are modelled in a video. Recently, Kwitt et al. [107] adopted this

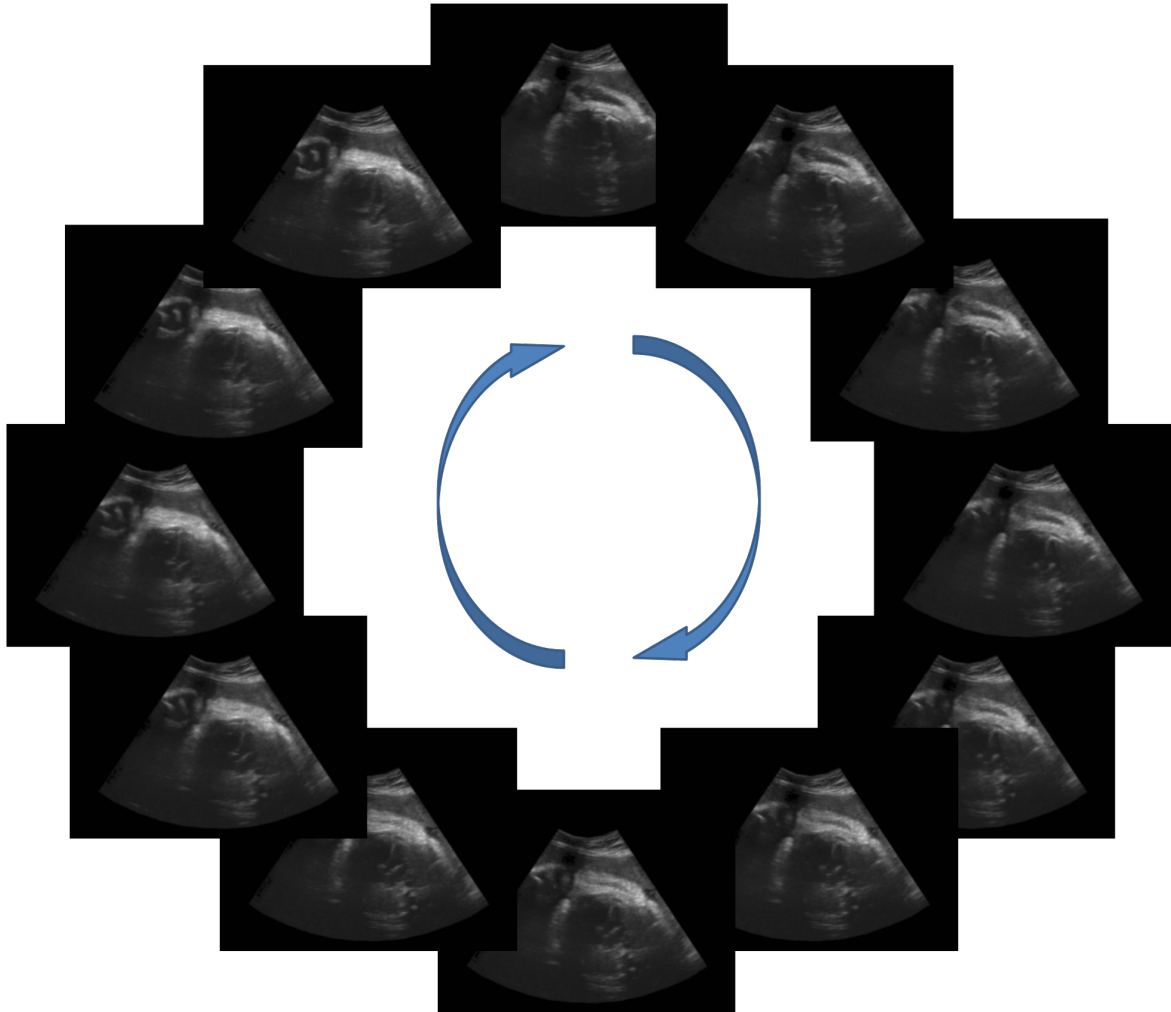


Figure 4.7: *Fetal heartbeat as a continuous sequence.* The motion patterns can play an important role in detection of certain sequences. For example in this sequence it is challenging to find and located the fetal heart on any single frame. However, it can be seen easily through a continuous sequence, through its distinct motion patterns.

approach for object localisation from phantom ultrasound video sequences. In their study, a number of *key* sequences that contain the structures of interest are acquired and their model parameters are computed and stored as reference points. During test time, model parameters are then calculated for the unseen test sequences and compared to those of the key sequences using the Martin distance [46, 133] metric, which is calculated in terms of the parameters of the state-space models and is explained later in this chapter. In the next section, we briefly explain the DT model, which is fundamental in understanding the KDT model used in our work.

4.4.1 Detecting Sections of Interest

Given an ultrasound image sequence of T video frames, $\{\mathbf{y}_t\}_{t=1}^T$, the feature trajectories of $\{\mathbf{y}_t\}_{t=1}^T$ can be modelled as the output of an LDS. We follow Doretto et al. [54] for computing the system parameters, which models pixel intensities in each frame as the output of an LDS. However, instead of using raw pixel intensities as in [54], our framework consists of a pre-processing stage to produce a contrast-invariant representation of anatomical structures by constructing a feature symmetry map for each frame.

Recall from Chapter 3 that the feature symmetry (FS) and signed feature symmetry (SFS) maps of a 2D image can be constructed using the monogenic signal [62], which is composed of an even-symmetric bandpass image filter (log-Gabor filter in this case) denoted by $even(x, y)$, and two odd-symmetric filters formed from its Riesz transform denoted by $odd(x, y)$. Formally FS is defined as:

$$FS(x, y) = \sum_s \frac{[|even_s(x, y)| - |odd_s(x, y)| - T_s]}{\sqrt{even_s(x, y)^2 + odd_s(x, y)^2 + \epsilon}}, \quad (4.1)$$

and SFS is defined as:

$$SFS(x, y) = \sum_s \frac{[|even_s(x, y)| - |odd_s(x, y)| - T_s]}{\sqrt{even_s(x, y)^2 + odd_s(x, y)^2 + \epsilon}} \cdot \text{sgn}(even_s(x, y)), \quad (4.2)$$

where s is the scale of the band-pass filter, ϵ is a small positive constant to avoid division by zero, $[\cdot]$ denotes replacing negative values with zeros, T_s is noise compensation threshold, and $\text{sgn}(x)$ is the *signum* function.

The Dynamical Model

In particular given an LDS model, the appearance of each video frame is determined through the observed variable $\mathbf{y}$, and the motion and dynamics in the video over a

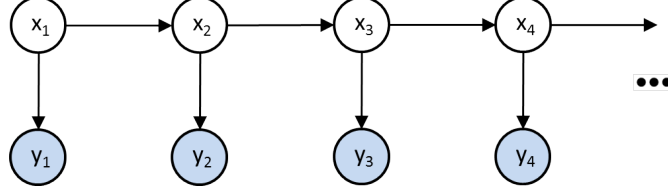


Figure 4.8: *The dynamic texture model graphical representation. In this model $\mathbf{x}_t$ and $\mathbf{y}_t$ are the hidden state and observed video frame at time t respectively.*

given time is determined through the hidden-state variables $\mathbf{x}$, which are sampled from a Gauss-Markov process. In such a model it is assumed that the observed frame at any given time can be constructed from a linear combination of the hidden state variables.

More formally, given an ultrasound sequence $\mathbf{S}$ of T video frames, let $\mathbf{S} = [\mathbf{y}_1, \dots, \mathbf{y}_T]$, where $\mathbf{y}_t \in \mathbb{R}^d$ is the frame observed at time t . It is assumed that at each time instant t , a noisy version of the image can be measured such that $\mathbf{y}(t) = \mathbf{S}(t) + \mathbf{w}(t)$, where $\mathbf{w}(t) \in \mathbb{R}^d$ is an independent and identically distributed (i.i.d.) sequence drawn from a known distribution, resulting in a sequence $\mathbf{y}(t) \in \mathbb{R}^d$ for $t = 1, \dots, T$. Therefore the state and observed variables are related through the LDS that is defined by:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t, \\ \mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{w}_t + \bar{\mathbf{y}}. \end{cases} \quad (4.3)$$

Here $\mathbf{x}_t \in \mathbb{R}^n$ is the state of the LDS and $\mathbf{y}_t \in \mathbb{R}^d$ is the observed pixel intensities at time t . Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the state transition matrix that describes the dynamics of the state evolution and $\mathbf{C} \in \mathbb{R}^{d \times n}$ is the output matrix (contains the principal components learned with [54]), and $\bar{\mathbf{y}}$ is the mean of the dynamical texture sequence. The state noise process, $\mathbf{v}_t$ has zero mean and has covariance $\mathbf{Q}$, $\mathbf{v}_t \sim N(\mathbf{0}, \mathbf{Q})$ where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is a positive definite matrix. Similarly the observation noise $\mathbf{w}_t$ is derived from a zero-mean Gaussian distribution with covariance $\mathbf{R}$, $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{R})$ for a positive definite matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$. The dynamic texture model can be represented by the graphical model in Figure 4.8.

In an LDS model such as Equation 4.3, it can be seen that the choice of matrices

$\mathbf{A}$, $\mathbf{C}$, and $\mathbf{Q}$ are not unique. Consequently, many such matrices could result in the same output $\mathbf{y}_t$, given an appropriate initial condition. As shown in [54], in order to identify a unique model from a sequence $\mathbf{y}_t$, a canonical model realisation needs to be selected to be able to uniquely represent each model. Therefore it is assumed that $d \gg n$ and $\text{rank}(\mathbf{C}) = n$ in Equation 4.3, thus a canonical model that makes the columns of $\mathbf{C}$ orthonormal is chosen such that: $\mathbf{C}^\top \mathbf{C} = \mathbf{I}_n$, where $\mathbf{I}$ is the identity matrix with dimensions $n \times n$.

The problem can now be formulated such that given a sequence of measurements, $\mathbf{S} = [\mathbf{y}_1, \dots, \mathbf{y}_T]$, we have to estimate $\hat{\mathbf{A}}, \hat{\mathbf{C}}, \hat{\mathbf{Q}}, \hat{\mathbf{R}}$, a canonical model of the process $\mathbf{S}$. We follow the least-squares algorithm proposed in [54] for estimating the system parameters. In this framework given the observed video sequence $\mathbf{S} = [\mathbf{y}_1, \dots, \mathbf{y}_T]$, the sample mean is estimated by:

$$\bar{\mathbf{y}} = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t,$$

and subtracted from the sequence such that $\tilde{\mathbf{S}} = [\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_T]$, where $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \bar{\mathbf{y}}, \quad \forall t$. From here on $\tilde{\mathbf{S}}$ will be denoted as $\mathbf{S}$ unless otherwise stated.

Model parameters are then estimated using singular value decomposition (SVD) such that, $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where the output matrix $\mathbf{C}$ is estimated from the c principal components of largest eigenvalue. Therefore $\mathbf{C} = [\mathbf{u}_1, \dots, \mathbf{u}_c]$, where $\mathbf{u}_i$ is the i -th column of $\mathbf{U}$, assuming the diagonal entries of $\mathbf{\Sigma}$ are in descending order. The state-space parameters are estimated such that:

$$\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_T] = \mathbf{C}^\top \mathbf{S},$$

which leads to the least-square estimate of the state-transition matrix

$$\mathbf{A} = \hat{\mathbf{X}}_{2:T}(\hat{\mathbf{X}}_{1:T-1})^\dagger.$$

Here $\mathbf{X}^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$, which represents the Moor-Penrose pseudo-inverse of $\mathbf{X}$.

Kernel Dynamic Texture Model

In this thesis, we use a kernel dynamic texture (KDT) model as initially suggested in [32, 107]. In this model, the evolution of the hidden states are kept linear but in order to capture the dynamics of the video the output matrix $\mathbf{C}$ is replaced by a non-linear observation function $C : \mathbb{R}^n \rightarrow \mathbb{R}^d$, utilising the kernel principal component analysis (KPCA). With PCA [57], data is projected to the linear subspace that best captures the variability of the data. However KPCA is able to project the data onto non-linear functions in the input space, thus transforming the data into some non-linear feature space and then applying standard PCA in feature space.

Given the same ordered ultrasound sequence $\mathbf{S}$ and a kernel function $k(\mathbf{y}_1, \mathbf{y}_2)$ with associated feature transformation $\langle \phi(\mathbf{y}_1), \phi(\mathbf{y}_2) \rangle$, the kernel principal components are the eigenvectors of the covariance matrix of the transformed data $\phi(\mathbf{y}_i)$. Assuming the data in the feature space is centred (has zero mean in the feature space), the c -th kernel principal component in the feature space can be represented as [182],

$$\mathbf{kp}_c = \sum_{i=1}^T \alpha_{i,c} \phi(\mathbf{y}_i), \quad (4.4)$$

where $\alpha_{i,c}$ represents the i -th component of the c -th KPCA weight vector and $\boldsymbol{\alpha}_c = \frac{1}{\sqrt{\lambda_c}} \mathbf{kp}_c$, assuming the eigenvectors are sorted in descending order of the eigenvalues $\{\lambda_c\}_{c=1}^T$. Here λ_c and $\mathbf{kp}_c$ are the c -th largest eigenvalue and eigenvector of the kernel matrix $\mathbf{K}$ with entries $[\mathbf{K}]_{i,j} = k(\mathbf{y}_i, \mathbf{y}_j)$. Moreover, the scaling of the $\mathbf{kp}_c$ ensures that the principal component in the feature-space has unit length. Finally, given a training data sample $\mathbf{y}_j$, the c -th KPCA coefficients $x_{c,j}$ can be computed as the projection of $\mathbf{y}_j$ in the feature space $\phi(\mathbf{y}_j)$ onto the principal component $\mathbf{kp}_c$ such that,

$$x_{c,j} = \langle \phi(\mathbf{y}_j), \mathbf{kp}_c \rangle = \sum_{i=1}^T \alpha_{i,c} k(\mathbf{y}_i, \mathbf{y}_j). \quad (4.5)$$

The KPCA coefficients $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$ and the state transition matrix $\mathbf{A}$ for the training set $\mathbf{Y}$ can then be estimated as,

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\alpha}^\top \mathbf{K}, \\ \mathbf{A} &= [\mathbf{x}_1, \dots, \mathbf{x}_{T-1}] [\mathbf{x}_0, \dots, \mathbf{x}_{T-2}]^\dagger, \end{aligned} \quad (4.6)$$

where $\alpha = [\alpha_1, \dots, \alpha_c]$ is the KPCA weight matrix of the first c -th principal components and $\mathbf{X}^\dagger$ is the pseudo-inverse of $\mathbf{X}$.

4.4.2 Distance Metrics for Comparing LDSs

Given a KDT model estimate for each sub-sequence, a suitable metric needs to be defined to assess similarity between any two models. Prior work on comparison metrics of LDSs vary from metrics based on subspace angles between the observability subspaces of the systems [46] to metrics based on the Binet-Cauchy kernels [14, 214] and finally metrics based on the KL-divergence of the probability distributions of the stochastic processes [31]. A full comparison of these classes of metrics is outside the scope of this thesis. However, Chaudhry et al. [35] showed that the metrics based on the Martin Distance and Binet-Cauchy maximum singular values kernel performed best on a number of different applications. In this framework during the training phase, a kernel function is defined based on a distance metric, between all the videos in the training set. The kernel function indicates the similarity between any two sequences and is used to train a classifier. During testing, the same kernel function is used to create a kernel distance matrix between the test videos and those in the training set. The trained classifier is then used to classify the sequences in the test dataset, based on the values of the kernel distance matrix. In what follows, the distance metrics used in this chapter are discussed in more detail.

Distance Metrics Based on Subspace Angles

Martin Distance

The Martin distance was initially defined by Martin [133] and later introduced by De Cock and De Moor [47] as a metric for measuring similarity between models of auto-regressive moving average (ARMA) processes. This metric was first used by Saisan [178] as a similarity measure between two dynamic texture models, where the metric is based on the principal angles between the observability subspaces of models.

Formally given $M_i = (\mathbf{A}_i, \mathbf{C}_i)$ for two dynamic texture models $i = 1$ and $i = 2$, defined by their parameters $\mathbf{A}_i$ and $\mathbf{C}_i$ with model order n , the Martin distance is defined as;

$$d_m(M_1, M_2)^2 = -\log \prod_{i=1}^n \cos^2(\theta_i) \quad (4.7)$$

where θ_i is the i th principal angle between the extended observability matrices $\mathbf{O}_1$ and $\mathbf{O}_2$ which are defined as;

$$\mathbf{O}_i = [\mathbf{C}_i^\top, (\mathbf{C}_i \mathbf{A}_i^\top), (\mathbf{C}_i \mathbf{A}_i^2)^\top, \dots] \quad (4.8)$$

for $i = 1, 2$. As shown in [47], the principal angles can be computed as the solution to the following generalised eigenvalue problem:

$$\begin{bmatrix} 0 & \mathbf{O}_{12} \\ (\mathbf{O}_{12})^\top & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{O}_{11} & 0 \\ 0 & \mathbf{O}_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

subject to $x^\top \mathbf{O}_{11} x = 1$ and $y^\top \mathbf{O}_{22} y = 1$, where

$$\mathbf{O}_{ij} = (\mathbf{O}_i)^\top \mathbf{O}_j = \sum_{t=0}^{\inf} (\mathbf{A}_i^t)^\top \mathbf{C}_i^\top \mathbf{C}_j \mathbf{A}_j^t \quad (4.9)$$

for $i = 1, 2$ and $j = 2, 1$. It was further shown that the first n largest eigenvalues are

the cosines of the principal angles, thus $\lambda_i = \cos\theta_i$ for $i = 1, 2, \dots, n$, therefore

$$d_m(M_1, M_2)^2 = -2 \sum_{i=1}^n \log(\lambda_i) \quad (4.10)$$

Furthermore the infinite sum in equation 4.9 can be obtained by solving the corresponding Sylvester equation, which may be solved using standard numerical packages:

$$\mathbf{O}_{ij} = \mathbf{A}_i^\top \mathbf{O}_{ij} \mathbf{A}_j + \mathbf{C}_i^\top \mathbf{C}_j \quad (4.11)$$

Extending the above definition of the Martin distance to a kernel dynamic texture can be intuitive by considering that the kernel dynamic texture is just the same as a dynamic texture model, with the difference that it has been learned in the feature space of a kernel. Therefore the inner product between the principal components in the linear model, $\mathbf{C}_1^\top \mathbf{C}_2$, can be replaced with a kernel principal components in the feature space. As show in [32] both kernels must however induce the same inner product in the same feature space.

To illustrate with an example, given two data sequences $\{\mathbf{y}_i^{(a)}\}_{i=1}^{N_a}$ and $\{\mathbf{y}_i^{(b)}\}_{i=1}^{N_b}$ with two corresponding kernel functions k_a and k_b with feature transformation functions $\phi(\mathbf{y})$ and $\psi(\mathbf{y})$ respectively, such that:

$$k_a(\mathbf{y}_1, \mathbf{y}_2) = \langle \phi(\mathbf{y}_1), \phi(\mathbf{y}_2) \rangle, \quad k_b(\mathbf{y}_1, \mathbf{y}_2) = \langle \psi(\mathbf{y}_1), \psi(\mathbf{y}_2) \rangle. \quad (4.12)$$

Applying Kernel PCA to each sequence with their corresponding kernels yields KPCA weight matrices α and β receptively, thus the c th and d th KPCA components in each feature space is given by:

$$\mathbf{u}_c = \sum_{i=1}^{N_a} \alpha_{i,c} \phi(\mathbf{y}_i^{(a)}), \quad \mathbf{u}_d = \sum_{i=1}^{N_b} \beta_{i,d} \psi(\mathbf{y}_i^{(a)}). \quad (4.13)$$

We can now define the inner-product between the KPCA components as follows:

$$\begin{aligned}
\langle \mathbf{u}_c, \mathbf{u}_d \rangle &= \left\langle \sum_{i=1}^{N_a} \alpha_{i,c} \phi(\mathbf{y}_i^{(a)}), \sum_{j=1}^{N_b} \beta_{j,d} \psi(\mathbf{y}_j^{(b)}) \right\rangle \\
&= \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \alpha_{i,c} \beta_{j,d} g(\mathbf{y}_i^{(a)}, \mathbf{y}_j^{(b)}) \\
&= \alpha_c^\top \mathbf{G} \beta_d.
\end{aligned}$$

where $g(\mathbf{y}_1, \mathbf{y}_2) = \langle \phi(\mathbf{y}_1), \psi(\mathbf{y}_2) \rangle$ and $\mathbf{G}$ is the corresponding matrix with entries $\mathbf{G}_{i,j} = g(\mathbf{y}_i^{(a)}, \mathbf{y}_j^{(b)})$. As show in [32], the function g can be computed if the two kernels have compatible feature-spaces, i.e. the two kernels induce the same inner-product in the feature-space. Therefore the inner product matrix between all KPCA components is:

$$\mathbf{F} = \alpha^\top \mathbf{G} \beta, \quad (4.14)$$

which takes the place of $\mathbf{C}_i^\top \mathbf{C}_j$ in Equations 4.9 and 4.11.

Distance Metrics Based on Trace Kernels

BC-Max Singular Value Kernel

As previously mentioned, the SVM classification algorithm can be used on distances as well as kernels. To capitalise on this property, Vishwanathan et al. introduced a family of kernels for LDSs, called the Binet-Cauchy (BC) kernels [214]. From this family of kernels, the trace kernel K_{LDS}^T for comparing two infinitely long centred Euclidean sequences represented by LDS models $M = (x_0, \mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R})$ and $M' = (x'_0, \mathbf{A}', \mathbf{C}', \mathbf{Q}', \mathbf{R}')$ is defined as:

$$\begin{aligned}
K_{LDS}^T(\{\mathbf{y}_{t=0}^\infty\}, \{\mathbf{y}'_{t=0}^\infty\}) &= K_{LDS}^T(\mathbf{M}, \mathbf{M}') \\
&:= \mathbb{E}_{v,w} \left[\sum_{t=0}^{\infty} \lambda^t \mathbf{y}_t^\top \mathbf{y}'_t \right].
\end{aligned}$$

where λ is a weight factor between 0 and 1 and $\mathbb{E}$ is the expected value of the infinite sum of inner products with respect to the joint probability distribution of $\mathbf{v}_t$ and $\mathbf{w}_t$. Furthermore it was shown in [214] that if the two systems have the same underlying independent noise process with covariance $\mathbf{Q}$ and $\mathbf{R}$ for the state and output process respectively, the BC trace kernel can be computed in closed form as:

$$K_{LDS}^T(\mathbf{M}, \mathbf{M}') = \mathbf{x}_0^\top \mathbf{P} \mathbf{x}'_0 + \frac{\lambda}{1-\lambda} \text{trace}(\mathbf{Q} \mathbf{P} + \mathbf{R}) \quad (4.15)$$

where

$$\mathbf{P} = \sum_{t=0}^{\infty} \lambda^t (\mathbf{A}^t)^\top \mathbf{C}^\top \mathbf{C}' \mathbf{A}^t \quad (4.16)$$

If $\lambda \|\mathbf{A}\| \|\mathbf{A}'\| < 1$, $\mathbf{P}$ can be computed using the Sylvester equation [214] as:

$$\mathbf{P} = \lambda \mathbf{A}' \mathbf{P} \mathbf{A}' + \mathbf{C}^\top \mathbf{C}' \quad (4.17)$$

As a result of learning the system parameters following [54], the solution involves computing the matrix of pairwise inner products of the principal components of the two data matrices. This was extended to non-linear dynamical systems (NLDS) by Chaudhry et al. [34], where the trace kernel is now calculated as the expected value of an infinite series of weighted inner products between the outputs in an embedded high-dimensional space using the map Ψ as:

$$\begin{aligned}
K_{NLDS}(\mathbf{M}, \mathbf{M}') &:= \mathbb{E}_{\mathbf{v}, \mathbf{w}} \left[\sum_{t=0}^{\infty} \lambda_t \Psi(\mathbf{y}_t)^\top \Psi(\mathbf{y}'_t) \right] \\
&= \mathbb{E}_{\mathbf{v}, \mathbf{w}} \left[\sum_{t=0}^{\infty} \lambda_t k(\mathbf{y}_t, \mathbf{y}'_t) \right]
\end{aligned}$$

Furthermore similar to the case of the Martin distance, $\mathbf{C}^\top \mathbf{C}'$ is now the product of matrices from kernel principal components rather than the principal components in the linear LDS. A BC-trace kernel for NLDSs is then suggested in [34] as:

$$K_{NLDS}(\mathbf{M}_1, \mathbf{M}_2) = \mathbf{x}_0^\top \bar{\mathbf{P}} \mathbf{x}_0 + \frac{\lambda}{1 - \lambda} \text{trace}(\mathbf{Q} \bar{\mathbf{P}} + \mathbf{R}), \quad (4.18)$$

where $\mathbf{x}_0$ is the initial state of the system, $\bar{\mathbf{P}} = \sum_{t=0}^{\infty} \lambda_t (\mathbf{A}_t^1)^\top \mathbf{F} \mathbf{A}_t^2$, $\mathbf{F}$ is the inner product matrix between all the KPCA components and $\mathbf{Q}$ and $\mathbf{R}$ are the state and output covariance matrices. To remove the dependency on the initial state and the noise process, Chaudhry et al. [34] proposed the BC maximum singular value kernel for NLDSs as $K_{NLDS}^\sigma = \max \sigma(\bar{\mathbf{P}})$, where σ represents the singular values kernel, to take into account only the dynamics of the NDLS. Thus a normalized kernel of the similarity values can be constructed such that $K(\mathbf{M}_1, \mathbf{M}_2) = 1$ if $\mathbf{M}_1 = \mathbf{M}_2$ as

$$K(\mathbf{M}_1, \mathbf{M}_2) = \frac{K(\mathbf{M}_1, \mathbf{M}_2)}{\sqrt{K(\mathbf{M}_1, \mathbf{M}_1) K(\mathbf{M}_2, \mathbf{M}_2)}} \quad (4.19)$$

A distance can now be computed as $d(\mathbf{M}_1, \mathbf{M}_2) = 2(1 - K(\mathbf{M}_1, \mathbf{M}_2))$ and classification is then carried out using these values.

Contact the author for images

Figure 4.9: *Snapshots of three key structures at two time points on the noodle phantom. Key structures are short sequences from arbitrary locations on the noodle phantom. Image adapted from [108].*

4.5 Experiments and Results

Experiment One

This experiment was designed to illustrate the capabilities of the proposed method on phantom and clinical data. The phantom data was provided by the authors of [107] and enables direct comparison with that work. The clinical data was generated specifically for this experiment and consisted of two 2D fetal ultrasound sweeps each acquired from 35 subjects. Clinical data acquisition was carried out as previously outlined in Section 4.2.

Phantom Study Comparison This experiment allowed a systematic comparison between Martin distance (MD) as utilised in [107] and the BC maximum singular values kernel distance proposed in our work, on data from a home-made *noodle phantom* made of gelatine with embedded Soba noodles. Figure 4.9 shows snapshots of the phantom. The task was to detect key structures in the noodle phantom. For this purpose, two different datasets were constructed. The first dataset is composed of short (40 frames) ultrasound key sequences captured by moving the transducer over three different *key structures* that are going to be detected. The key structures are set arbitrary from random parts of the noodle phantom and do not signify any particular structures. The second set contained longer sequences of ultrasound data acquired along multiple paths at arbitrary directions, which simulated search paths for any of the key sequences. Moreover the method described in [107] was re-implemented - without applying our feature symmetry data enhancement step - to estimate the MD model parameters and

to carry out sequence classification. As in [107], Average Precision (AP), mean AP and mini/max AP are reported, and the experiment was repeated for an LDS model with 2, 5, and 8 states.

Results from the phantom study experiment are summarised in Table 4.2a. Consistent with the findings of [107] for the Martin distance, a 5-state KDT model yielded the best performance. The highest mAP achieved on the phantom dataset using the Martin Distance was 51.14% (5-state KDT model). In general, results obtained using the BC similarity metric were much higher compared to those of the Martin distance. The highest mAP achieved using the BC metric was 70.14% (8-state KDT model).

Clinical Data Comparison A similar experimental protocol to the phantom study comparison was followed for the clinical data evaluation with the following changes. The goal here was to detect the sequences that contain the fetal head. The acquisition protocol was as described in Section 4.3. The training and testing sequences were selected to be 20 frames long, and were labelled positive if they contained the fetal skull and negative otherwise by the author and confirmed by an expert obstetrician. The data was constructed from a pool of 70 videos, obtained from 35 participants (two sweeps from each participant). Furthermore, an SVM classifier was trained to identify sequences of frames that contained the fetal skull. For the purposes of measuring the performance of the classifier, the average precision (AP) value was calculated for each classification outcome as in [108], and to summarise these statistics, the mean average precision (mAP), and the mean classification accuracy (mCA) are calculated over a 10-fold cross-validation.

Tables 4.2b and 4.2c summarise classification results for sub-sequences of 20 frames long. As shown in table 4.2b without the pre-processing step, the best results were achieved with a 5-state KDT model (mAP 90.56%, mCA 85.03%) in comparison to when the frames are pre-processed, where the motion in the video is best described with a 2-state KDT model (mAP 95.85%, mCA 90.21%) using the BC singular value

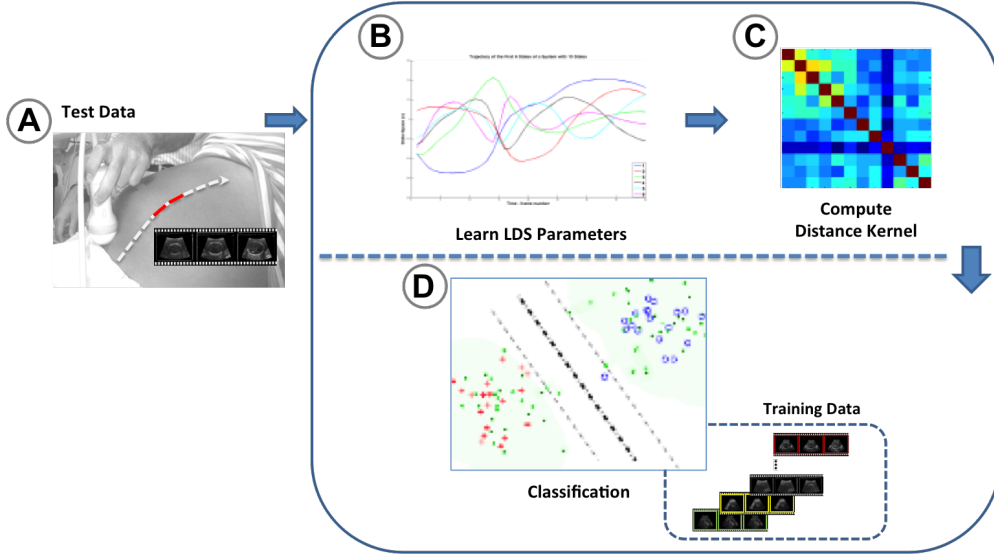


Figure 4.10: Method schematic overview. Given an unseen test video as in step A, it is treated as an LDS and its model parameters are estimated in step B. A distance kernel is computed between the test and training model parameters using an appropriate distance metric in step C. Finally, in step D a pre-trained classifier is used to classify the test video using the distance kernel from step C.

kernel. Generally when filtering is not applied, increasing the number of states from 2 to 5 increases the classification accuracy, however a further increase to 8 states leads to a decrease in performance. This can be explained by the fact that when KPCA is used, the main dynamics of the video are best described using the first 2/5 eigenvalues for the filtered/raw frames respectively, therefore additional eigenvalues capture a very small portion of the variation in the feature space, thus resulting in noisier KDT model parameter estimates.

Experiment Two

The experiments described in this section are designed to show the gain in classification accuracy when feature symmetry and signed feature symmetry maps - as explained in Section 4.4.1 - are utilised on real-world clinical data. The proposed classification framework is schematically illustrated in Figure 4.10. As such, the framework was designed to identify sequences that contain the fetal skull from the rest using an SVM

Table 4.2: *Classification on the phantom dataset(a) and the raw clinical (b) and pre-processed clinical (c) datasets using 10-fold cross validation as previously reported in [129].*

(a) Phantom study comparison				
	State (n)	mAP (%)	min/max AP (%)	P/R, $\alpha=0.05$ (%)
D_m	2	45.2	08.5/72.9	53.4/45.3
	5	51.1	18.2/84.5	45.6/50.4
	8	32.0	01.2/72.4	61.9/34.0
D_b	2	62.2	24.9/80.8	67.4/59.9
	5	68.0	27.7/88.3	69.9/68.3
	8	70.1	26.8/90.4	77.4/51.8

(b) Clinical data (no preprocessing)				
	State (n)	mAP (%)	min/max AP (%)	mCA (%)
D_m	2	23.9	20.3/25.9	65.8
	5	21.5	20.3/22.9	67.5
	8	23.8	22.7/26.3	63.6
D_b	2	87.3	79.6/91.4	82.8
	5	90.6	87.0/93.9	85.0
	8	76.3	57.7/87.9	77.0

(c) Clinical data (preprocessed)					Keys	
	State (n)	mAP (%)	min/max AP (%)	mCA (%)	Precision	TP/(TP+FP)
D_m	2	17.3	14.0/19.6	61.8	Recall	TP/(TP+FN)
	5	29.5	26.9/33.4	44.7	AP	Average Precision
	8	20.7	19.0/22.0	62.7	mAP	Mean Average Precision
D_b	2	95.9	94.1/96.8	90.2	P/R	Precision/Recall
	5	86.6	79.2/95.1	78.6	D_m	Martin Distance Metric
	8	72.8	49.8/94.6	81.2	D_b	Binet-Cauchy
					mCA	Singular Value Metric
						mClassification Accuracy

classifier. To train the system, these sequences have been extracted from long ultrasound sequences acquired from a predefined free-hand ultrasound sweep of a mother's abdomen, as described in Section 4.3.

A dataset of positive and negative examples was constructed for this experiment. From the overall number of ultrasound videos acquired, a total of 162 sub-sequences

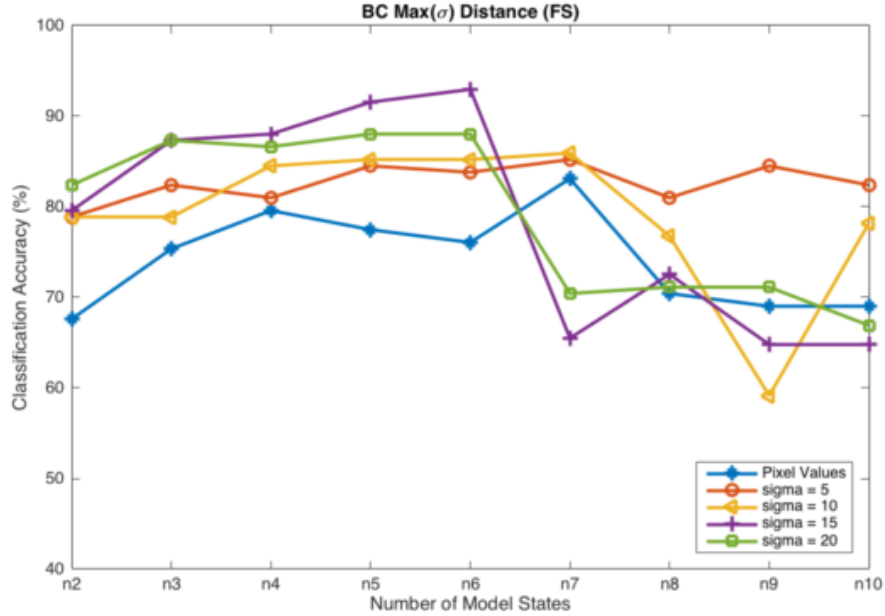
with a duration of 40 frames were selected as the positive samples and 314 sub-sequences of the same length as the negative samples. Here, positive samples are defined as sequences that contain the fetal skull in all the consecutive frames and negative samples are defined as sequences that do not have any frames of the fetal skull. For the purposes of measuring the performance of the classifier, the classification accuracy and the average precision value was calculated for each classification outcome. The experiment was repeated multiple times with the number of model states ranging from 2 to 10 and the centre-wavelength of the log-Gabor bandpass filter used for constructing the feature symmetry was varied from 5 to 20 pixels.

Table 4.3: *Classification results for the clinical dataset using Martin distance and BC-max singular values kernel metrics.*

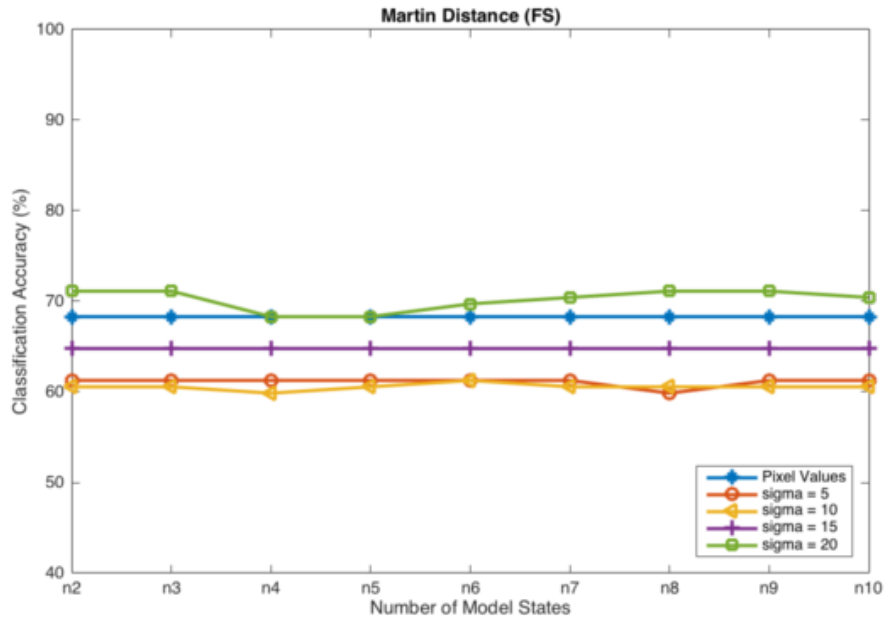
	n2	n3	n4	n5	n6	n7	n8	n9	n10
Martin Distance Metric									
RPI	68.3	68.3	68.3	68.3	68.3	68.3	68.3	68.3	68.3
FS _($\sigma=10$)	60.6	60.6	59.9	60.6	61.3	60.6	60.6	60.6	60.6
FS _($\sigma=15$)	64.8	64.8	64.8	64.8	64.8	64.8	64.8	64.8	64.8
SFS _($\sigma=10$)	61.3	64.8	64.8	61.3	66.2	64.8	61.3	63.4	64.8
SFS _($\sigma=15$)	61.3	71.1	69.0	61.3	71.1	71.1	61.3	71.1	71.1
Binet-Cauchy Metric									
RPI	67.6	75.4	79.6	77.5	76.1	83.1	70.4	69.0	69.0
FS _($\sigma=10$)	78.9	78.9	84.5	85.2	85.2	85.9	76.8	59.2	78.2
FS _($\sigma=15$)	79.6	87.3	88.0	91.5	93.0	65.5	72.5	64.8	64.8
SFS _($\sigma=10$)	85.2	85.9	86.6	85.2	89.4	87.3	86.6	88.0	90.8
SFS _($\sigma=15$)	87.3	88.7	87.3	85.9	90.1	89.4	86.6	90.8	87.3
RPI: Raw Pixel Intensity									
(S)FS: (Signed) Feature Symmetry									

The results illustrate that using the signed feature symmetry instead of the raw pixel intensities increases the classification accuracy of the proposed framework as shown in Figures 4.11 and 4.12. This can be explained by the fact that such preprocessing provides a smooth intensity invariant representation of structures in ultrasound sequences and can reduce the impact of speckle patterns and contrast variations that occur due to shadowing between images.

Similarly the bottom row of Figure 4.12 illustrates the results for using the signed feature symmetry with σ values of 10, 15, 20 and 30 pixels and model orders of $N = 2$ to 10. Although the highest classification accuracy (93%) has been achieved using FS with $\sigma = 15$ and $N=6$, the classification results are consistently higher for signed feature symmetry data as the number of model states increases. Here the highest accuracy for SFS has been achieved for a σ value of 15 and a model with 9 states (90.8%). The settings that have given the best performance are summarised in Table 4.3 and their corresponding precision-recall curves shown in Figure 4.14.

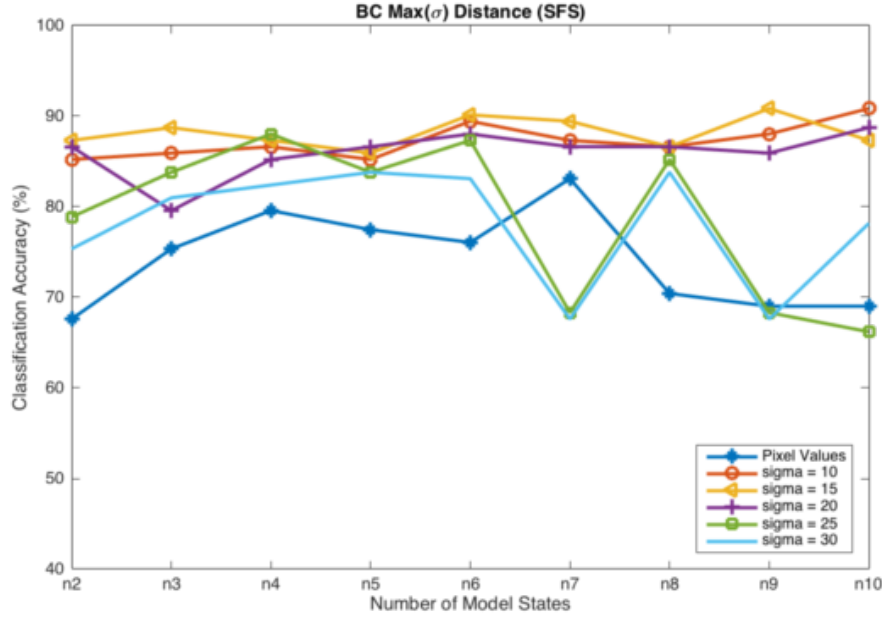


(a) Classification accuracy obtained using the BC max singular values kernel.

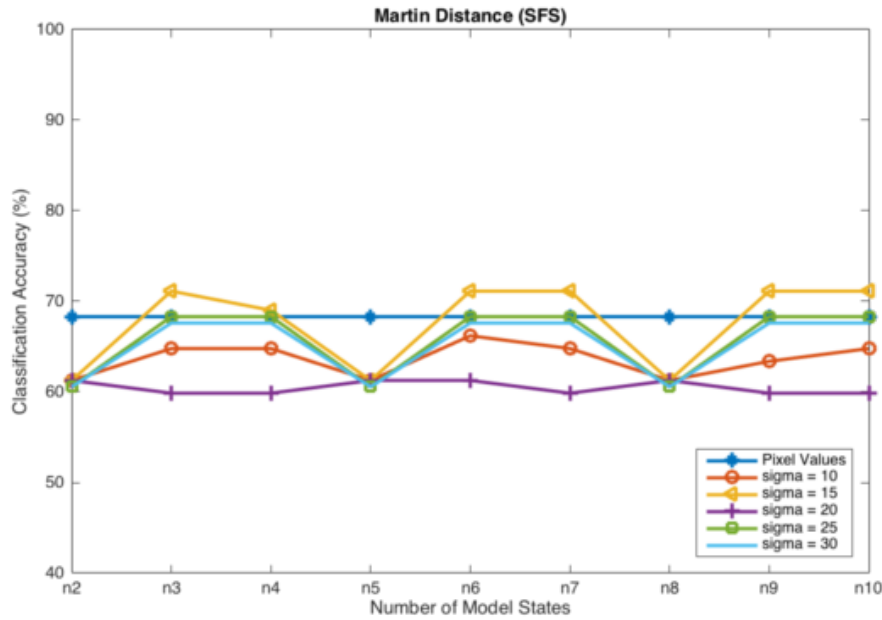


(b) Classification accuracy obtained using the Martin distance.

Figure 4.11: Variation of classification accuracy for the two distance metrics when using feature symmetry. Classification accuracy is shown (y-axis) using a system order of $n=2$ to $n=10$ (x-axis). The plot shows the effect of changing model order when using the pixel intensities versus using the feature symmetry maps of centre-wavelength=5 to 20.



(a) Classification accuracy obtained using the BC max singular values kernel.



(b) Classification accuracy obtained using the Martin distance.

Figure 4.12: Variation of classification accuracy for the two distance metrics when using signed feature symmetry. Classification accuracy is shown (y-axis) using a system order of $n=2$ to $n=10$ (x-axis). The plot shows the effect of changing model order when using the pixel intensities versus using the signed feature symmetry maps of centre-wavelength=10 to 30.

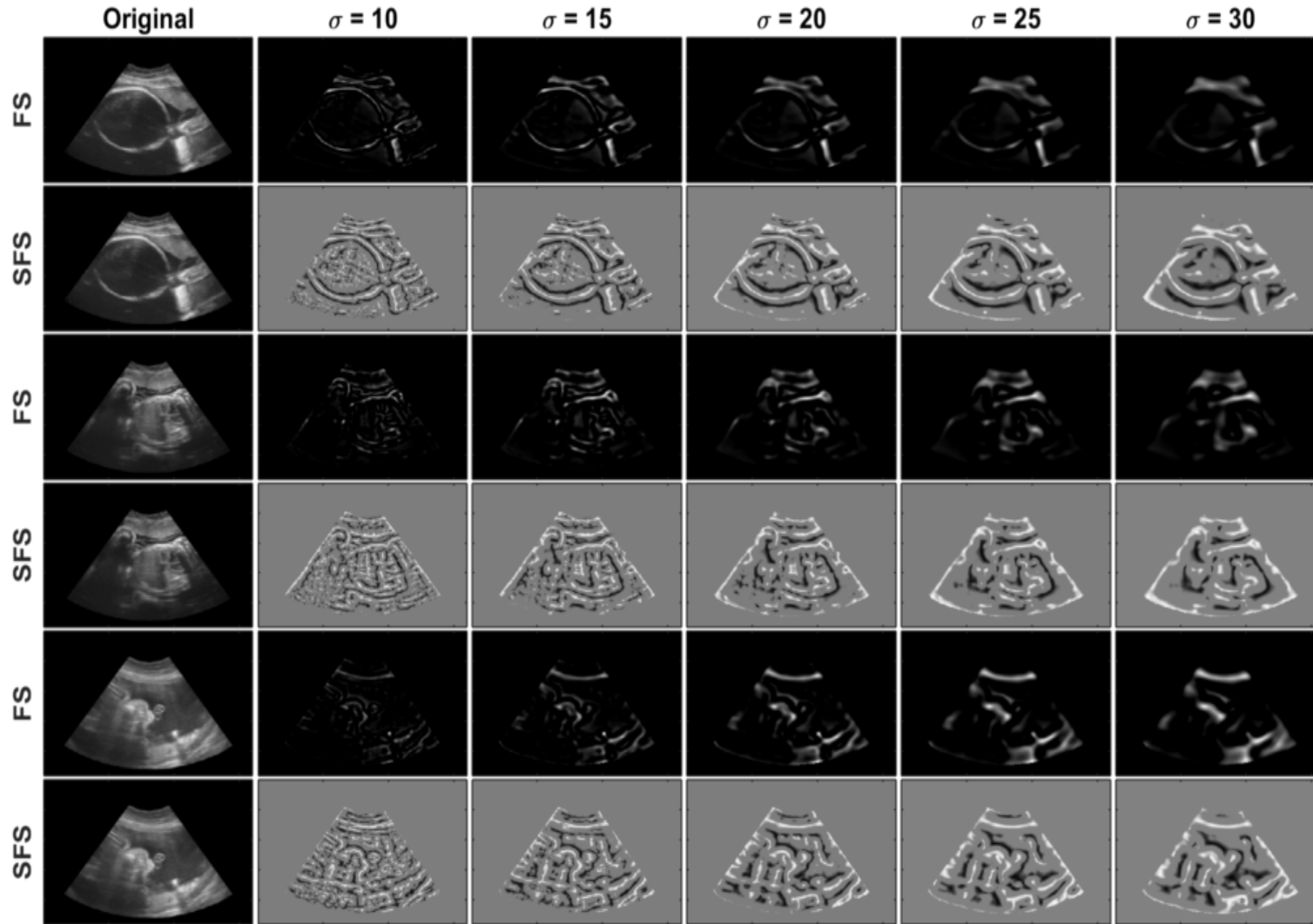


Figure 4.13: *Feature symmetry (FS) and signed feature symmetry (SFS) example plots with increasing value of the log-Gabor centre frequency parameter σ . SFS value lies in the range -1 to 1, where -1 indicates a highly symmetric trough (black), 0 indicates no symmetry (grey), and 1 indicates a highly symmetric peak (white). FS values do not take into account the sign, so symmetric local structures are in white and unsymmetrical structures are in black.*

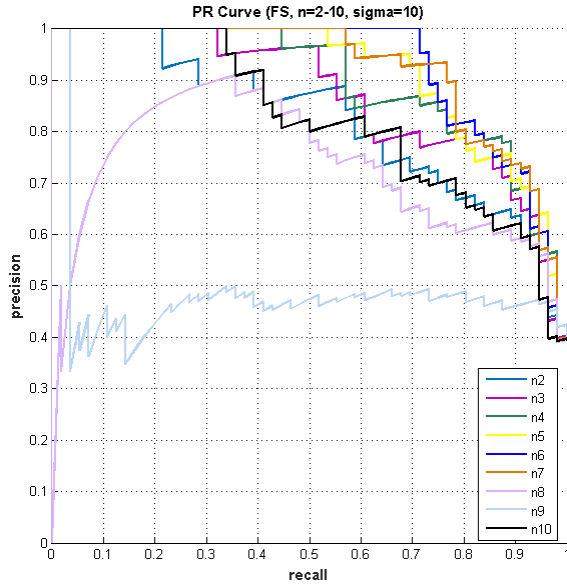
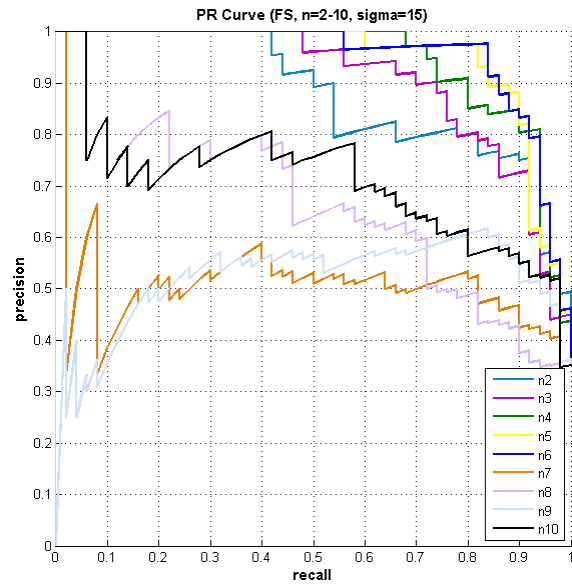
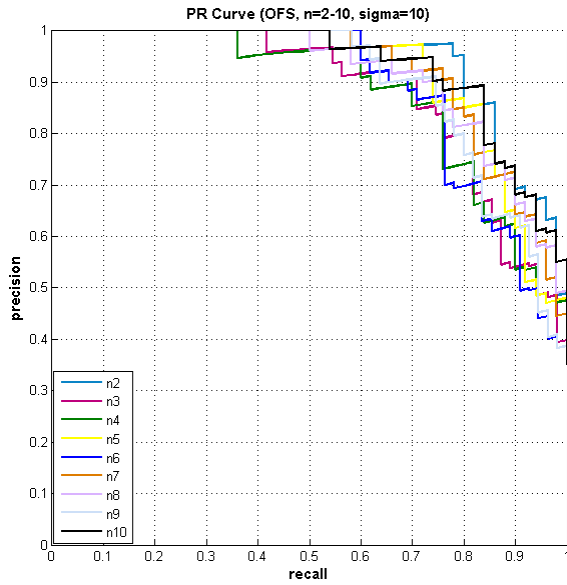
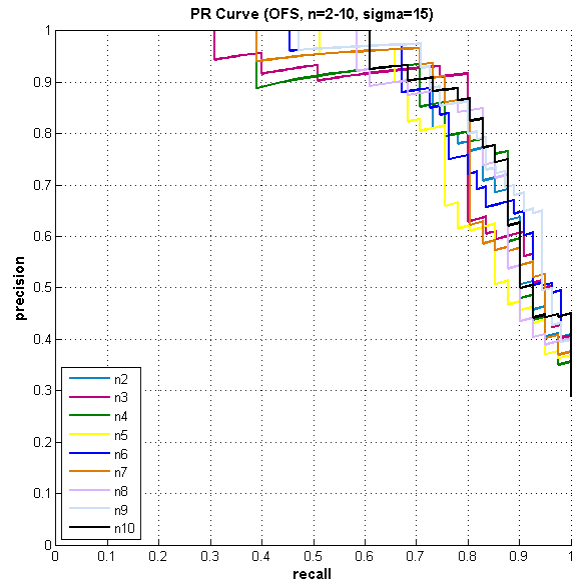
(a) *pr bcinit n all fs sigma 10*(b) *pr bcinit n all fs sigma 15*(c) *pr bcinit n all ofs sigma 10*(d) *pr bcinit n all ofs sigma 15*

Figure 4.14: Precision-Recall Curves. FS with center-wavelength of 10(top left) and 20 (top right) for Model order of $n=2$ to 10 and SFS curves with the same settings at the bottom.

It is important to note that large σ values will smooth the image excessively and this can have an adverse effect on the classification outcome. This is illustrated in Figure 4.11. Furthermore, Figure 4.13 illustrates the effect of various σ values when

using feature symmetry and signed feature symmetry on a number of example frames. Moreover, given that the duration of the sequences are not long, increasing the system orders leads to a decrease in accuracy and potentially over-fitting as the LDS systems become unstable.

4.6 Conclusion

Ultrasound image acquisition and interpretation can be a challenging task for non-experts. This chapter presented a method, validated on both phantom and real-world clinical data, to identify short sequences of interest as a first step to tackle this grand challenge. The results demonstrate that our proposed method outperforms previous work on structure detection in ultrasound videos.

In summary, we have proposed a general framework for ultrasound video analysis, motivated by our interest in ultrasound in pregnancy and perinatal care, which couples standardised, and possibly non-conventional scanning protocols, with image analysis methods designed to extract predefined useful information from ultrasound videos. As a specific application, this framework can be used for the detection of breech presentation, in pregnancies of 28 week and older, given the data predefined scanning protocol. We emphasise that the accuracy achievable in other applications will be dependent on both the choice of scanning protocol and to a lesser extent (in our experience) on the ultrasound equipment.

Chapter 5

The Clinical Application Of Analysing Linear Ultrasound Video Sweeps: Detecting Fetal Presentation and Heartbeat

5.1 Introduction

Confirmation of pregnancy viability (presence of fetal cardiac activity) and diagnosis of fetal presentation (head or buttock in the maternal pelvis) are the first essential components of ultrasound assessment in obstetrics. The former is useful in assessing the presence of an ongoing pregnancy and the latter is essential for labour management. In this chapter, we describe an automated framework for detection of fetal presentation and heartbeat from a predefined free-hand ultrasound sweep of the maternal abdomen, based on the framework presented in Chapters 4 and 3. This is a step towards enabling novice sonographers to identify *high-risk* pregnancies. Throughout our experiments, we illustrate the clinical value of the proposed framework on a challenging ultrasound

video dataset.

An early version of the work presented in this chapter has been reported in:

Maraci MA, Bridge CP, Napolitano R, Norris T, Aye CYL, Molloholli M, Papageorghiou A, Noble JA. Towards Automating the Six Step Basic Ultrasound Scan. Proceedings of the 25th World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2015, 11-14 October 2015, Montreal, Canada. (Clinical Abstract, **Selected for Oral Communication and Conference Press Release**)

Maraci MA, Bridge CP, Napolitano R, Papageorghiou A, Noble JA. A Framework to Analyse Linear Ultrasound Videos for Detecting Fetal Presentation and Heartbeat. Submitted to *Medical Image Analysis*.

5.2 The Framework

The full algorithm we have developed is shown schematically in Figure 5.1. In step A, a multi-class discriminative classifier - trained using the data in T^A - is deployed to categorise ultrasound data into four classes of fetal structure: skull, abdomen, heart and "other". At test time, given a set of unseen video frames, a pre-trained classifier is used to categorise the data into the four classes. Considering the typical heartbeat frequency of a fetus and scan speed (30 fps) employed in this work, it is assumed that fetal heart motion can be captured in at least 30 frames, if it indeed exists. Therefore, if 30 or more consecutive frames are classified as a fetal heart in step A, they are passed on to step B to identify whether the fetal heart beats or not. In this step, a kernel dynamic texture classifier is trained based on training sequences in T^B , where positive samples in the training set are short videos of a beating fetal heart, and negative samples are sequences that do not contain a fetal heartbeat. Moreover, it is important to note that as the ultrasound videos are intentionally kept simple and general, the likelihood of having a long sequence of a fetal heartbeat is low. The next subsections describe each of steps in more detail.

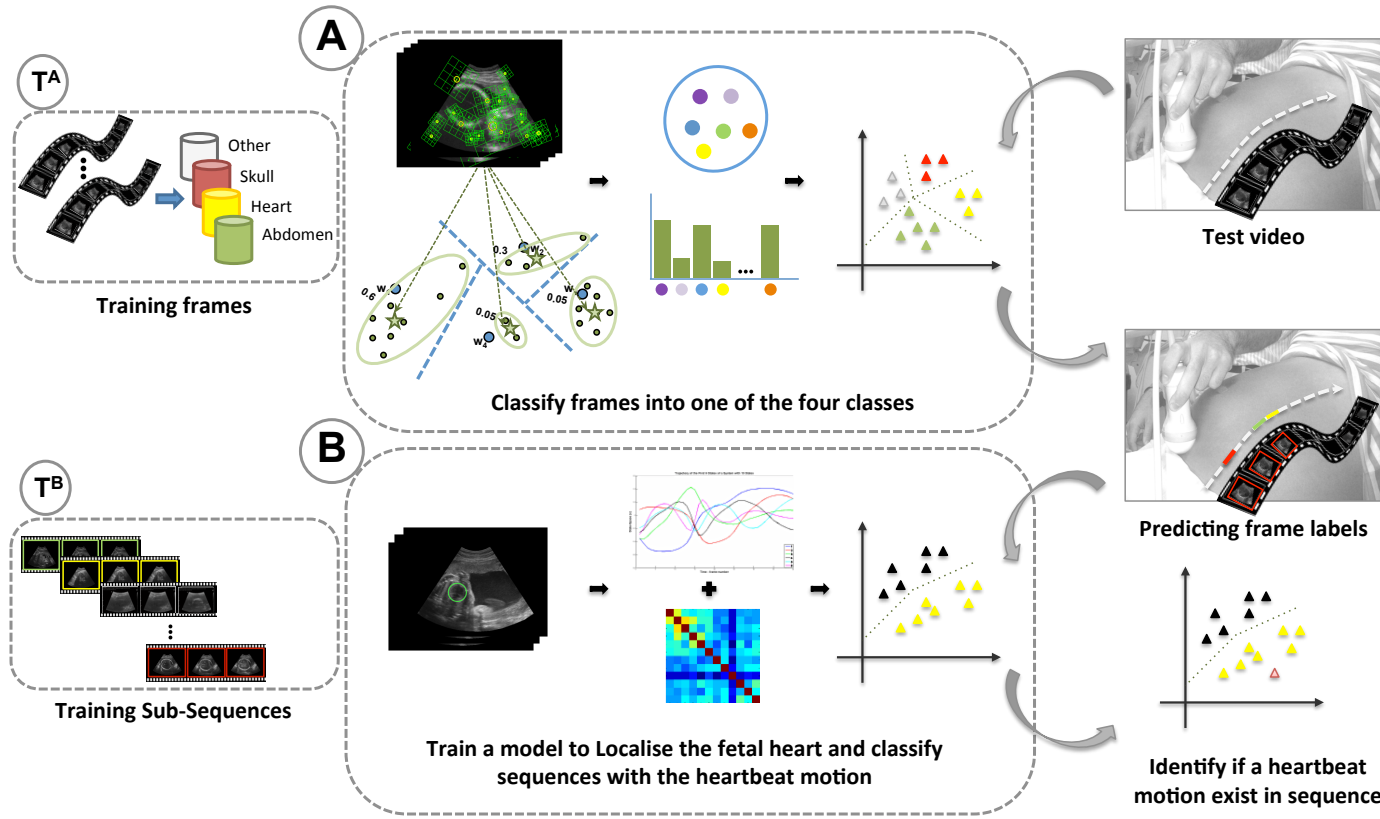


Figure 5.1: *The main steps of the framework.* Given a new training video, all the frames are first classified into “skull”, “abdomen”, “heart” or “other”. If a set of consecutive fetal heart frames are detected in step A, they are further analysed in step B to identify whether a heartbeat can be found. In step T^A , the green colour represents the training dataset of frames corresponding to the fetal abdominal class, yellow indicates the training dataset of fetal hearts, red indicates the dataset of fetal skulls and white indicates the dataset of frames which belong to the “other” class.

Classifying Ultrasound Video Frames

As part of the first step, each video frame is represented through a FV encoding [160,161] of a set of dense rootSIFT features as described in Chapter 3. A pre-trained SVM classifier then categorises each video frame into *fetal skull*, *fetal abdomen*, *fetal heart* or *other structures*. Furthermore, the results are concatenated to characterise the entire video.

As the data used in this study consists of an ordered sequence of frames, temporal information is used to regularise the classification results. In order to utilise this temporal information a conditional random field (CRF) graphical model [109] is constructed, where each frame of the video is represented as a node in the graph. Here the classification scores for each frame are converted into probabilities and used as the node potential in the graph. This setting will smooth out the classifier scores by taking into account the neighbouring frames, where the joint probability of an assignment to all the nodes f_i (variables) is defined as the normalised product of a set of non-negative potential functions as follows,

$$p(f_1, f_2, \dots, f_N) = 1/Z \prod_{i=1}^N \phi_i(f_i) \prod_{e=1}^E \phi_e(f_{e_j}, f_{e_k}). \quad (5.1)$$

Here we have a potential function for each node i , $\phi_i()$, and edge e , $\phi_e()$, in the graph where (f_{e_j}, f_{e_k}) represents an edge between nodes j and k . As each frame of the video is treated as a node in our graphical model, the node potential $\phi_i()$ for that frame is set to the probability scores obtained from the first step. The edge potential function $\phi_e()$ between any two nodes is the probability of a node transitioning from one state to another and has been empirically set based on the videos in the training dataset. Moreover, Z is the normalisation constant to ensure the distribution sums to one over all possible joint configurations of the variables. Finally, the Viterbi [66] algorithm is used to find the most probable classification result for each frame.

Localising Fetal Heart

The frame classification procedure described above is able to identify the frames containing the fetal abdomen. But in order to assess fetal viability, it is desirable to detect the location of the heart within these frames. This task is complicated by the fact that, in our predefined-freehand scans, the orientation of the heart relative to the probe is variable and unknown. We therefore chose to make use of rotation invariant detection methods as suggested in [21].

Rotation Invariant Features

The method is based on the use of a set of complex-valued rotation invariant basis functions, b , that have a particular form that is described in polar coordinates (r, θ) by the product of a radial profile $p(r)$ and a Fourier basis on the angular coordinate, θ [21, 118]:

$$b_{j,k}(r, \theta) = p_j(r)e^{ik\theta} \quad (5.2)$$

for $0 \leq r < R$, $0 \leq \theta < 2\pi$, where j indexes a set of different *radial profiles*, $p_j(r)$. Notice that, while the form of the radial profile is general, the angular part of the separable form of the basis function must be the Fourier basis in order to achieve the desired rotation invariance. Figure 5.2 illustrates a set of basis functions.

In order to use the framework with a vector-valued image representation (such as the gradient), $v(\mathbf{p})$, we must first express the orientation of the vectors in a Fourier orientation histogram. This represents an orientation histogram as a truncated set of M Fourier series coefficients, rather than a set of discrete bins. The m^{th} coefficient at image position $\mathbf{p}$ is:

$$c_m(\mathbf{p}) = \|v(\mathbf{p})\|e^{-im \arg(v(\mathbf{p}))}, \quad m = 0, 1, \dots, M. \quad (5.3)$$

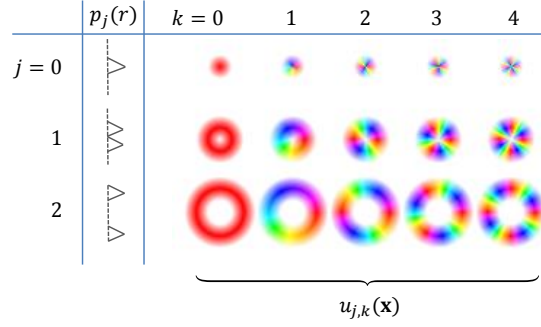


Figure 5.2: *Set of profiles and basis functions with $j = 3$, $k = 4$ (only $k \geq 0$ displayed). The saturation and hue represent the complex magnitude and argument respectively. Image adapted from [21]*

When working with discrete images, we sample the basis functions on a rectangular grid and use them as a filter kernel on the Fourier histogram images. One feature with parameters j, k, m describing the window centred at position $(\mathbf{x})$ is therefore given by,

$$D_{j,k,m}(\mathbf{p}) = b_{j,k}(\mathbf{p}) * c_m(\mathbf{p}), \quad (5.4)$$

and a complete description of a window is built up by using a number of such basis functions. In our experiments, parameters j, k, m are empirically set to 6, 4, 4 respectively. As a result of the shift property of the Fourier series, the complex magnitude of the resulting features are analytically invariant to the orientation of the underlying image.

For classification of each window as heart or non-heart we use a linear SVM classifier with the rotation invariant quantities as input. At test time, each pixel in each frame is assigned a classification score as the output of the SVM classifier, reflecting the probability of belonging to a heart. For each image location, we simply sum these scores across frames to get a total score for each pixel, and choose the pixel with the highest score to be the location of the centre of the heart.

Detecting the fetal heartbeat

Once the video frames that contain a fetal heart are identified and the heart is localised, consecutive frames are compiled into shorter sub-sequences. To identify whether a fetal heartbeat motion can be detected in these sub-sequences, the feature trajectories (dynamics) of the sequences of frames, $\{y_t\}_{t=1}^T$, are modelled as the output of a linear dynamical system (LDS). We follow Doretto et al. [54] for the system identification of the model as described in Chapter 4. In this model, the appearance of each video frame is determined through the observed variable and the motion and dynamics in the video over a given time is determined through the hidden-state variables, which are sampled from a Gauss-Markov process. Furthermore, the observed frame at any given time can be constructed from a linear combination of the hidden state variables. Estimating KDT model parameters is explained in more detail in Chapter 4.

Distance Metrics

Given a KDT model estimate for each sub-sequence, a suitable metric now needs to be defined to assess similarity between any two sub-sequence models. As discussed in Chapter 4, prior work on comparison metrics of LDSs show on a number of applications that the similarity metrics based on the Martin Distance and Binet-Cauchy maximum singular values kernel produced the best results [35]. Moreover, we have previously shown [129] that the Binet-Cauchy maximum singular values kernel produced superior results on medical ultrasound data, and it was also used here in our experiments.

5.3 Experiments and Results

Experiments were designed to evaluate the accuracy of the proposed framework to detect the fetal presentation and heartbeat. From the total of 323 ultrasound video sweeps, 60% (194 videos) were used for training in the framework and the remaining

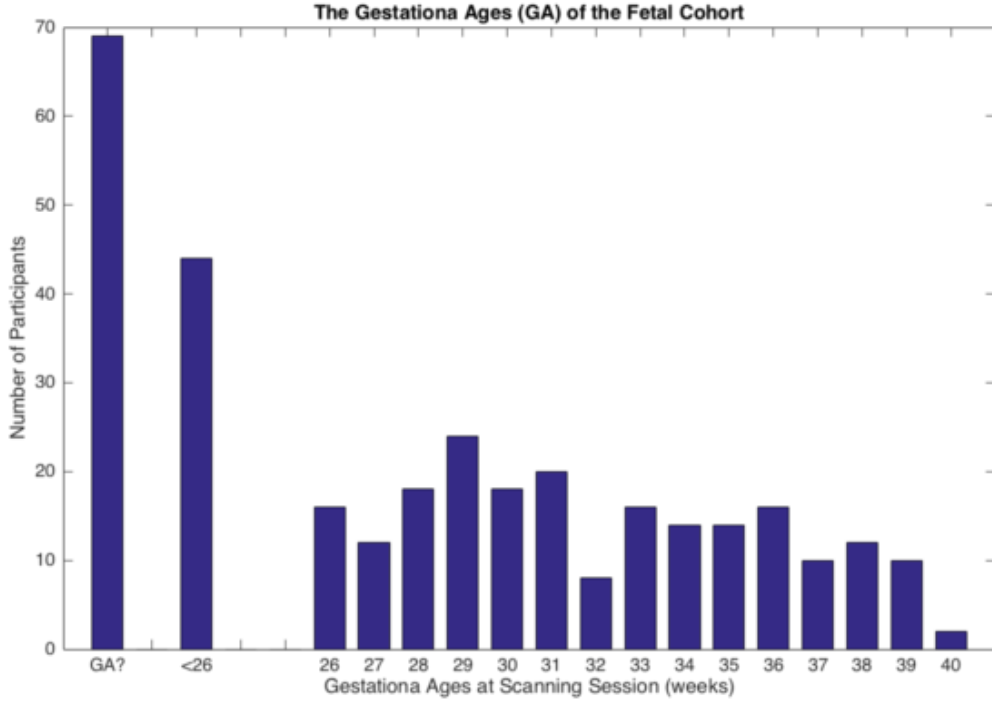


Figure 5.3: *Gestational age spread of the diverse clinical fetal cohort.* Bar chart of the number of subjects with unknown gestational age and gestational age of younger than 26 weeks (on the left) and bar chart of the number of subjects at each gestational age (on the right).

40% (129 videos) were used for testing. The participants were from a diverse set of ethnic backgrounds and at various stages of pregnancy. The data acquisition was carried out using a Philips ultrasound probe as described in Chapter 3. Furthermore, the distribution of the fetal ages in this dataset are illustrated in Figure. 5.3.

From 129 test videos, 58.9% contain the fetal skull and abdomen structure, which are essential for detecting the fetal presentation. Moreover, 37.9% of the test videos did not contain either the fetal skull (11.6%) or the fetal abdomen (26.3%) and the presentation could not be detected in them. Finally, only 26% or (84/323) of the entire dataset contain a fetal heart sequence of 30 frames or longer. In what follows, the first experiment evaluates the accuracy of the frame classification task. Given the entire set of collected ultrasound videos, a dataset of 4 classes (fetal skull, abdomen, heart, other) was constructed, where all the individual frames containing any of the four classes are extracted and placed in the appropriate class and the classification accuracy

reported. The second experiment evaluates the accuracy of fetal heart localisation in the relevant frames and the third experiment evaluates the accuracy of detecting the fetal heartbeat.

5.3.1 Classification of Individual Frames

SIFT patches of 8×8 , 16×16 , 32×32 and 64×64 pixels with a stride of 4 pixels on each 240×320 image were calculated, repeated at 9 different scales with a scaling factor of $\sqrt{2}$. Moreover as ultrasound data is only visible within the ultrasound fan (field of view), SIFT features were only computed within a bounding box around this region to reduce the chances of calculating redundant features. A schematic of the different size descriptors is presented in Figure 5.4a, where the red box indicates the bounding box regions in which the SIFT descriptors are computed. Classification results are summarised in Table 5.1. As in [96], the features are decorrelated and reduced in dimensions from 128D to 40D and also square rooted as this has been shown to improve classification results [8]. Moreover as some structures of interest are relatively small, such as the fetal heart and abdominal structures, the image was divided into 1×1 and 2×2 spatial subdivisions as shown in Figure 5.4 and corresponding FV computed and stacked as one, to investigate whether detection accuracy could be improved.

In order to ensure training and test data are independent, 60% of the videos were randomly selected for training and the remaining 40% for testing. The model was therefore trained on 4500 randomly selected images from each class ($\approx 18k$ frames over all classes) and tested on 480 randomly selected unseen images from each class (≈ 1600 frames over all classes). The mean accuracy and mean average precision (mAP) have been calculated and summarised in Table 5.2. In the table, L_1 indicates no spatial subdivisions and L_5 indicates the additional 4 spatial subdivisions.

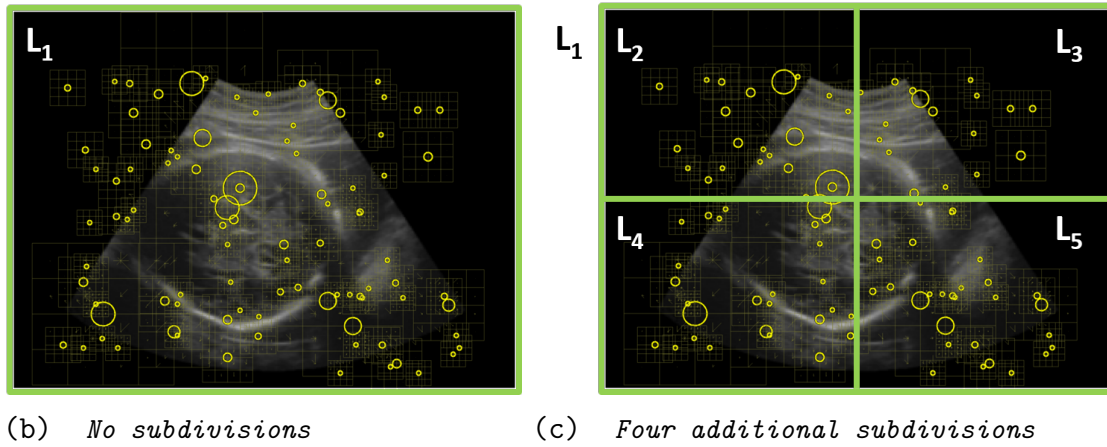
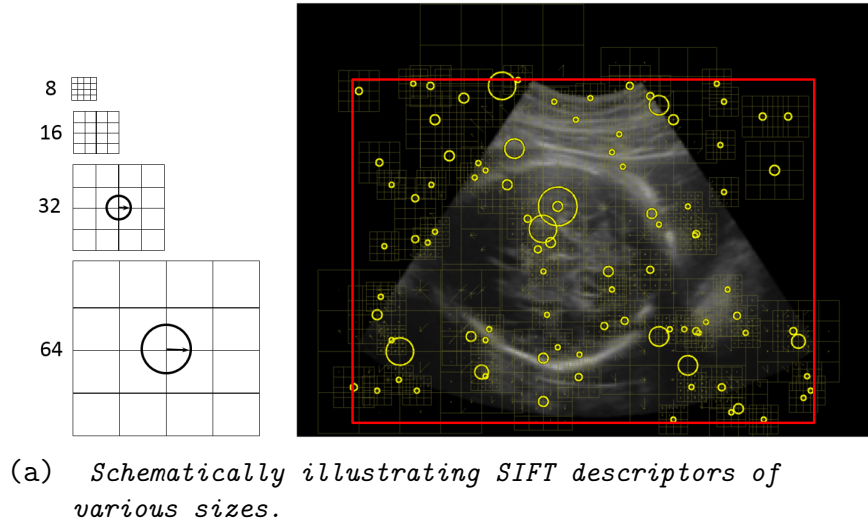


Figure 5.4: (a) illustrates SIFT descriptors of various sizes including patches of 8×8 , 16×16 , 32×32 , and 64×64 pixels. Moreover SIFT descriptors are only computed within the red box, to stop computing redundant descriptors. (b) and (c) illustrate the spatial subdivisions. L_1 indicates no spatial subdivisions and L_5 indicates a 4×4 subdivision in addition to the L_1 . A small number of randomly selected descriptors are plotted for illustration purpose and not to clutter the image.

Table 5.1: Video frame classification results (Step A). The effect of increasing the SIFT descriptor size from an 8×8 to a 16×16 patch is shown, where the number of the GMM is set to 80 modes and the PCA is used to reduce feature dimensions to 40.

SIFT	Skull Class.	Abdomen Class.	Heart Class.	Other Class.	Mean	Mean Ave.
Patch Size	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Precision (%)
8×8	94.38	89.17	35.21	94.58	78.33	90.01
16×16	95.83	91.04	50.42	97.71	83.75	93.37
32×32	96.46	92.08	60.63	97.92	86.77	94.75
64×64	96.25	86.13	72.92	97.92	87.55	95.25

Table 5.2: Video frame classification results (Step A). The number of GMMs are fixed to 80 words to analyse the effect of varying the feature dimensions from 10 to 80 with L5 and without tiling L1. With the L5 tiling, each image is broken down into 4 parts where a set of feature are computed for that tile. The features are the concatenated together.

No. PCA	Skull Class.	Abdomen Class.	Heart Class.	Other Class.	Mean	Mean Ave.
Components	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Precision (%)
10_{L1}	93.96	76.88	62.92	94.79	82.13	89.93
40_{L1}	95.42	89.38	60.00	97.50	85.57	92.59
80_{L1}	95.42	89.58	49.38	97.29	82.92	92.77
10_{L5}	95.42	83.75	57.71	97.29	83.54	93.32
40_{L5}	96.46	92.08	60.63	97.92	86.77	94.75
80_{L5}	95.21	92.71	51.67	98.96	84.63	94.98

L1 No subdivisions

L5 Five subdivisions

5.3.1.1 Visualising the Results

In this part, a visual representation of the FV encodings and the classification results for a number of examples are presented.

Visualising the FV Encodings

Figures 5.5,5.6,5.7,5.8 illustrate the FV classification scores on a number of examples. Here dense SIFT is calculated on each image with a stride of one pixel (for better visualisation effect) and encoded using FV encoding. Furthermore, a pre-trained classifier is used to categorise them, where positive classification scores are shown in purple and negative scores are shown in green. The class name with the highest classification score is written on the top left corner of each image. This is followed by the second highest

classification score on the second line. If the highest classification score is the correct class, the label font is set to green colour and if it has been identified incorrectly in red.

As can be seen, in each category the positively scored pixels correspond to meaningful anatomical structures. For example in each case the purple response in the top left image has highlighted the fetal skull boundary, and similarly for the abdomen, and heart.

Visualising the Classification

Moreover applying the CRF model to the classification scores regularises the results and eliminates sudden peaks. This is illustrated in Figures 5.10, 5.9, 5.11, and 5.12 where the top bar illustrates the obtained classification scores.

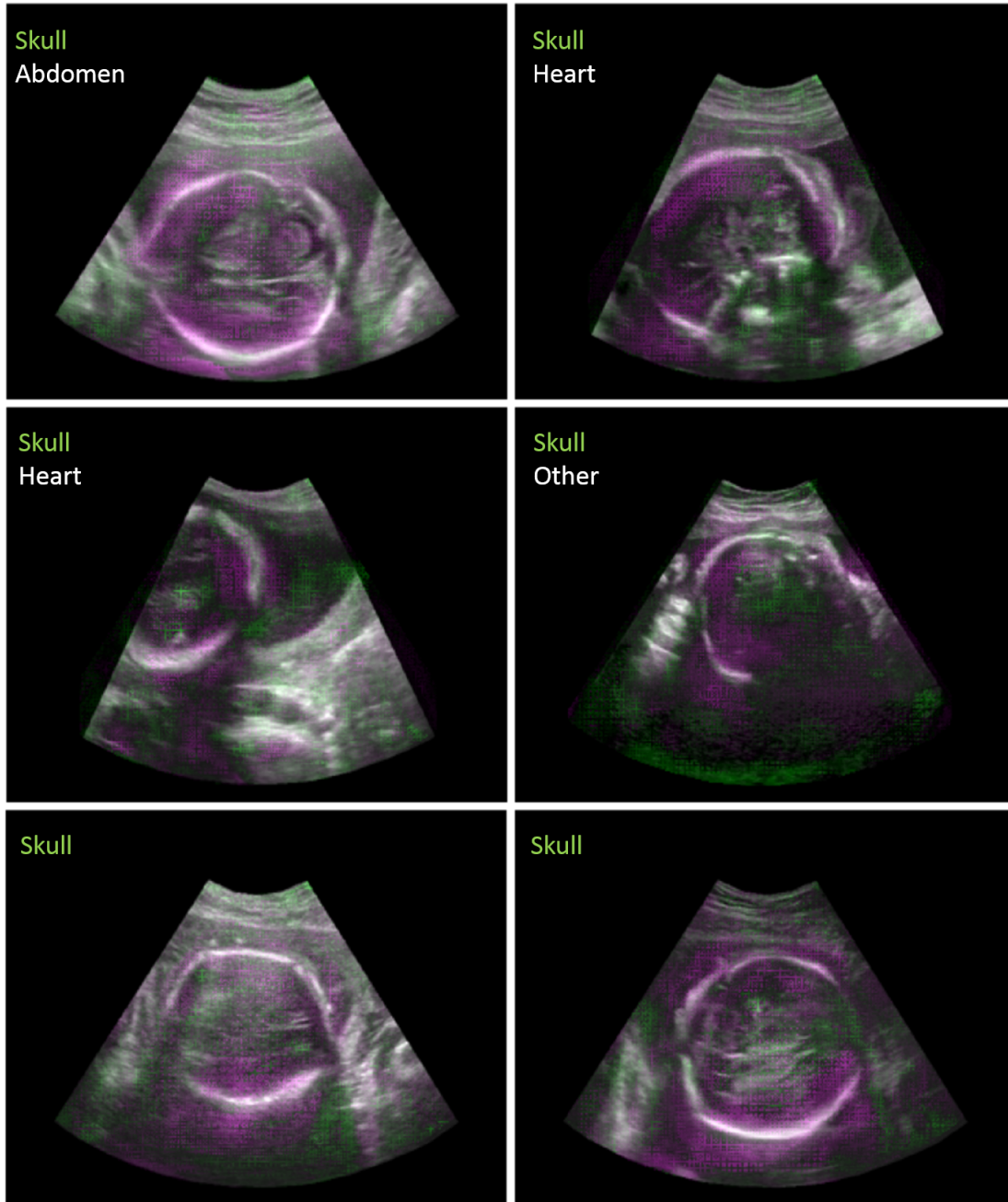


Figure 5.5: *Visualising FV responses on the six examples of the fetal skull. Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores.*

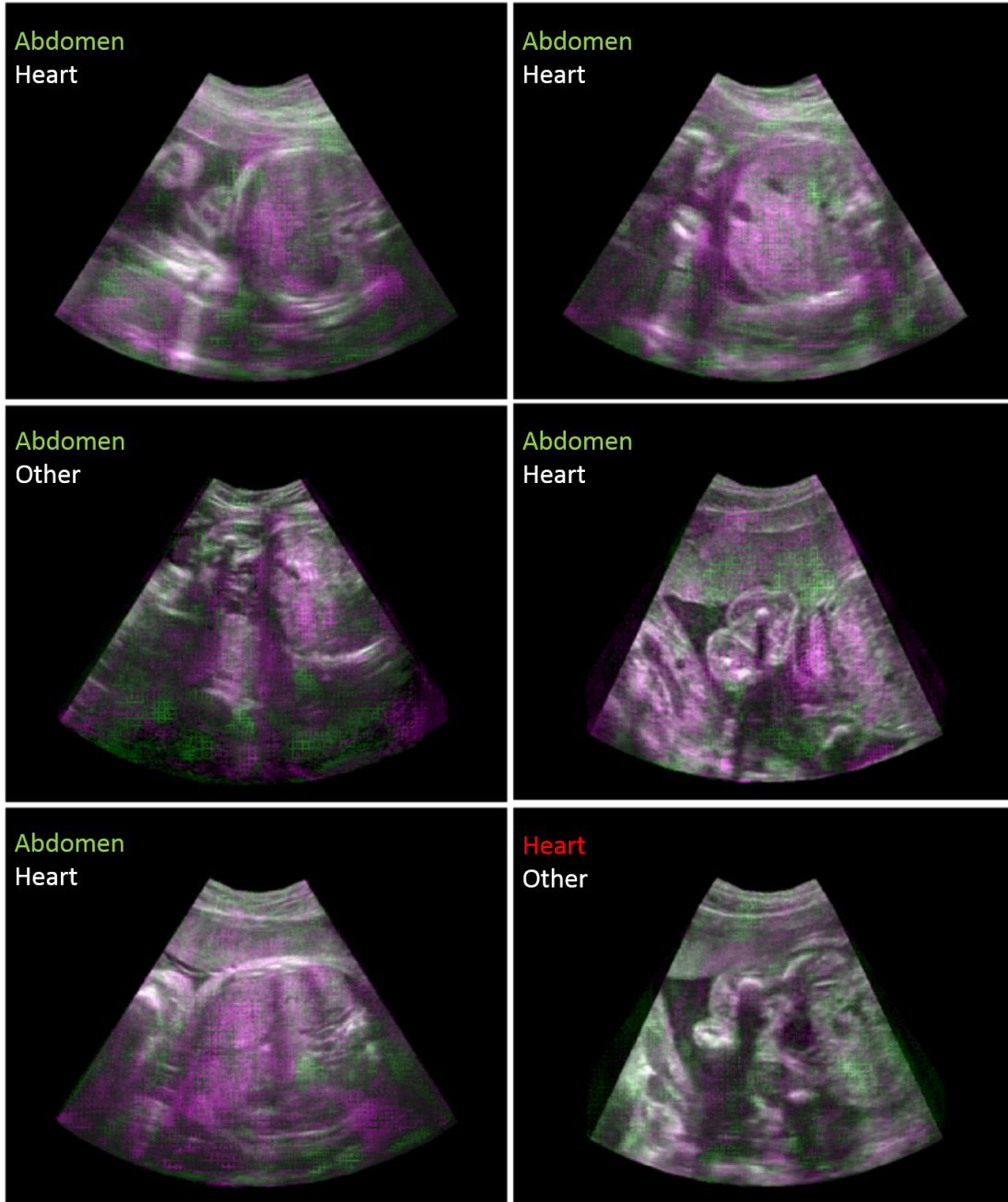


Figure 5.6: *Visualising FV responses on the six examples of the fetal abdomen. Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores. Here the frame at the bottom right corner has been classified incorrectly.*

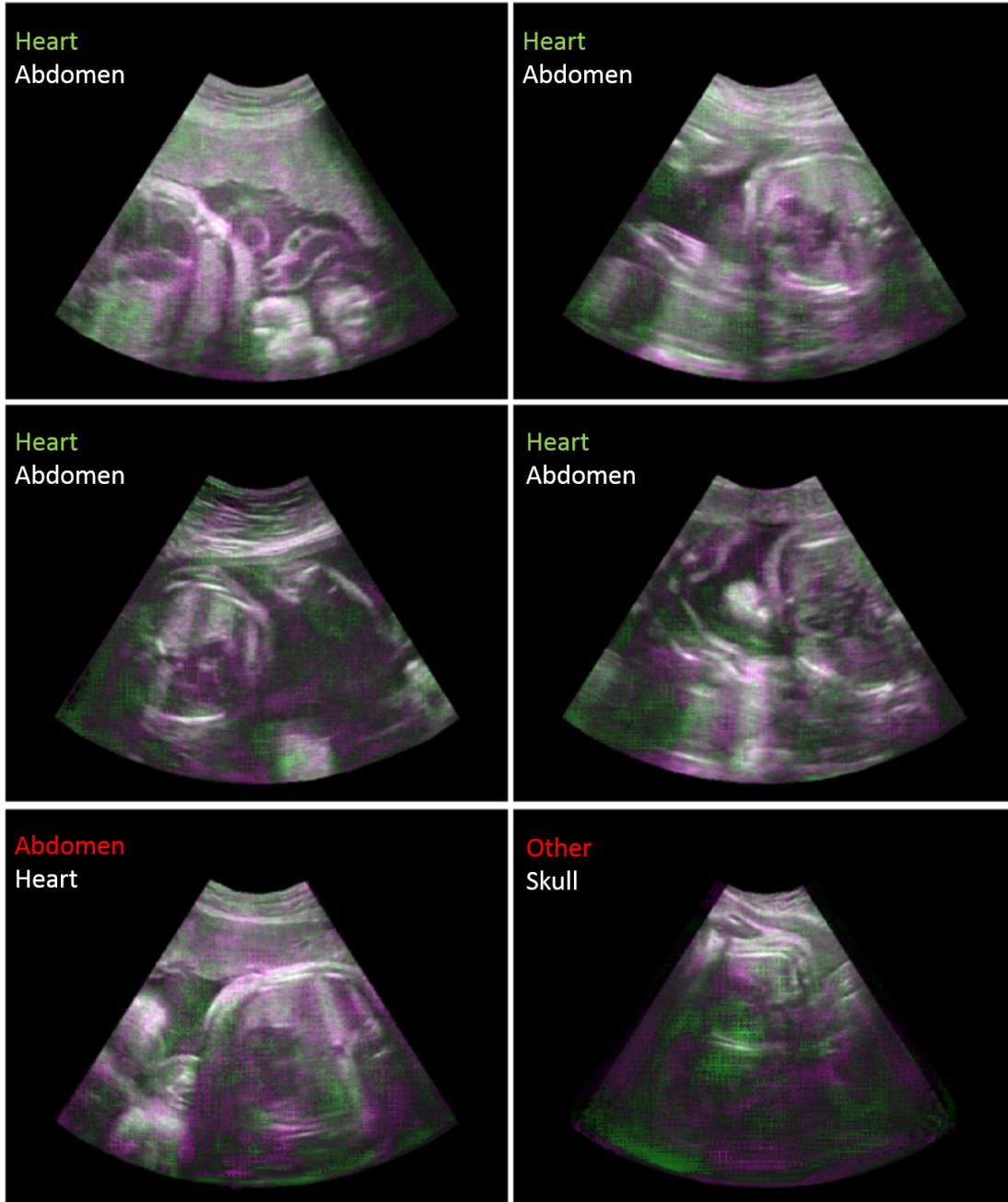


Figure 5.7: *Visualising FV responses on the six examples of the fetal heart. Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores. Here the last two frames have been misclassified.*

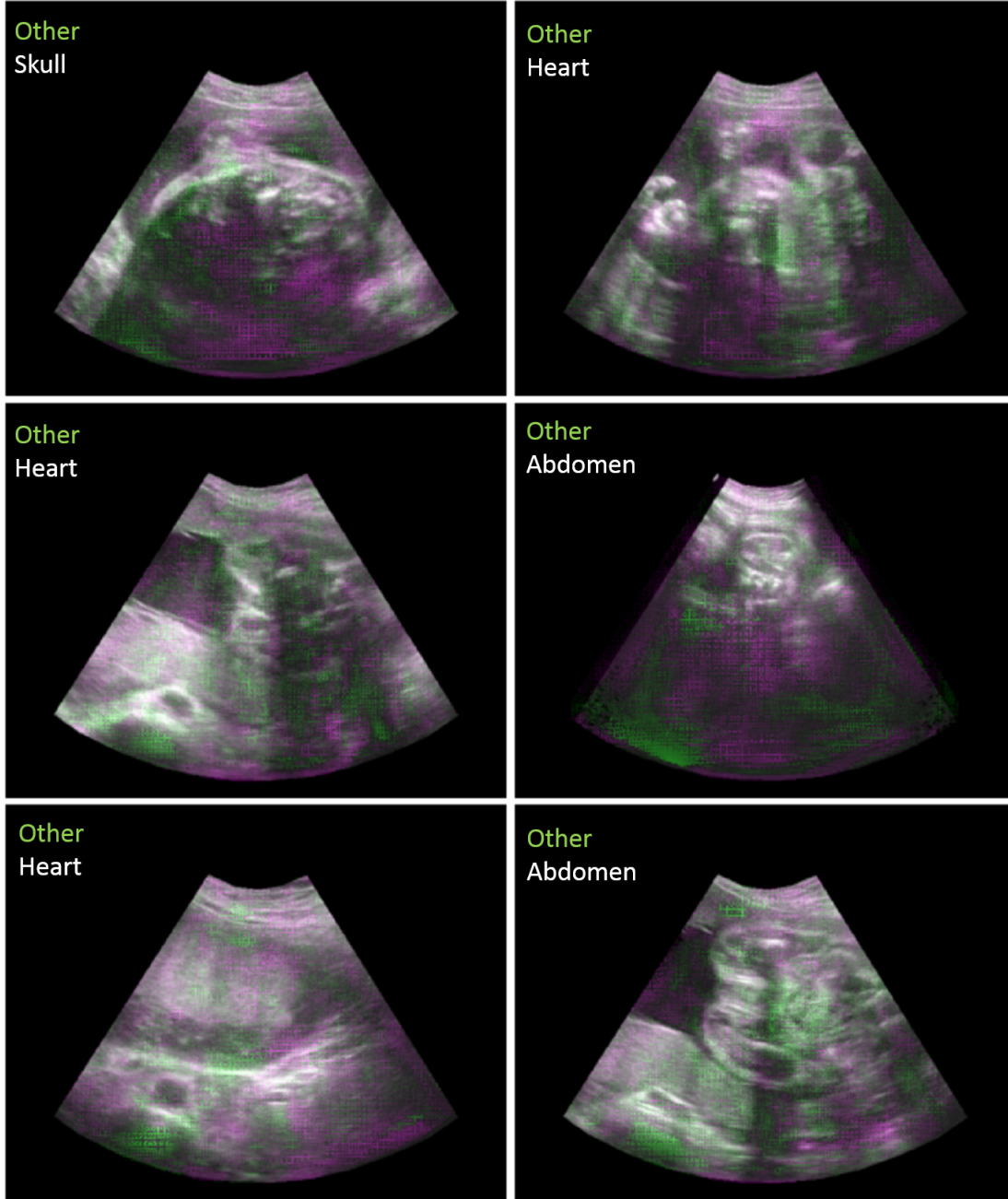


Figure 5.8: *Visualising FV responses on the six examples of the "other" fetal structures.* Dense SIFT features are constructed with a stride of one pixels and then encoded using the FV encoding. A pre-trained classifier scores each FV. Purple indicates the positive and green correspond to negative scores.

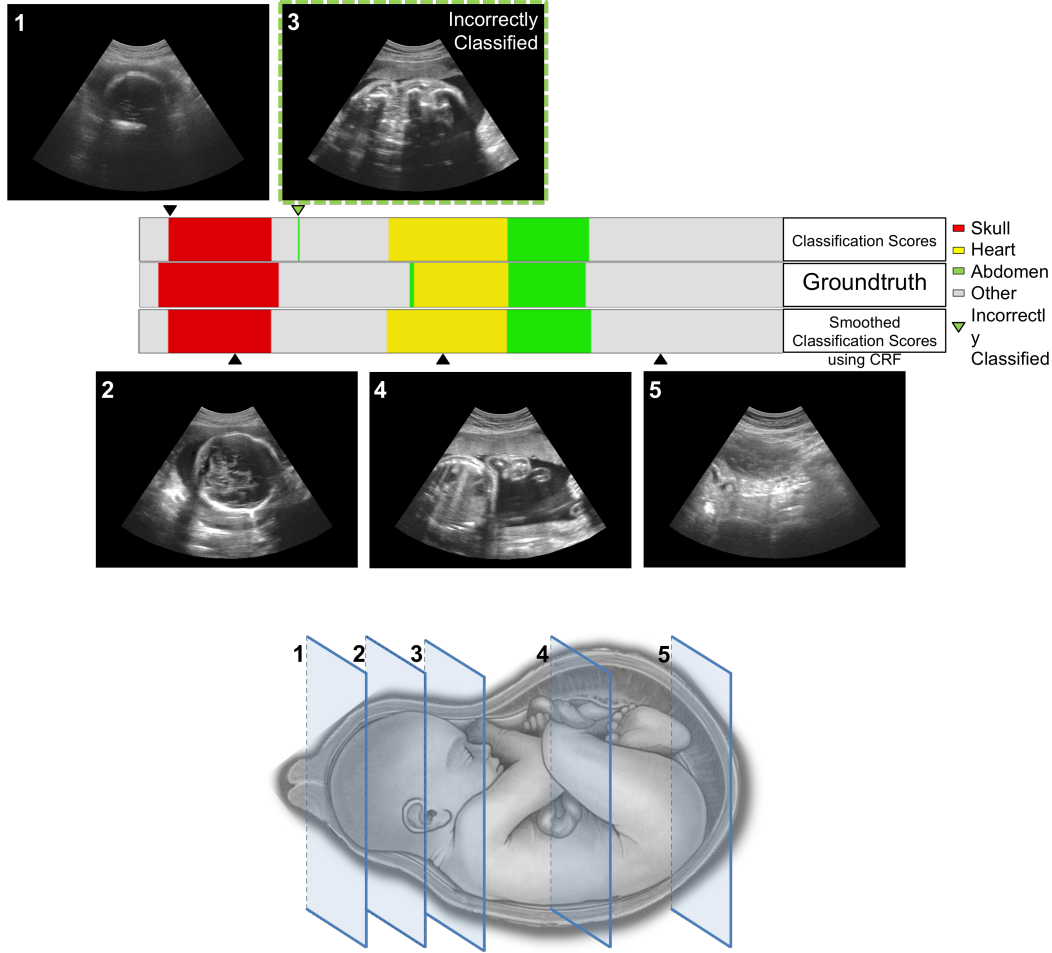


Figure 5.9: *Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). The example frames illustrate the visual appearance of the video at various time points. As it is illustrated, the misclassified frame (green spike on the top bar) has been smoothed out after applying the CRF model (on the bottom bar).*

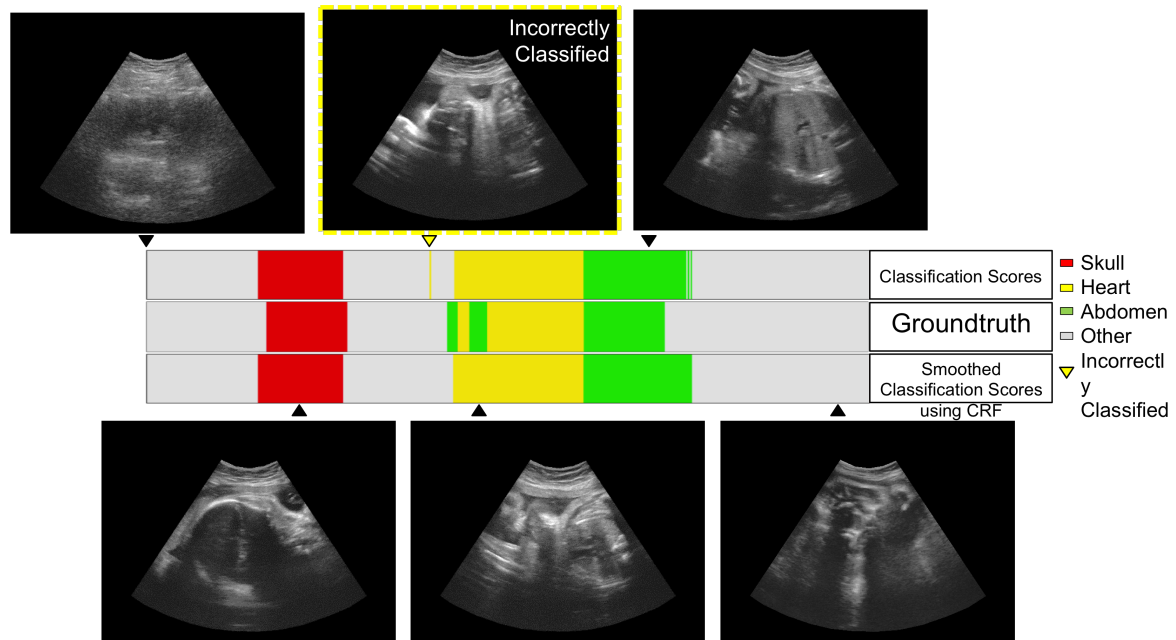


Figure 5.10: *Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). The six sample frames correspond to various time points on the video. As is it shown, the misclassified frame (second frame on the top bar) has been corrected for after applying the CRF model.*

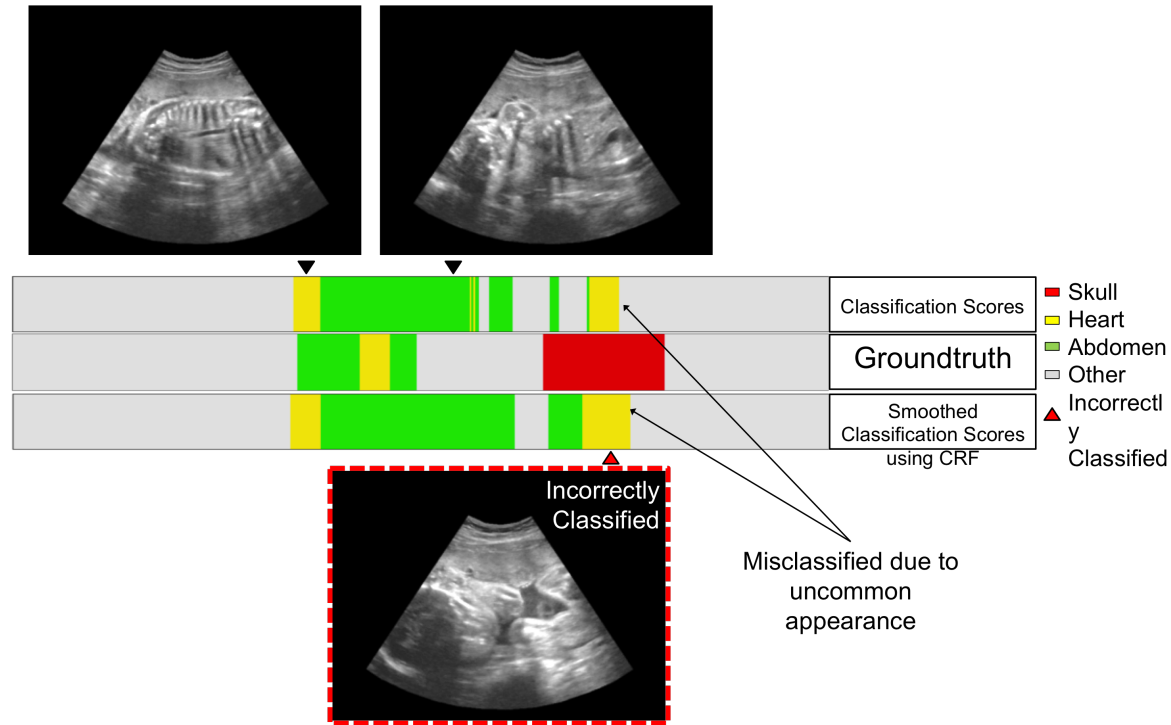


Figure 5.11: *Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). The heart (first yellow block) and abdomen (first green block) are correctly identified, however the second block of frames classified as heart have been incorrectly labelled. The smoothing process does not rectify this error as expected.*

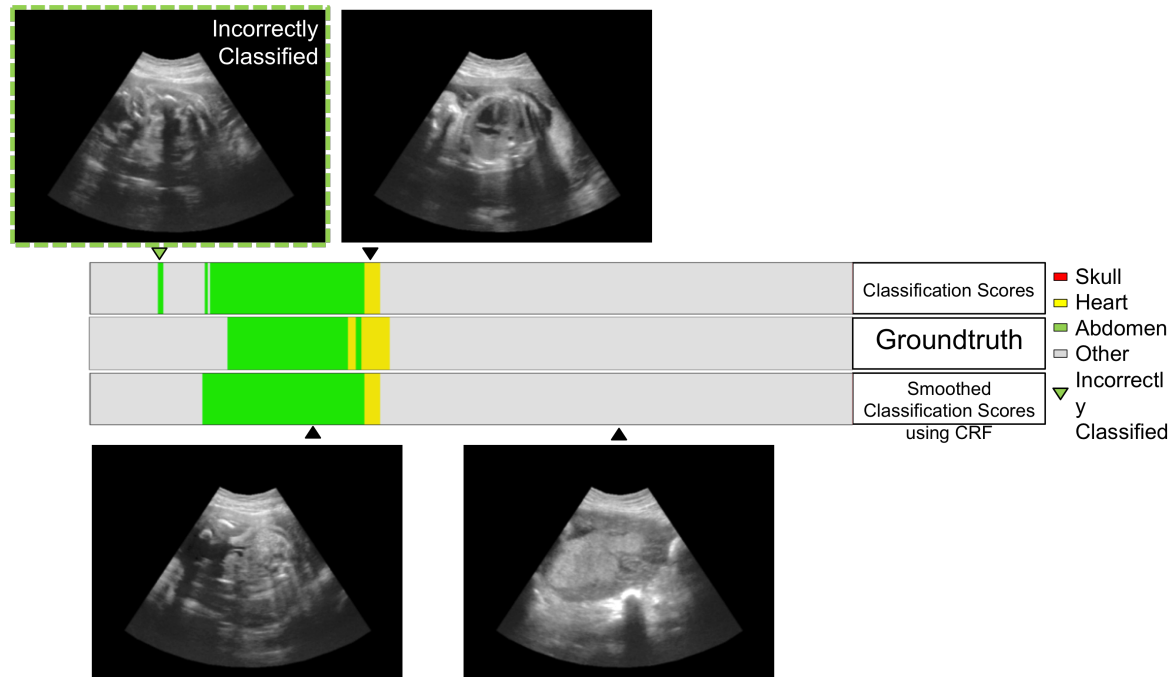


Figure 5.12: *Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). As in the other examples, the green spike on the top bar has been removed as desired after applying the CRF smoothing.*

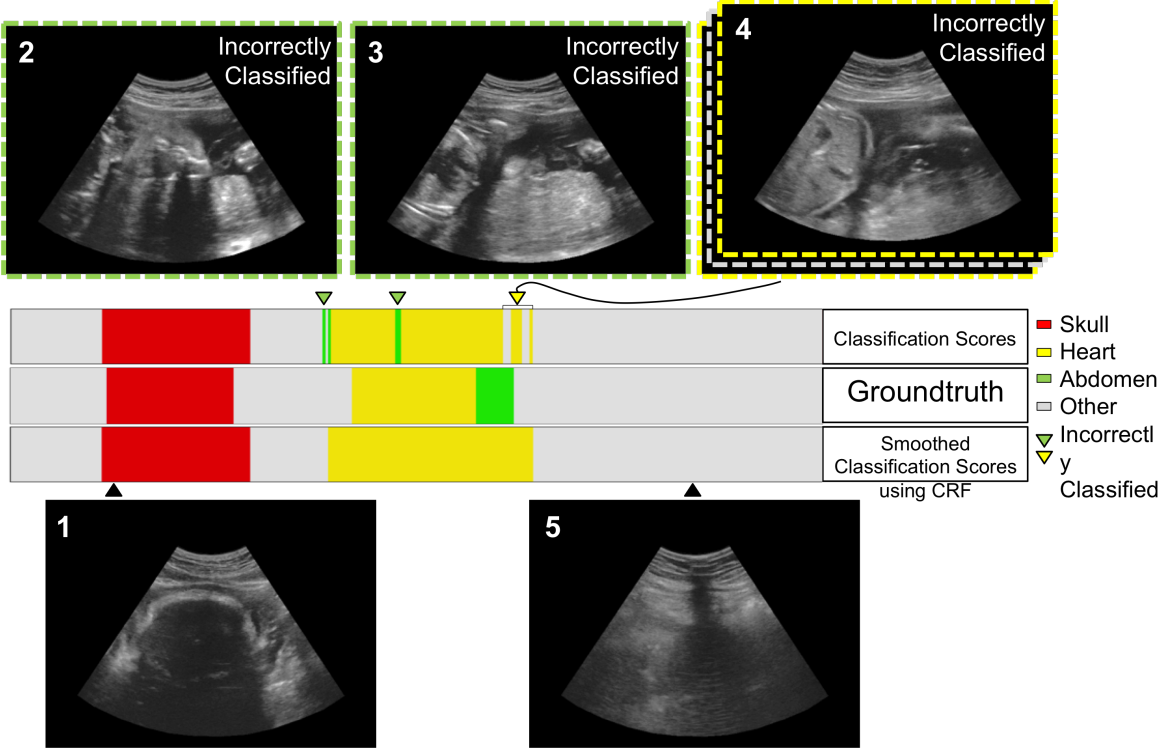


Figure 5.13: Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). As in the other examples, the green spike on the top bar has been removed as desired after applying the CRF smoothing.

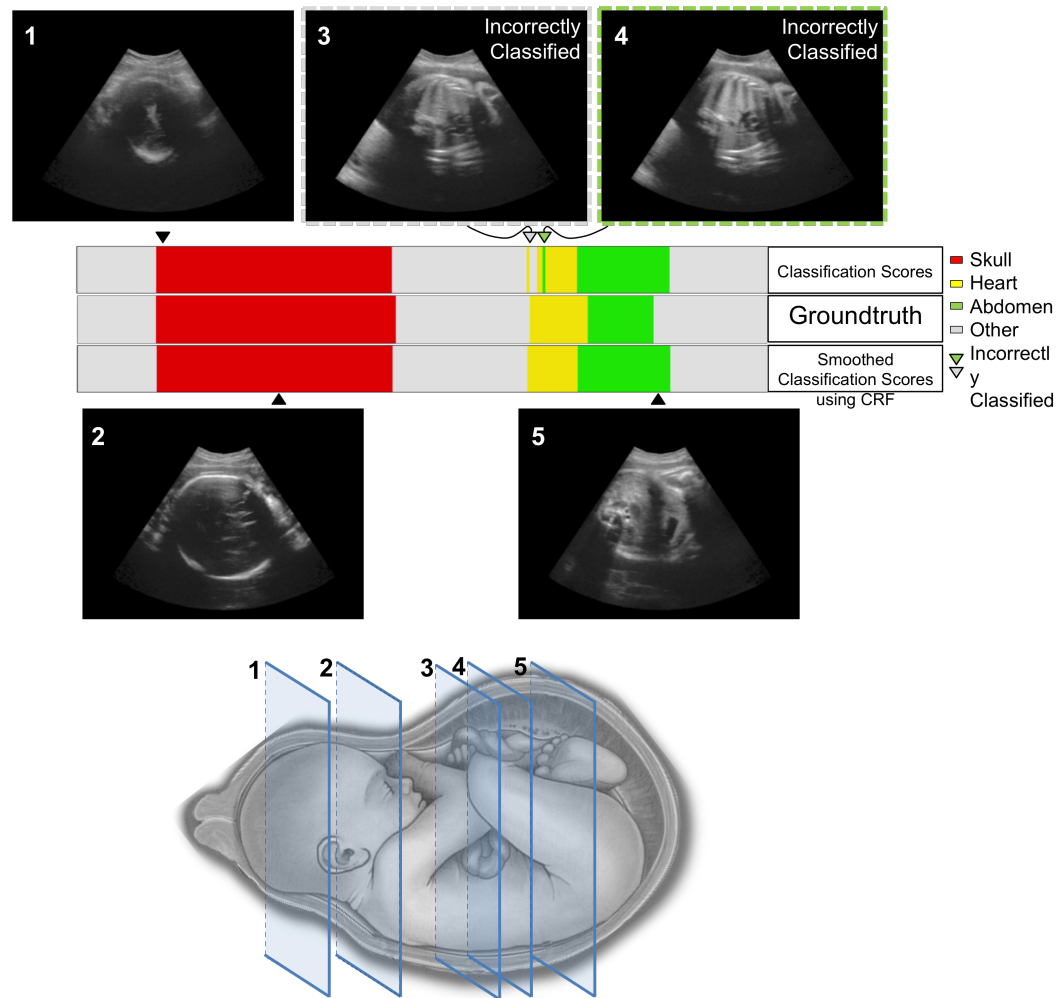


Figure 5.14: *Visualising the classification results (top bar) and smoothed classification results using the CRF (bottom bar). As in the other examples, the green spike on the top bar has been removed as desired after applying the CRF smoothing.*

5.3.2 Detecting the Fetal heartbeat

The algorithm described in Section 5.2 was applied to identify video sweeps that contain at least 30 consecutive frames of the fetal heart. The extracted frames were then compiled into a short sub-sequence with 30 frames. Finally a pre-trained classifier was used to identify whether these short sequences contain a valid heartbeat motion. In order to evaluate the motion detector framework, two experiments were conducted. The first experiment used the entire ultrasound frame (see Figure 5.15a), and in the second experiment the fetal heart was localised first, following the method described above, and each frame was cropped based on a empirically adjusted 60 pixels radius around the detected heart (see Figure 5.15b). Given the size of the fetal heart in our dataset, a 60 pixels radius ensures that the fetal heart is always part of the cropped image.

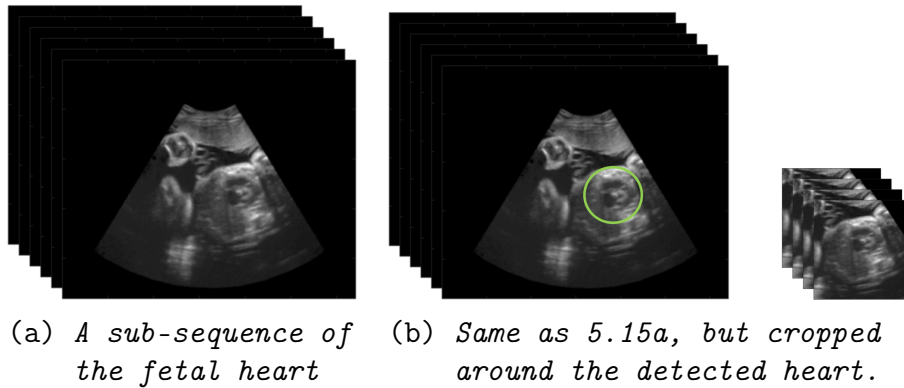


Figure 5.15: *The sequence is cropped around the detected fetal heart in 5.15b to investigate whether this improved the classification accuracy*

Tables 5.3 and 5.4 summarise classification results for these fetal heartbeat sub-sequences. From the videos in the testing dataset, 136 short videos of a fetal heartbeat (of 30 frames long) were extracted. The dynamics of these subsequences was then further analysed to identify whether a fetal heartbeat could be detected.

As shown in Table 5.3 without localising the fetal heart, the best results were achieved with a 3-state model and $\sigma = 50$ and a 4-state model and $\sigma = 45$ for the

Table 5.3: *Classification results for detecting the fetal heartbeat without localising the fetal heart (Step B). The sequence dimensions are $240 \times 320 \times 30$.*

n	3	4	5	6	7	8	9	10
$\sigma = 30$	80.46	81.61	83.91	83.91	79.31	50.57	54.02	56.32
$\sigma = 35$	80.46	81.61	83.91	83.91	83.91	50.57	88.51	58.62
$\sigma = 40$	52.87	82.76	51.72	52.87	45.98	57.47	54.02	85.06
$\sigma = 45$	83.91	88.51	85.06	57.47	68.97	51.72	52.87	52.87
$\sigma = 50$	88.51	86.21	86.21	83.91	52.87	74.71	54.02	51.72
$\sigma = 55$	55.17	52.87	57.47	52.87	55.17	54.02	65.52	56.32
$\sigma = 60$	85.06	56.32	56.32	62.07	52.87	44.83	57.47	57.47

Table 5.4: *Classification results for detecting the fetal heartbeat after cropping the frames around the localised fetal heart (Step B). The sequence dimensions are $120 \times 120 \times 30$, cropped around the detected fetal heart.*

n	3	4	5	6	7	8	9	10
$\sigma = 30$	81.61	80.46	80.46	83.91	54.02	85.06	81.61	91.95
$\sigma = 35$	87.36	88.51	63.22	87.36	86.21	85.06	47.13	83.91
$\sigma = 40$	49.43	52.87	79.31	78.16	52.87	52.87	50.57	54.02
$\sigma = 45$	89.66	85.06	52.87	52.87	64.37	77.01	50.57	52.87
$\sigma = 50$	93.10	50.57	89.66	70.11	54.02	91.95	68.97	90.80
$\sigma = 55$	78.16	51.72	52.87	51.72	74.71	52.87	62.07	73.56
$\sigma = 60$	63.22	57.47	55.17	64.37	42.53	82.76	50.57	62.07

signed feature symmetry filter (detection accuracy of 88.51%). By comparison when the frames were cropped around the detected fetal heart, a 3-state model and $\sigma = 50$ for the signed feature symmetry filter (detection accuracy of 93.99%) produced the highest results. Generally increasing the number of states up to 7 increases the classification accuracy, however, a further increase leads to a decrease in performance. This can be explained by the fact that when KPCA is used, the main dynamics of the video are best described using the first 6 or 7 eigenvalues, therefore additional eigenvalues capture a very small portion of the variation in the feature space, thus resulting in noisier KDT model parameter estimates.

5.4 Conclusions

In this study we have looked at the problem of locating certain anatomical features in ultrasound and in specific, identification of breech presentation and fetal viability. Breech delivery can significantly increase survival risk for neonates [82]. However, planned vaginal breech deliveries where antenatal ultrasound is available can be associated with a better outcome than reported in randomised trials [74]. Moreover basic obstetric ultrasound skills such as identification of fetal presentation and the pregnancy viability are well recognised skills to be acquired by experienced sonographers. Yet for less experienced or less trained sonographers, identifying anatomical features might be a challenging task. Therefore we believe this framework is a step towards empowering less experienced or trained sonographers to effectively identify structures of interest and interpret the images more confidently.

In addition, we believe the process of ultrasound image acquisition in search of certain anatomical objects could also be simplified through this method. As an example although the ultrasound sweep videos for this study were acquired by a number of experienced obstetrician, a simplified protocol has been followed during the data acquisition to ensure the proposed framework is not only kept simple but also is suited for the application. The data used in this study was obtained from the INTERGROWTH-21st project [211], which contains mothers at different stages of pregnancy. Therefore this data include a variety of representations of the fetal structures with different sizes and shadowing. This provides a set of rich features for the dataset of the positive sequences. However, a larger dataset of videos may be required to build a robust classifier.

One of the main limitations of this approach is that as the scanning protocol is kept simple, fetal skull, abdomen or heart may not be captured or observed as part of the acquired sequence, specially if the fetus is small or at an early gestational age. This is depicted schematically in Figure 5.16. This is generally not a problem in our

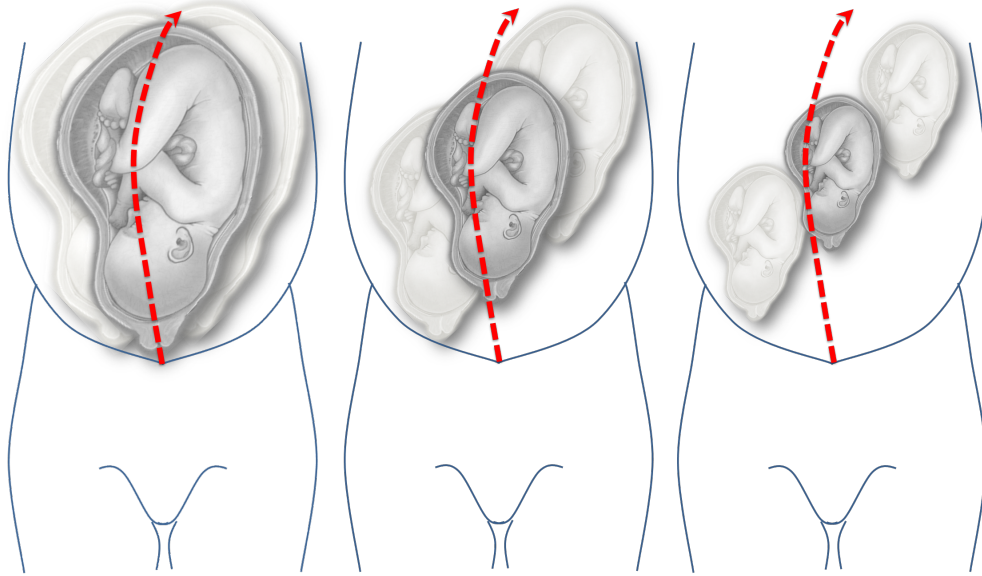


Figure 5.16: Protocol limitations. *As the fetus size gets smaller in proportion of the available space in the womb, the chances of capturing key anatomical structures of interest decreases due to the flexibility of movement of the fetus in the womb.*

target population, at a gestational age of 28 weeks and above, as the fetus is grown to a relatively large size. However, a simple solution to rectify this potential shortcoming is to deploy multiple scans at various positions. This will potentially enhance the results in other cases too, and will indeed be incorporated at the next phase of experiments.

Another potential limitation of this study is that all the fetuses in our database had healthy cardiac activity, thus cases with no cardiac activity were not analysed. This may not be a major issue as the dynamics of a static heart are indeed different to one which is constantly beating. We aim to collect static fetal cardiac data to address this issue in the future. Moreover, sensitivity to scanner settings is one of the major complicating factors in ultrasound images. Therefore, effect of image artefacts as well as variations in ultrasound machine's settings such as frame rate and image persistence need to be further researched. Finally, the designed framework successfully detected fetal presentation and viability as shown by the results. The framework successfully combines simple scanning protocols with image analysis methods and has promising accuracy for real-world deployment.

Chapter 6

Discussion and Future Work

In summary, ultrasound is a safe and relatively low-cost imaging modality that can be effectively used to identify pregnancy complications. Recent technological advances have improved device portability, image quality and cost efficacy, which has triggered hopes for more widespread use of this imaging modality in limited-resource settings. However, navigating the probe and obtaining fetal measurements is a non-trivial task. Moreover, current methods for identifying pregnancy complications and fetal health require expert knowledge for acquisition and interpretation of complex ultrasound patterns. Therefore, for the uptake of this technology in limited-resource countries it is necessary to develop new solutions to empower non-experts to acquire and interpret data.

This thesis has presented a simple yet effective scanning protocol; simple enough to be learned in a short time by non-expert sonographers and information-rich enough to allow automated analysis for detection of the fetal presentation. Moreover, a frame by frame detection framework has been designed to automatically identify the fetal skull, abdomen and heart. Sequences that contain the fetal heart are further analysed to identify whether a fetal heartbeat can be detected. This chapter summarises the main contributions in this thesis and discusses the current limitations and potential opportunities for future improvements.

6.1 Contributions and Limitations

The main contributions of this thesis are as follows;

Framework For Automatic Detection Of Frames Of Interest In Ultrasound

Videos: In Chapter 3, two original methods for automatic detection of frames of interest in an ultrasound video sweep were introduced. Moreover, the merits of pre-processing images on classification accuracy, on two datasets of variable sizes, were investigated. The technical contributions of this chapter were three-fold, namely: (1) to introduce the LP-SIFT (Local Phase based SIFT) features, (2) to illustrate the efficacy of pre-processing ultrasound images before classification and to propose a structured edge representation framework based on Structured Random Forests (SRF) for learning better edge representations, and (3) to propose a classification framework for ultrasound 2D frames based on the improved Fisher vector (FV) encoding of low-level features. The results illustrate that learning edge maps using the SRF framework can improve detection accuracy on the smaller datasets by a margin of approximately 10%.

Automatic Detection Of Sequences Of Interest In An Ultrasound Video,

Based On Sequence Dynamics: Although ultrasound image analysis has been an active area of research in recent years, research into solutions which assist during data acquisition of a 2D scan is a largely unexplored area. In Chapter 4, the general problem of identifying video segments that contain structures of interest based on their motion patterns was addressed. To illustrate clinical applicability, this approach was used to automatically identify the fetal presentation (whether it is breech or not) during the third trimester of pregnancy from a predefined freehand ultrasound sweep. The sweeps were partitioned into smaller sub-sequences. A kernel dynamic texture model was learned for all the video sequences and a similarity metric was defined between the estimated model parameters. The similarity kernel was then used to train

an SVM classifier. The classification accuracy achieved using this framework was 93%.

Clinical Application: Detection Fetal Presentation And Heartbeat From An Ultrasound Video, A Step Towards Automatic Detection Of High-Risk Pregnancies: In **Chapter 5**, a framework for automatic detection of the fetal presentation and heartbeat from a predefined-freehand ultrasound video sweep was proposed. To our knowledge this has not been done before. The clinical value of the proposed framework was shown on a real-world application and a challenging ultrasound video dataset, to identify *high-risk* pregnancies. Our proposed framework achieved an accuracy of approximately 88%.

limitations and assumptions Given the underlying interest of this thesis in developing solutions for limited-resource settings, utilising simplified yet specific and targeted ultrasound scans play a key role in enabling midwives, nurses, and volunteers to undertake certain diagnostic tasks. However, it is important to note that a simplified ultrasound scan inherently contain a limited amount of information and may not be applied to all types of conditions. Moreover, another limitation to have in mind is the possibility of missing information from a simplified scan. Therefore, the analysis of data of this kind would require taking such limitations into account. One way to tackle this limitations is by taking multiple measurements similar to the work of [3] and would have to be studied further.

Another potential limitation is that all the fetuses in our database had healthy cardiac activity, thus cases with no cardiac activity were not analysed. Moreover, the effect of image artefacts as well as variations in ultrasound machine's settings such as frame rate and image persistence may need to be further studied. Additionally, the lack of high-quality ground truth data is one of the main challenges that is yet to be addressed effectively. Furthermore, comparison of the relative performance of a computer algorithm with a human observer, with various levels of training, in the

classification of ultrasound images would be a necessary next step to take. Such a study would clarify whether or not, a human can perform as well as or better than a computer at classification, when minimal training is provided.

6.2 Exploratory Work and Future Directions

We believe there are several exciting extensions to this thesis that could improve both the performance and clinical application of interest. The suggested areas of potential improvement are 1) deploying a low-cost ultrasound probe for data acquisition, 2) learning features from the data through deep convolutional neural networks (CNNs) rather than the popular hand-engineering features, and 3) optimising the code for real-time performance. These suggestions are further discussed here. Furthermore, attempts have been made towards the first two suggested extensions. Our initial findings have been reported in:

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Fetal Head Detection on Images from a Low-Cost Portable USB Ultrasound Device. Conference in Medical Image Understanding and Analysis (MIUA), 17-19 July 2013, Birmingham, UK.

Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Automatic Fetal Head Detection on video clips from a Low-Cost Portable USB Ultrasound (US) Device. Proceedings of the 23rd World Congress on Ultrasound in Obstetrics and Gynecology - ISUOG 2013, 6-9 October 2013. (Clinical Abstract)

Gao Y, **Maraci MA**, Noble JA. Describing Ultrasound Video Content Using Deep Convolutional Neural Networks. International Symposium on Biomedical Imaging (ISBI), 2016, *Accepted paper*.

6.2.1 Using a Portable Probe

As discussed in Section 2, the use of ultrasound imaging is far less common in limited-resource settings. Emerging advances in this field have paved way for smaller ultrasound machines with acceptable image qualities. This has subsequently led to the development of portable ultrasound devices to plug into laptops and smart phones, facilitating use in limited-resource settings. In light of the factors that can affect the quality of ultrasound image analysis and interpretation, we have carried out an initial investigation into whether a combined machine learning and data acquisition approach to fetal head detection using a low-cost USB probe is equivalent to the same analysis on data from a mid-range probe. Initial results show that the algorithm works well on images obtained from both devices and that statistically there is no significant difference between the performance of the algorithms on the two sets [127, 128].

It is important to note that currently the low-cost portable probes that can be powered from the USB port of a laptop are simpler, with less sophisticated beam-forming and post-processing. This means the images can look quite different and potentially for some tasks diagnostically inferior. An example of a low-powered portable ultrasound device, and a regular ultrasound probe is shown in Figure 6.1. Furthermore, Figure 6.2 illustrates two examples of a fetal skull, obtained from two participants, using the two different ultrasound probes. As can be seen, images from the two different sources show a different level of detail of the anatomical structures. However, we argue that although the images are different, for certain applications such as detection of the fetal skull, this difference in quality is not a limitation.

Data Acquisition

The 2D fetal ultrasound images used in this study were acquired from subjects participating in a fetal growth project [211]. Data acquisition was carried out using a mid-range ultrasound machine (Philips HD9 with a V7-3 transducer *denoted as A*),

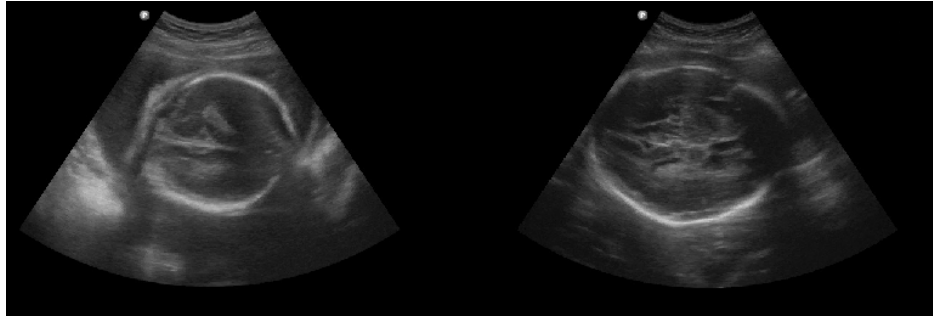
(a) *The Philips transducer.*(b) *Interson Seemore USB probe.*

Figure 6.1: *The two transducers used. The Philips V7-3 transducer is shown on the left and the Interson Seemore USB transducer on the right. Philips image adapted from <http://ultrasound-store.com/products/qz-6073-qzew>*

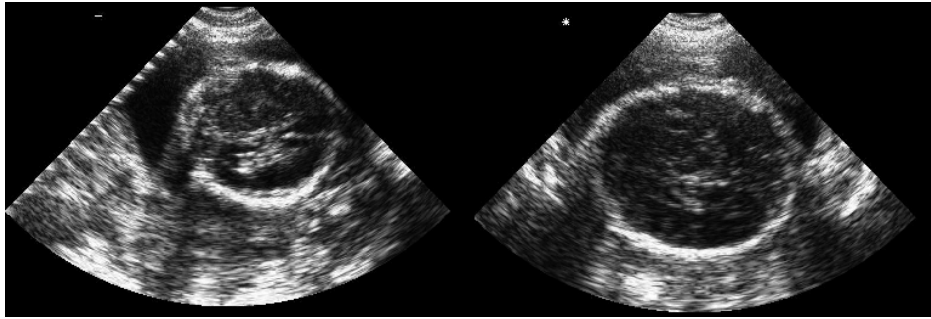
and a low-end portable USB ultrasound machine (Interson Seemore *denoted as B*), by an obstetrician trained to follow standardised procedures (for this refer to Chapter 1, Figure 1.1). The participants were fifteen healthy pregnant women, aged 20 to 38 with the fetus at a gestational age of 16 to 39 weeks. During data acquisition, probe *A* was used twice for each participant resulting in a total of 30 images obtained using *A*. Similarly probe *B* was also used twice for each participant resulting in another 30 images acquired using *B*. Image acquisition was carried out by the same obstetrician and during the same session. The participants were scanned with the two ultrasound probes with the intention to include the same anatomical features while keeping external conditions constant.

Analysis

Automatic head detection and localisation was carried out using the boundary fragment model (BFM) as suggested by Stebbing and McManigle [193] and allows an object to be represented by its scale-normalised edge responses. A pre-trained classifier was used to detect the fetal skull in unseen images. Initially, the scale and centroid of the skull was detected. Fragments from a pretrained edge fragment library were then utilised to weekly describe the shape of the skull. An iterative ellipse fitting algorithm [192] was then used to fit an ellipse on the identified skull edges. The reader



(a) *Images obtained using the Philips ultrasound machine and transducer.*



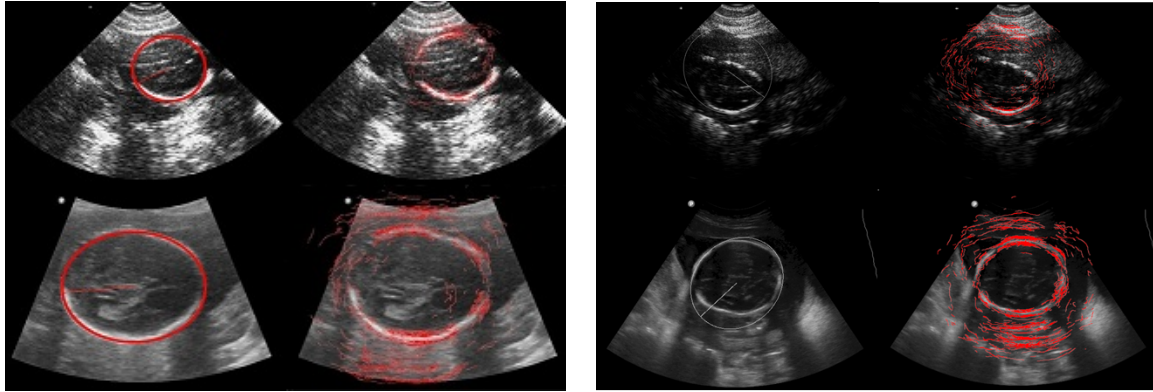
(b) *Images obtained using the Interson Seemore USB transducer.*

Figure 6.2: *Images with different quality. Two sample images are shown obtained using the Philips HD9 and the V7 – 3 transducer and Interson SeeMore probe.*

is referred to [193] and [192] for an extensive explanation and discussion of the two methods.

Validation Methods

The images were graded using a Likert-scale system [132] as previously reported in the literature [16, 185]. The Likert-scale used in this study is set from 1 – 3, where the grades represent *poor*, *fair* and *good* respectively from 1 to 3. The results were graded with the assumption that all the grades have the same weight. Thus images are divided into two classes; class 1 includes images that are rated as *poor* and class 2 includes images rated *fair* or *good*. To analyse the results, the images were assessed prior to any processing. This is to measure 1) sharpness of the images and 2) the visibility of the region of interest (*ROI*) to get a better insight into the accuracy of



(a) *Fetal skull acquired using probe A (bottom) and B (top).*

(b) *Some failure cases where head was not detected correctly.*

Figure 6.3: a) The Interson Seemore and Philips images are shown on the top and bottom rows respectively with the Ellipse fitting results on the left and BFM on the right - The fitted ellipse has been highlighted for clarity. (b) The edge detection has not worked well due to the bright attenuation marks at the top mainly on participants who are at earlier gestations ages where the skull for example is small.

the model, given the initial difference in appearance between the images obtained from the two probes. This assessment was carried out by the author, and with visual checks by an obstetrician. A two tailed Wilcoxon signed-rank test with a 95% confidence interval was used to indicate if there was a statistically significant difference between the result of the head detection algorithm on images acquired using *A* and *B*.

Image Quality Analysis

The Likert-scale grading show that from the 30 images that are acquired from *A* in this experiment, 6 images are graded as poor in terms of visibility and 2 are graded as poor in terms of sharpness. In comparison, the results obtained from *B* suggest there is a clear difference in the appearance and visibility of the region of interest (ROI) as 19 images are rated as poor and the rest are rated fair.

Head Detection Analysis

The results show that 7 images acquired via *A* are in class 1 and the other 23 in class 2 suggesting an overall success of 76.7 % in identifying the head boundaries. The results

Table 6.1: *The results of visibility & sharpness of the ROI in images obtained with the A and B probes, before and after head detection.*

	Good No%	Fair No%	Poor No%	p Value Wilcoxon Rank Test
Visibility ^a	40.0	20.0	20.0	$6.69 * 10^{-4}$
Visibility ^b	0.0	36.7	63.3	
Sharpness ^a	46.7	46.7	6.6	$1.23 * 10^{-4}$
Sharpness ^b	0.0	43.3	56.7	
Head Detection ^a	46.67	30.0	23.33	0.2012
Head Detection ^b	73.33	6.67	20.0	
<i>a.</i> Philips Probe				
<i>b.</i> SeeMore Probe				

for B are also very high with 6 images in class 1 and 24 images as class 2, indicating an accuracy of 80%. Figure 6.3a illustrates the result of the BFM and Ellipse fitting algorithm on an image obtained from the two probes, A and B. Also an example where the algorithms have not worked so well can be seen in Figure 6.3b.

The test was carried out with a two tailed 95% confidence interval. The Wilcoxon Signed-Rank test results from Table 6.1 show that statistically no significant difference between the accuracy of the results were found ($p=0.2012$). However there is a significant difference in the sharpness ($p=0.00067$) and visibility ($p=0.00012$) before the images are processed, as expected. This suggests that although the visibility and sharpness of the images from a low-cost probe may not be as high as a mid-range probe, the performance of the head detection algorithm is not affected by the source of the images.

Initial Conclusions

The statistical analysis shows that the fetal head detection and segmentation algorithms work well with the images obtained from the low-cost USB probe. Given the initial results, it would be interesting to perform such analysis on a larger dataset and for potentially different applications. Currently the main limitation is that data acquisition using the low-cost USB probe requires non-conventional arrangements and

can be challenging due to logistical reasons. These include the USB probe's short cable, and the requirement for it to be connected to a separate laptop to facilitate data acquisition.

6.2.2 Deep Learning Approach

Recently deep learning frameworks have shown outstanding state-of-art accuracy in many applications. The great success on large scale image classification [104, 134, 188] has been partly due to very large collections of labelled natural images. Moreover, recent medical imaging studies have shown promising results in using deep convolutional neural network (CNN) and transfer learning, for example in automated fetal abdominal plane finding [36]. CNNs are highly complex and non-linear models, where the depth and breadth of the network directly affects their feature representation power. As part of an exploratory experiment and in a collaboration with Gao et al. [71], a CNN was trained for the classification of four different classes of ultrasound images using a transfer learning approach. My main contribution to this work was providing and explaining the dataset, describing the significance of the clinical application, and discussions on the architecture of the CNN.

Figure 6.4 illustrates the overall architecture of the CNN used in the work. The design adopts relatively large kernels and stride in the first two convolutional layers. This dramatically reduces the spatial resolution of the data and can improve the computational efficiency of the process. However, it has to be noted that as a consequence of this design decision, the lower level features may become less detailed thus affecting the overall classification accuracy [134].

Transfer Learning

Although CNNs have the advantage of learning powerful feature representations, the fully-supervised deep architectures could substantially over-fit due to the limited avail-

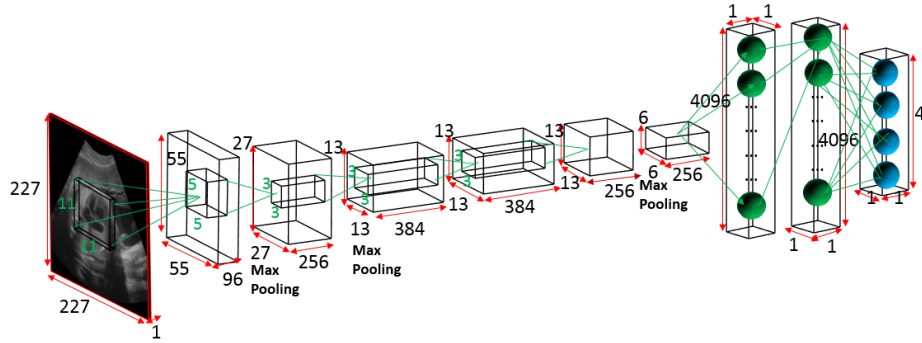


Figure 6.4: *A schematic of the CNN used in this thesis. There are 4096 neurons in each of the fully connected layers. There are 4 neurons in the output layer. Image from [70]*

able training data. A number of studies [53, 172, 220] have demonstrated that transfer learning can be a powerful tool to reduce over-fitting. This involves training a base network on a separate (and large) dataset, and then transferring the learned features of the base network to a new target network to be trained on a target dataset. But do features extracted from a CNN, trained on natural images, generalise well to ultrasound data?

To answer this question a transfer learning approach was used as schematically illustrated in Figure 6.5. The base CNN (denoted by A-CNN) was pre-trained on the 2012 ImageNet database (1.2 million high-resolution images with 1000 different classes of natural images). The learned parameters of the first n layers ($n = 1, 2, \dots, 7$) of the base CNN were then transferred to our target CNN, which is denoted as B-CNN. The remaining layers of B-CNN were randomly initialised from a Gaussian with zero mean and standard deviation of 0.01. To evaluate this approach, the parameters of B-CNN were frozen and also fine-tuned using the ultrasound dataset.

Figure 6.6 illustrates the performance of the standard model. It shows that heart is the most frequently miss-classified class in the test set (approximately 59%). In contrast, the fetal skull achieved the best performance with a classification accuracy of approximately 92%. Abdomen and Other structures achieved an accuracy of around 82% and 90% respectively. It is worth noting that the overall classification performance is improved by approximately 8% after applying a 50% drop-out to the network and

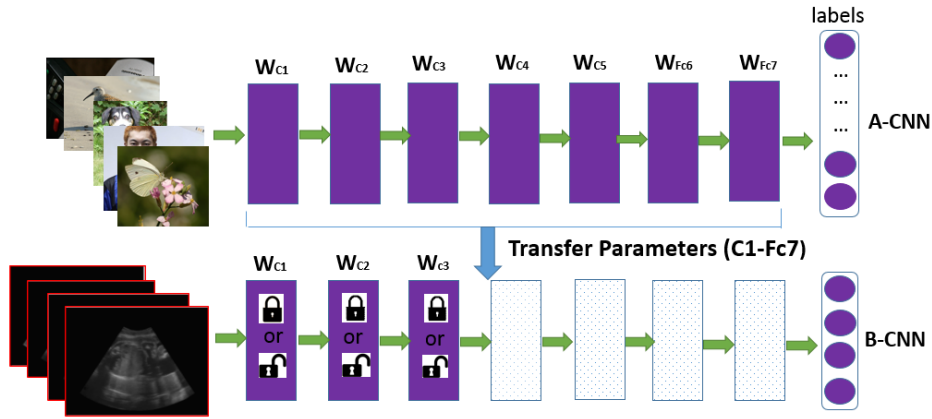


Figure 6.5: Transfer learning architecture. Top row: The CNN-A is pre-trained on the ImageNet database for classification. The labelled purple rectangles (e.g. W_{C1}) represent the weights learned by that layer. Bottom row: In the CNN-B, the first n weight layers (in this example, $n = 3$) are copied from the base network A. The remaining layers are randomly initialised, and then the entire network is trained on ultrasound datasets with the transferred parameters locked (frozen) or unlocked (fine-tuned). Image from [70]

augmenting the training data. Particularly, fetal heart detection is improved by approximately 10%, compared to the standard model.

As can be seen from Figure 6.7, the standard model suffers from substantial overfitting, where the low level convolutional kernels seem to have learned the low-level noise in the data. However, applying drop-out and augmenting the training data helps to reduce overfitting, where clear patterns can be seen in the low-level features. The improvements are partly due to better spatial feature as a result of augmentation. Moreover, utilising drop-out during the learning phase can prevent complex co-adaptations of neurons and force the network to learn robust features that are useful in conjunction with many different random subsets of other neurons.

Initial Conclusions

The results of transfer learning experiments are summarised in Figure 6.8. It can be seen that transferring the lower layers of the target network without fine-tuning, equates to an overall drop of performance (about 8% across layers 1 to 7). Moreover, the high performance of the first two layers indicates that the network has learned

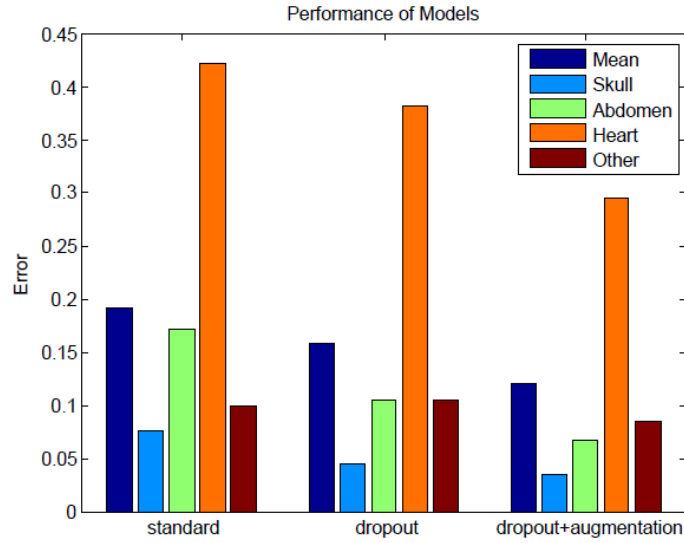


Figure 6.6: *Classification performance of models directly trained on ultrasound images. Image from [70]*

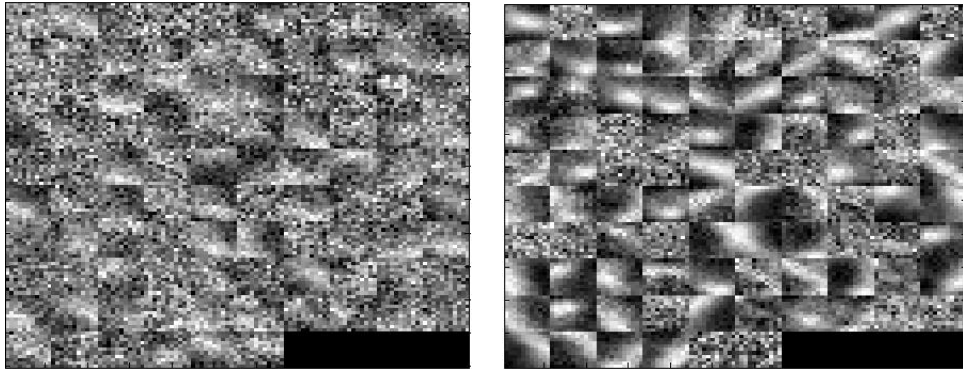


Figure 6.7: *Convolutional kernels learned by the first convolution layer: (left) standard model, (right) added dropout and augmentation. Image from [70]*

general features. Layers 3 to 5 illustrate a slight drop in performance, and layers 5 to 7 show a more significant drop, across all classes. This potentially signifies that features learned in the upper layers of the network may have become too specific. However, transferring features and additional fine-tuning of the network, can provide a more generalised model.

The initial results show that classification accuracy can be enhanced in a CNN by initialising a network with features learned from natural images. However, a major challenge in our dataset are the cardiac frames, partly due to low inter-class variation

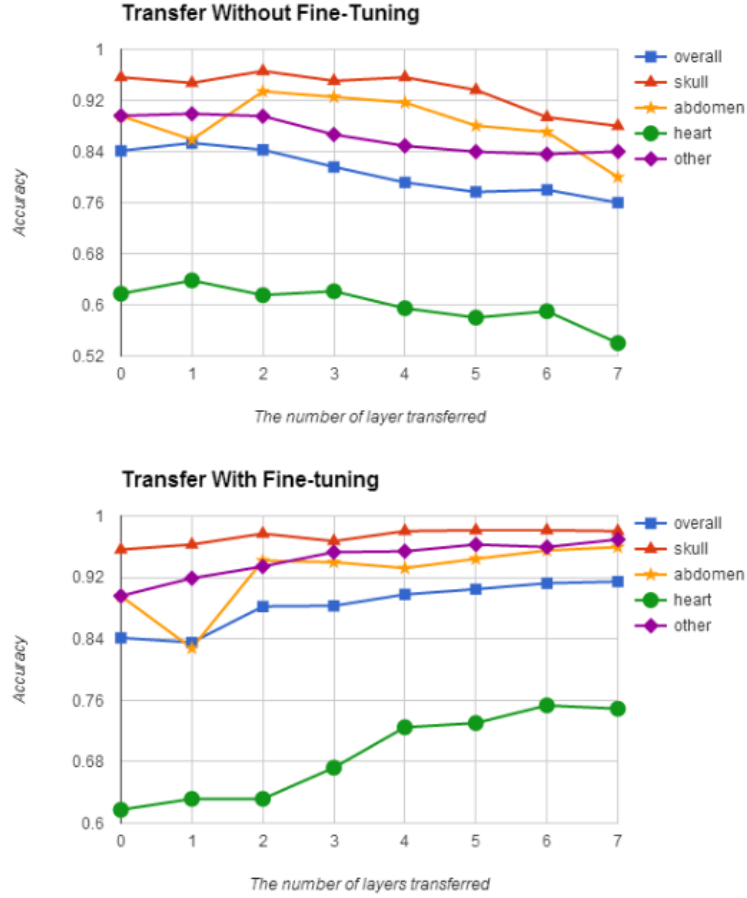


Figure 6.8: *Classification performance of transfer learning models. Image from [70]*

between heart and abdomen, and high intra-class differences in terms of appearance. This can be partly addressed with more training examples. Moreover, investigating motion cues in consecutive heart frames can also be a great advantage and an interesting area to be further investigated.

6.2.3 Real-Time Performance

This work has shown great potential for deployment in low-income settings. An additional feature which would facilitate the uptake of this technology is to ensure feedback can be provided to the sonographers in real-time. This can enhance their scanning experience and also facilitate the use of this technology for non-experts both as a

diagnostic device as well as a training device.

6.2.4 Summary

The work presented in this thesis represents an inter-disciplinary effort towards combining machine learning algorithms and ultrasound imaging to develop a framework with potential for clinical use in limited-resource settings (**Chapter 5**). Specifically, we have proposed a simple ultrasound scanning protocol, in which an ultrasound video is acquired through a linear sweep of the maternal abdomen. The framework was designed to automatically infer the fetal presentation by detecting the fetal skull and abdomen from the sweep. Fetal viability was then assessed in instances where a suitable consecutive sequence of frames of the fetal heart were detected through the method discussed in **Chapter 4**. To our knowledge this has not been done before. Moreover, we have identified three exiting possible extensions to this thesis, where we have illustrated some preliminarily promising results **Chapter 6**.

Bibliography

- [1] Sonographic examination of the fetal central nervous system: guidelines for performing the basic examination and the fetal neurosonogram. *Ultrasound in Obstetrics and Gynecology*, 29(1):109–116, 2007.
- [2] Alfred Abuhamad, editor. *Ultrasound in Obstetrics and Gynecology: A Practical Approach*. GLOWM, 2014.
- [3] Alfred Abuhamad, Yili Zhao, Sharon Abuhamad, Elena Sinkovskaya, Rashmi Rao, Camille Kanaan, and Lawrence Platt. Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination. *American journal of perinatology*, 2015.
- [4] Juan Luis Alcazar, Linder Diaz, Paula Florez, Stefano Guerriero, and Matias Jurado. Intensive training program for ultrasound diagnosis of adnexal masses: protocol and preliminary results. *Ultrasound in Obstetrics & Gynecology*, 42:218–223, Feb 2013.
- [5] Ruth Anderson, Beth Kolko, Laura Schlenke, Waylon Brunette, Alexis Hope, Rob Nathan, Wayne Gerard, Jacqueline Keh, and Michael Kawooya. The midwife’s assistant: designing integrated learning tools to scaffold ultrasound practice. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD ’12, pages 200–210, New York, NY, USA, 2012. ACM.
- [6] Jérémie Anquez, Elsa D Angelini, and Isabelle Bloch. Segmentation of fetal 3d ultrasound based on statistical prior and deformable model. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 17–20, May 2008.
- [7] L Ap Cynydd, Nigel W John, Franck P Vidal, Derek A Gould, Elizabeth Joeques, and Peter Littler. Cost effective ultrasound imaging training mentor for use in developing countries. *Stud Health Technol Inform*, 142:49–54, 2009.
- [8] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [9] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

- [10] David P. Bahner and Nelson A. Royall. Advanced ultrasound training for fourth-year medical students: a novel training program at the ohio state university college of medicine. *Academic Medicine*, 88(2):206–213, Feb 2013.
- [11] Max L Baker and Glenn V Dalrymple. Biological effects of diagnostic ultrasound: a review. *Radiology*, 126(2):479–483, 1978.
- [12] John L Barron, David J Fleet, and SS Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.
- [13] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.
- [14] Alessandro Bissacco, Alessandro Chiuso, and Stefano Soatto. Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(11):1958–1972, 2007.
- [15] Michael J. Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75 – 104, 1996.
- [16] Michael Blaivas, Larry Brannam, and Daniel Theodoro. Ultrasound image quality comparison between an inexpensive handheld emergency department (ed) ultrasound machine and a large mobile ed ultrasound system. *Academic Emergency Medicine*, 11(7):778–781, Jul 2004.
- [17] Michael Blaivas, Walter Kuhn, Brad Reynolds, and Larry Brannam. Change in differential diagnosis and patient management with the use of portable ultrasound in a remote setting. *Wilderness & environmental medicine*, 16(1):38–41, 2005.
- [18] Tobias Blum, Andreas Rieger, Nassir Navab, Helmut Friess, and Marc Martignoni. A review of computer-based simulators for ultrasound training. *Simulation in Healthcare*, 8(2):98–108, Apr 2013.
- [19] Stephen Bly, Michiel C. Van den Hof, Society of Obstetricians , Diagnostic Imaging Committee, and Gynaecologists of Canada. Obstetric ultrasound biological effects and safety. *Journal of Obstetrics and Gynecology Candada*, 27(6):572–580, Jun 2005.
- [20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [21] Christopher P Bridge and J Alison Noble. Object localisation in fetal ultrasound images using invariant features. In *Biomedical Imaging, International Symposium on*, April 2015.

- [22] Hannah E. Brindle, Theresa J. Allain, Sam Kampondeni, Noel Kayange, Brian Faragher, Imelda Bates, and Elizabeth Joeques. Utilization of ultrasound in medical inpatients in malawi. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 107(7):405–410, Jul 2013.
- [23] Phillip G. Brown, Joseph Alsousou, Ashley Cooper, Mark S. Thompson, and J. Alison Noble. The autoqual ultrasound elastography method for quantitative assessment of lateral strain in post-rupture achilles tendons. *Journal of Biomechanics*, 46(15):2695 – 2700, 2013.
- [24] Andres Bruhn, Joachim Weickert, and Christoph Schnorr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3):211–231, 2005.
- [25] Waylon Brunette, Wayne Gerard, Matthew A. Hicks, Alexis Hope, Mitchell Ishimitsu, Pratik Prasad, Ruth E. Anderson, Gaetano Borriello, Beth E. Kolko, and Robert Nathan. Portable antenatal ultrasound platform for village midwives. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 23:1–23:10, New York, NY, USA, 2010. ACM.
- [26] R Byaruhanga, D G Bassani, A Jagau, P Muwanguzi, A L Montgomery, and J E Lawn. Use of wind-up fetal doppler versus pinard for fetal heart rate intermittent monitoring in labour: a randomised clinical trial. *BMJ Open*, 5(1), 2015.
- [27] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, 1986.
- [28] Gustavo Carneiro, Fernando Amat, Bogdan Georgescu, Sara Good, and Dorin Comaniciu. Semantic-based indexing of fetal anatomies from 3-d ultrasound data using global/semi-local context and sequential sampling. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.
- [29] Gustavo Carneiro, Bogdan Georgescu, Sara Good, and Dorin Comaniciu. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree. *IEEE Transactions on Medical Imaging*, 27(9):1342–1355, Sep 2008.
- [30] Vikram Chalana, Thomas C. Winter III, Dale R. Cyr, David R. Haynor, and Yongmin Kim. Automatic fetal head measurements from sonographic images. *Academic Radiology*, 3(8):628–635, Aug 1996.
- [31] Antoni B Chan and Nuno Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 846–851. IEEE, 2005.

- [32] Antoni B Chan and Nuno Vasconcelos. Classifying video with kernel dynamic textures. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.
- [33] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926, 2008.
- [34] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and Rene Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE, 2009.
- [35] Rizwan Chaudhry and Rene Vidal. Recognition of visual dynamical processes: Theory, kernels and experimental evaluation. *Department of Computer Science, John Hopkins University, Technical Report*, -:09–01, 2009.
- [36] Hao Chen, Dong Ni, Jing Qin, Shenli Li, Xin Yang, Tianfu Wang, and Pheng Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *Biomedical and Health Informatics, IEEE Journal of*, 19(5):1627–1636, Sept 2015.
- [37] Mary Cheng and Mary Hannah. Breech delivery at term: a critical review of the literature. *Obstetrics & Gynecology*, 82(4 Pt 1):605–618, Oct 1993.
- [38] Catherine Cluver, G Justus Hofmeyr, Gillian Ml Gyte, and Marlene Sinclair. Interventions for helping to turn term breech babies to head first presentation when using external cephalic version. *Cochrane Database Syst Rev*, 1:CD000184, 2012.
- [39] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [40] Emanuele Coviello, Adeel Mumtaz, Antoni B Chan, and Gert RG Lanckriet. Growing a bag of systems tree for fast and accurate classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1979–1986, 2012.
- [41] Marveen Craig. Sonographer shortages: A day late and a dollar short? *Journal of Diagnostic Medical Sonography*, 19(4):261–271, 2003.
- [42] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

- [43] Rémi Cuingnet, Oudom Somphone, Benoit Mory, Raoul Prevost, Mohammad Yaqub, Raffaele Napolitano, Aris Papageorghiou, David Roundhill, J Alison Noble, and Roberto Ardon. Where is my baby? a fast fetal head auto-alignment in 3d-ultrasound. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 768–771, 2013.
- [44] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, 2005.
- [45] Robert Dalla Pozza, Markus Loeff, Rainer Kozlik-Feldmann, and Heinrich Netz. Hand-carried ultrasound devices in pediatric cardiology: clinical experience with three different devices in 110 patients. *Journal of the American Society of Echocardiography*, 23(12):1231–1237, Dec 2010.
- [46] Katrien De Cock and Bart De Moor. Subspace angles and distances between arma models. In *Proc. of the Intl. Symp. of Math. Theory of networks and systems*, volume 1. Citeseer, 2000.
- [47] Katrien De Cock and Bart De Moor. Subspace angles between arma models. *Systems & control letters*, 46(4):265–270, 2002.
- [48] Oya Demirci, Ahmet Semih Tugrul, Abdulkadir Turgut, Sirin Ceylan, and Sadiye Eren. Pregnancy outcomes by mode of delivery among breech births. *Archives of gynecology and obstetrics*, 285(2):297–303, Feb 2012.
- [49] Meghna Desai, Penelope A. Phillips-Howard, Frank O. Odhiambo, Abraham Katana, Peter Ouma, Mary J. Hamel, Jackton Omoto, Sheila Macharia, Annemieke van Eijk, Sheila Ogwang, Laurence Slutsker, and Kayla F. Laserson. An analysis of pregnancy-related mortality in the kemri/cdc health and demographic surveillance system in western kenya. *PLoS One*, 8(7):e68733, 2013.
- [50] Kristen K DeStigter, G Eli Morey, Brian S Garra, Matthew R Rielly, Martin E Anderson, Michael G Kawooya, Alphonsus Matovu, and Frank R Miele. Low-cost teleradiology for rural ultrasound. In *Global Humanitarian Technology Conference (GHTC), 2011 IEEE*, pages 290–295. IEEE, 2011.
- [51] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1841–1848. IEEE, 2013.
- [52] JoãoS. Domingos, RichardV. Stebbing, Paul Leeson, and J.Alison Noble. Structured random forests for myocardium delineation in 3d echocardiography. In Guorong Wu, Daoqiang Zhang, and Luping Zhou, editors, *Machine Learning in Medical Imaging*, volume 8679 of *Lecture Notes in Computer Science*, pages 215–222. Springer International Publishing, 2014.

- [53] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.
- [54] Gianfranco Doretto, Alessandro Chiuso, Ying Nian Wu, and Stefano Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.
- [55] Gianfranco Doretto, Daniel Cremers, Paolo Favaro, and Stefano Soatto. Dynamic texture segmentation. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1236–1242. IEEE, 2003.
- [56] Francis A. Duck. Is it safe to use diagnostic ultrasound during the first trimester? *Ultrasound in Obstetrics and Gynecology*, 13(6):385–388, 1999.
- [57] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (Pt.1)*. Wiley-Interscience, 2 edition, November 2000.
- [58] Sabaratnam Arulkumaran E. Malcolm Symonds, Ian M. Symonds, editor. *Essential Obstetrics and Gynaecology*. Churchill Livingstone, 2013.
- [59] Issam El-Naqa, Yongyi Yang, Miles N Wernick, Nikolas P Galatsanos, and Robert M Nishikawa. A support vector machine approach for detection of microcalcifications. *Medical Imaging, IEEE Transactions on*, 21(12):1552–1563, 2002.
- [60] Arthur T Evans. *Manual of obstetrics*. Lippincott Williams & Wilkins, 2007.
- [61] Sultan S Farjad, G Iohom, and G Shorten. Effect of feedback content on novices learning ultrasound guided interventional procedures. *Minerva Anesthesiol*, Epub ahead of print:Epub ahead of print, Jun 2013.
- [62] M. Felsberg and G. Sommer. The monogenic signal. *Signal Processing, IEEE Transactions on*, 49(12):3136–3144, 2001.
- [63] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.
- [64] A.W. Fitzgibbon. Stochastic rigidity: image registration for nowhere-static scenes. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 662–669 vol.1, 2001.
- [65] National Collaborating Centre for Womens and Childrens Health (UK). *Antenatal Care: Routine Care for the Healthy Pregnant Woman.*, volume No. 62. RCOG Press London, 2008.
- [66] G David Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

- [67] Judith A Fortney, Inne Susanti, Saad Gadalla, Saneya Saleh, Susan M Rogers, and Malcolm Potts. Reproductive mortality in two developing countries. *American journal of public health*, 76(2):134–138, Feb 1986.
- [68] Roger K Freeman, Thomas J Garite, Michael P Nageotte, and Lisa A Miller. *Fetal heart rate monitoring*. Lippincott Williams & Wilkins, 2012.
- [69] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
- [70] Y Gao, Mohammad Ali Maraci, and J Alison Noble. Describing ultrasound video content using deep convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 787–790. IEEE, 2016.
- [71] Yuan Gao. Describing ultrasound video content using deep learning. Technical report, University of Oxford, 2015. Internal report.
- [72] Darius M Gavrilă. Pedestrian detection from a moving vehicle. In David Vernon, editor, *Computer Vision - ECCV 2000*, volume 1843 of *Lecture Notes in Computer Science*, pages 37–49. Springer Berlin Heidelberg, 2000.
- [73] Syed Nasir Ghani, James N. Kirkpatrick, Kirk T. Spencer, Gwendolyn L. Smith, Martin C. Burke, Susan S. Kim, Aseem D. Desai, and Bradley P. Knight. Rapid assessment of left ventricular systolic function in a pacemaker clinic using a hand-carried ultrasound device. *Journal of Interventional Cardiac Electrophysiology*, 16(1):39–43, Jun 2006.
- [74] François Goffinet, Marion Carayol, Jean-Michel Foidart, Sophie Alexander, Serge Uzan, Damien Subtil, and Gérard Bréart. Is planned vaginal delivery for breech presentation at term still an option? results of an observational prospective survey in france and belgium. *American Journal of Obstetrics and Gynecology*, 194(4):1002 – 1011, 2006.
- [75] Steven R Goldstein. Embryonic ultrasonographic measurements: crown-rump length revisited. *American journal of obstetrics and gynecology*, 165(3):497–501, 1991.
- [76] Vicente Grau, Harald Becher, and J Alison Noble. Phase-based registration of multi-view real-time three-dimensional echocardiographic sequences. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pages 612–619. Springer, 2006.
- [77] Vicente Grau, Harald Becher, and J Alison Noble. Registration of multiview real-time 3-d echocardiographic sequences. *Medical Imaging, IEEE Transactions on*, 26(9):1154–1165, 2007.

- [78] H Greenspan, S Belongie, R Goodman, P Perona, S Rakshit, and CH Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 222–228. IEEE, 1994.
- [79] Rebecca Hall, Carolyn Coffin, Dale Cyr, Wayne Persutte, Doug Roberts, Jean Lea Spitz, Alan Waggoner, Ultrasound Practitioner Commission, et al. The ultrasound practitioner: A proposal response to the sdms for the development of a middle care provider in ultrasound imaging. *Journal of Diagnostic Medical Sonography*, 15(4):140–156, 1999.
- [80] Mike Halliwell. A tutorial on ultrasonic physics and imaging techniques. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 224(2):127–142, 2010.
- [81] Christine W Hanna and Abou Bakr M Youssef. Automated measurements in obstetric ultrasound images. In *Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 3 - Volume 3*, ICIP '97, pages 714–, Washington, DC, USA, 1997. IEEE Computer Society.
- [82] Mary E Hannah, Walter J Hannah, Sheila A Hewson, Ellen D Hodnett, Saroj Saigal, Andrew R Willan, and Term Breech Trial Collaborative. Planned caesarean section versus planned vaginal birth for breech presentation at term: a randomised multicentre trial. term breech trial collaborative group. *Lancet*, 356(9239):1375–1383, Oct 2000.
- [83] R.M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6):610–621, 1973.
- [84] Benjamin Bejar Haro, Luca Zappella, and Rene Vidal. Surgical gesture classification from video data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 34–41. Springer, 2012.
- [85] Hauke S Heinzow, Hendrik Friederichs, Philipp Lenz, Andre Schmedt, Jan C Becker, Karin Hengst, Bernhard Marschall, and Dirk Domagk. Teaching ultrasound in a curricular course according to certified efsumb standards during undergraduate medical education: a prospective study. *BMC Medical Education*, 13(1):84, 2013.
- [86] Reli HersHKovitz, Eyal Sheiner, and Moshe Mazor. Ultrasound in obstetrics: a review of safety. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 101(1):15 – 18, 2002.
- [87] Durlin E Hickok, Diane C Gordon, John A Milberg, Michelle A Williams, and Janet R Daling. The frequency of breech presentation by gestational age at birth: a large population-based study. *American journal of obstetrics and gynecology*, 166(3):851–852, Mar 1992.

- [88] T. Hoehn and M. Hoppenz. Neonatal and childhood mortality rates in myanmar. *Klin Padiatr*, 221(4):266–268, 2009.
- [89] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1):185–203, 1981.
- [90] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.
- [91] Sandra MGVB Jardim and Mário AT Figueiredo. Segmentation of fetal ultrasound images. *Ultrasound in Medicine & Biology*, 31(2):243 – 250, 2005.
- [92] Sandra VB Jardim and Mário AT Figueiredo. Automatic contour estimation in fetal ultrasound images. In *Image Processing, 2003. ICIIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–1065–8 vol.3, 2003.
- [93] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006.
- [94] Michael J Kadour, Rosie Adams, Ruth English, Vaishali Parulekar, Susan Christopher, and J Alison Noble. Slip imaging: reducing ambiguity in breast lesion assessment. *Ultrasound in medicine & biology*, 36(12):2027–2035, 2010.
- [95] Michael J Kadour and J Alison Noble. Assisted-freehand ultrasound elasticity imaging. *Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on*, 56(1):36–43, 2009.
- [96] Andrea Vedaldi Ken Chatfield, Victor Lempitsky and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference*, pages 76.1–76.12. BMVA Press, 2011. <http://dx.doi.org/10.5244/C.25.76>.
- [97] Heidi Harbison Kimberly, Alice Murray, Maria Mennicke, Andrew Liteplo, Jason Lew, J Stephen Bohan, Lynda Tyer-Viola, Roy Ahn, Thomas Burke, and Vicki E Noble. Focused maternal ultrasound by midwives in rural zambia. *Ultrasound in medicine & biology*, 36(8):1267–1272, 2010.
- [98] Bruce J. Kimura and Anthony N. DeMaria. Technology insight: hand-carried ultrasound cardiac assessment—evolution, not revolution. *Nature Clinical Practice Cardiovascular Medicine*, 2(4):217–23; quiz 224, Apr 2005.
- [99] Sergio L Kobal, Lawrence SC Czer, Peter C Czer, Zhanna Feldsher, Robert Hamilton, and Robert J Siegel. Making an impossible mission possible. *CHEST Journal*, 125(1):293–296, 2004.

- [100] Sergio L. Kobal, Luca Trento, Simin Baharami, Kirsten Tolstrup, Tasneem Z. Naqvi, Bojan Cercek, Yoram Neuman, James Mirocha, Saibal Kar, James S. Forrester, and Robert J. Siegel. Comparison of effectiveness of hand-carried ultrasound to bedside cardiovascular physical examination. *American Journal of Cardiology*, 96(7):1002–1006, Oct 2005.
- [101] P. Kotschieder, S. Rota Bulo, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2190–2197, Nov 2011.
- [102] Simon Kotlyar and Christopher L. Moore. Assessing the utility of ultrasound in liberia. *Journal of Emergencies, Trauma and Shock*, 1(1):10–14, Jan 2008.
- [103] Peter Kovesi. Symmetry and asymmetry from local phase. In *Tenth Australian joint conference on artificial intelligence*, pages 185–190. Citeseer, 1997.
- [104] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [105] AM Kumar and KS Shriram. Automated scoring of fetal abdomen ultrasound scan-planes for biometry. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 862–865. IEEE, 2015.
- [106] Mehmet Kuntalp and Orkun Akar. A simple and low-cost internet-based tele-consultation system that could effectively solve the health care access problems in underserved areas of developing countries. *Computer Methods and Programs in Biomedicine*, 75(2):117 – 126, 2004.
- [107] Roland Kwitt, Nuno Vasconcelos, Sharif Razzaque, and S Aylward. Localizing target structures in ultrasound video - a phantom study. *Medical Image Analysis*, 17(7):712 – 722, 2013. Special Issue on the 2012 Conference on Medical Image Computing and Computer Assisted Intervention.
- [108] Roland Kwitt, Nuno Vasconcelos, Sharif Razzaque, and Stephen Aylward. Recognition in ultrasound videos: where am i? In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 83–90. Springer, 2012.
- [109] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001.
- [110] T Lander et al. *Neonatal and perinatal mortality: country, regional and global estimates*. World Health Organization, 2006.
- [111] Kenneth Ivan Laws. Textured image segmentation. Technical report, DTIC Document, 1980.

- [112] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [113] Wesley Lee, Andrea N. Hodges, Sterling Williams, Ivana M. Vettraino, and Beverley McNie. Fetal ultrasound training for obstetrics and gynecology residents. *Obstetrics and Gynecology*, 103(2):333–338, Feb 2004.
- [114] Victor Lempitsky, Michael Verhoek, J.Alison Noble, and Andrew Blake. Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography. In Nicholas Ayache, Herve Delingette, and Maxime Sermesant, editors, *Functional Imaging and Modeling of the Heart*, volume 5528 of *Lecture Notes in Computer Science*, pages 447–456. Springer Berlin Heidelberg, 2009.
- [115] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Generalized boundaries from multiple image interpretations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1312–1324, July 2014.
- [116] David Liang and Ingela Schnittger. Accuracy of hand-carried ultrasound. *Echocardiography*, 20(5):487–490, Jul 2003.
- [117] Che-Bing Liu, Ruei-sung Lin, and Narendra Ahuja. Modeling dynamic textures using subspace mixtures. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1378–1381. IEEE, 2005.
- [118] K. Liu, H. Skibbe, et al. Rotation-invariant HOG descriptors using Fourier analysis in polar and spherical coordinates. *IJCV*, 106(3):342–364, 2014.
- [119] Abhay Lodha, Qiaohao Zhu, Shoo K. Lee, Prakesh S. Shah, and "Canadian Neonatal Network". Neonatal outcomes of preterm infants in breech presentation according to mode of birth in canadian nicus. *Postgraduate medical journal*, 87(1025):175–179, Mar 2011.
- [120] David G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2, 1999.
- [121] Wei Lu, Jinglu Tan, and Randall Floyd. Automated fetal head detection and measurement in ultrasound images by iterative randomized hough transform. *Ultrasound in medicine and biology*, 31(7):929–936, Jul 2005.
- [122] Brian P. Lucas, Carolina Candotti, Bosko Margeta, Benjamin Mba, Rudolf Kumapley, Abdo Asmar, Ricardo Franco-Sadud, Joshua Baru, Christine Acob, Shane Borkowsky, and Arthur T. Evans. Hand-carried echocardiography by hospitalists: a randomized trial. *The American journal of medicine*, 124(8):766–774, Aug 2011.

- [123] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.
- [124] H. Ludwig. Christian gerhard leopold (1846–1911). *Der Gynakologe*, 37(10):961–966, 2004.
- [125] M Mahran and M Omram. The impact of diagnostic ultrasound on the prediction of intrauterine growth retardation in developing countries. *International Journal of Gynecology & Obstetrics*, 26(3):375–378, 1988.
- [126] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693, 1989.
- [127] Mohamad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, and J Alison Noble. Fetal head detection on images from a low-cost portable usb ultrasound device. In *Proc. 17th Conference on Medical Image Understanding and Analysis (MIUA)*, 2013.
- [128] Mohamad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, and J Alison Noble. Op15.02: Automatic fetal head detection on video clips from a low-cost portable usb ultrasound (us) device. *Ultrasound in Obstetrics & Gynecology*, 42(s1):89–89, 2013.
- [129] Mohamad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, and J Alison Noble. Object classification in an ultrasound video using lp-sift features. In *MICCAI Workshop on Machine Learning in Medical Imaging*. Springer, 2014.
- [130] Mohamad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, and J Alison Noble. Searching for structures of interest in an ultrasound video sequence. In *MICCAI Workshop on Medical Computer Vision*. Springer, 2014.
- [131] Mohamad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, and J Alison Noble. Fisher vector encoding for detecting objects of interest in ultrasound videos. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 651–654. IEEE, 2015.
- [132] Gary Michael Maranell. *Scaling: A Sourcebook for Behavioral Scientists*. Transaction Publishers, October 2007.
- [133] Richard J Martin. A metric for arma processes. *Signal Processing, IEEE Transactions on*, 48(4):1164–1170, 2000.
- [134] Xiang Zhang Michael Mathieu. Pierre sermanet david eigen xiang zhang michael mathieu rob fergus yann lecun. *arXiv preprint arXiv:1312.6229*, 2013.
- [135] George K Matsopoulos and S Marshall. Use of morphological image processing techniques for the measurement of a fetal head from ultrasound images. *Pattern Recognition*, 27(10):1317–1324, 1994.

- [136] I. Mbachu, G. O. Udigwe, C. I. Okafor, O. S. Umeonunihu, C. Ezeama, and G. U. Eleje. The pattern and obstetric outcome of hypertensive disorders of pregnancy in nnewi, nigeria. *Nigerian Journal of Medicine*, 22(2):117–122, 2013.
- [137] S Mbuyita, R Tillya, R Godfrey, U Janeen, K Mukherji, et al. Uptake of training on vsan by midlevel providers working in rural health facilities in tanzania: Implications for reliability. *J Biosafety Health Educ*, 2(123):2332–0893, 2014.
- [138] WarrenS. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- [139] Margaret B McNay and John E Fleming. Forty years of obstetric ultrasound 1957-1997: from a-scope to three dimensions. *Ultrasound in Medicine and Biology*, 25(1):3–56, Jan 1999.
- [140] Miguel Mulet-Parada and J.Alison Noble. 2d+t acoustic boundary detection in echocardiography. *Medical Image Analysis*, 4(1):21 – 30, 2000.
- [141] Ana IL Namburete, Bahbibibi Rahmatullah, and J Alison Noble. Nakagami-based choroid plexus detection in fetal ultrasound images using adaboost. In *Proceedings of the Medical Image Understanding and Analysis*, pages 31–36, 2012.
- [142] Ana IL Namburete, Richard V Stebbing, Bryn Kemp, Mohammad Yaqub, Aris T Papageorghiou, and J Alison Noble. Learning-based prediction of gestational age from ultrasound images of the fetal brain. *Medical image analysis*, 21(1):72–86, 2015.
- [143] Ana Ineyda Luisa Namburete. *Machine Learning to Assess the Fetal Brain from Ultrasound Images*. PhD thesis, Department of Engineering Science, University of Oxford, 2015.
- [144] Clay Naomi. Strategies for eliminating the sonographer shortage: Recruitment, retention, and educational perspectives. *Journal of Diagnostic Medical Sonography*, 20(6):408–413, 2004.
- [145] James P Neilson. Fetal electrocardiogram (ecg) for fetal monitoring during labour. *The Cochrane Library*, 2013.
- [146] Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848, 2001.
- [147] Tien Dung Nguyen, Sang Hyun Kim, and Nam Chul Kim. Surface extraction using svm-based texture classification for 3d fetal ultrasound imaging. In *Communications and Electronics, 2006. ICCE '06. First International Conference on*, pages 285–290, Oct 2006.

- [148] Stephane Nicolau, Anant Vemuri, Hurng-Sheng Wu, Min-Ho Huang, Yolin Ho, Arnaud Charnoz, Alexandre Hostettler, Luc Soler, and Jacques Marescaux. A low cost simulator to practice ultrasound image interpretation and probe manipulation: Design and first evaluation. In *VR Innovation (ISVRI), 2011 IEEE International Symposium on*, pages 37–40. IEEE, 2011.
- [149] J Alison Noble and Djamal Boukerroui. Ultrasound image segmentation: a survey. *IEEE Trans Med Imaging*, 25(8):987–1010, Aug 2006.
- [150] J Alison Noble, Nassir Navab, and H Becher. Ultrasonic image analysis and image-guided interventions. *Interface focus*, 1(4):673–685, 2011.
- [151] Nawal M Nour. An introduction to maternal mortality. *Reviews in Obstetrics and Gynecology*, 1(2):77, 2008.
- [152] Sebastian Nowozin and Christoph H Lampert. Structured learning and prediction in computer vision. *Foundations and Trends In Computer Graphics and Vision*, 6(3–4):185–365, 2011.
- [153] American College of Obstetricians, Gynecologists, et al. Acog practice bulletin no. 101: Ultrasonography in pregnancy. *Obstetrics and Gynecology*, 113(2 Pt 1):451, 2009.
- [154] SangMin Oh, JamesM. Rehg, Tucker Balch, and Frank Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *International Journal of Computer Vision*, 77(1-3):103–124, 2008.
- [155] Andreas Opelt, Axel Pinz, and Andrew Zisserman. A boundary-fragment-model for object detection. In *Computer Vision–ECCV 2006*, pages 575–588. Springer, 2006.
- [156] Harald Ostensen. Developing countries. *Ultrasound in Medicine & Biology*, 26, Supplement 1(0):S159 – S161, 2000.
- [157] Aris T Papageorgiou, Eric O Ohuma, Douglas G Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A Jaffer, Enrico Bertino, Michael G Gravett, Manorama Purwar, J Alison Noble, Ruyan Pang, Cesar G Victora, Fernando C Barros, Maria Carvalho, Laurent J Salomon, Zulfiqar A Bhutta, Stephen H Kennedy, and José Villar. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. *The Lancet*, 384(9946):869 – 879, 2014.
- [158] Munna Lal Patel, Rekha Sachan, Radheshyam, and Pushpalata Sachan. Acute renal failure in pregnancy: Tertiary centre experience from north indian population. *Nigerian medical journal*, 54(3):191–195, May 2013.
- [159] Sayan Dev Pathak, Vikram Chalana, and Yongmin Kim. Interactive automatic fetal head measurements from ultrasound images using multimedia computer technology. *Ultrasound in Medicine & Biology*, 23(5):665 – 673, 1997.

- [160] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8. IEEE, 2007.
- [161] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156. Springer Berlin Heidelberg, 2010.
- [162] N Petrenko. Nursing care during pregnancy. childbirth and perinatal education. Online article, accessed on 2015.10.08.
- [163] GV Ponomarev, MS Gelfand, and MD Kazanov. A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images. *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, pages 17–19, 2012.
- [164] Kara-Lee Pool, Brian S Garra, and Dorothy I Bulas. Volume sweep imaging: open-source technology for pediatric global health collaboration. *Pediatric radiology*, 44(6):677–678, 2014.
- [165] Bahbibibi Rahmatullah. *Assessment of Obstetric Ultrasound Images using Machine Learning*. PhD thesis, Department of Engineering Science, University of Oxford, 2012.
- [166] Bahbibibi Rahmatullah, Aris Papageorghiou, and J. Alison Noble. Automated selection of standardized planes from ultrasound volume. In Kenji Suzuki, Fei Wang, Dinggang Shen, and Pingkun Yan, editors, *Machine Learning in Medical Imaging*, volume 7009 of *Lecture Notes in Computer Science*, pages 35–42. Springer Berlin / Heidelberg, 2011. 10.1007/978-3-642-24319-6_5.
- [167] Bahbibibi Rahmatullah, Aris T Papageorghiou, and J Alison Noble. Integration of local and global features for anatomical object detection in ultrasound. In Nicholas Ayache, Hervé Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012*, volume 7512 of *Lecture Notes in Computer Science*, pages 402–409. Springer Berlin Heidelberg, 2012.
- [168] Bahbibibi Rahmatullah, Ippokratis Sarris, Aris Papageorghiou, and J Alison Noble. Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using adaboost. In *Proc. IEEE Int Biomedical Imaging: From Nano to Macro Symp*, pages 6–9, 2011.
- [169] Kashif Rajpoot, Vicente Grau, and J. Alison Noble. Local-phase based 3d boundary detection using monogenic signal and its application to real-time 3-d echocardiography images. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, pages 783–786, June 2009.
- [170] Kashif Rajpoot, Alison Noble, Vicente Grau, and Nasir Mahmood Rajpoot. Feature detection from echocardiography images using local phase information. In *Medical Image Understanding and Analysis (MIUA), Dundee, UK*, 2008.

- [171] Arunkumar Ravichandran, Rizwan Chaudhry, and René Vidal. View-invariant dynamic texture recognition using a bag of dynamical systems. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1651–1657, 2009.
- [172] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.
- [173] E Albert Reece, Efstratios Assimakopoulos, Xue-Zhong Zheng, Zion Hagay, and John C Hobbins. The safety of obstetric ultrasonography: concern for the fetus. *Journal of Obstetric and Gynecology*, 76(1):139–146, Jul 1990.
- [174] Xiaofeng Ren, Charless C. Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 614–627. Springer Berlin Heidelberg, 2006.
- [175] M. J. Rijken, S. J. Lee, M. E. Boel, A. T. Papageorghiou, G. H. A. Visser, S. L. M. Dwell, S. H. Kennedy, P. Singhasivanon, N. J. White, F. Nosten, and R. McGready. Obstetric ultrasound scanning by local health workers in a refugee camp on the thai burmese border. *Ultrasound in Obstetrics and Gynecology*, 34(4):395–403, 2009.
- [176] Andrew B. Ross, Kristen K. DeStigter, Matthew Rielly, Sonia Souza, Gabriel Eli Morey, Melissa Nelson, Eric Z. Silfen, Brian Garra, Alphonsus Matovu, and Michael Grace Kawooya. A low-cost ultrasound program leads to increased antenatal clinic visits and attended deliveries at a health care clinic in rural uganda. *PLoS ONE*, 8(10):e78450, 10 2013.
- [177] Payam Sabzmeydani and Greg Mori. Detecting pedestrians by learning shapelet features. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.
- [178] Payam Saisan, Gianfranco Doretto, Ying Nian Wu, and Stefano Soatto. Dynamic texture recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–58. IEEE, 2001.
- [179] L. J. Salomon, Z. Alfrevic, V. Berghella, C. Bilardo, E. Hernandez-Andrade, S. L. Johnsen, K. Kalache, K.-Y. Leung, G. Malinger, H. Munoz, F. Prefumo, A. Toi, W. Lee, and on behalf of the ISUOG Clinical Standards Committee. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics & Gynecology*, 37(1):116–126, 2011.
- [180] LJ Salomon, JP Bernard, M Duyme, B Doris, N Mas, and Y Ville. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound in obstetrics & gynecology*, 27(1):34–40, 2006.

- [181] I. Sarris, C. Ioannou, M. Dighe, A. Mitidieri, M. Oberto, W. Qingqing, J. Shah, S. Sohoni, W. Al Zidjali, L. Hoch, D. G. Altman, A. T. Papageorghiou, International Fetal , and Newborn Growth Consortium for the 21st Century. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. *Ultrasound in Obstetrics & Gynecology*, 38(6):681–687, Dec 2011.
- [182] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.
- [183] Sachita Shah, Blaise A Bellows, Adeyinka A Adedipe, Jodie E Totten, Brandon H Backlund, and Dana Sajed. Perceived barriers in the use of ultrasound in developing countries. *Critical ultrasound journal*, 7(1):1–5, 2015.
- [184] Sachita P Shah, Henry Epino, Gene Bukhman, Irene Umulisa, JMV Dushimiyimana, Andrew Reichman, and Vicki E Noble. Impact of the introduction of ultrasound services in a limited resource setting: rural rwanda 2008. *BMC international health and human rights*, 9(1):4, 2009.
- [185] Robert S Shapiro, Jon Wagreich, RB Parsons, A Stancato-Pasik, Hsu-Chong Yeh, and Roger Lao. Tissue harmonic imaging sonography: evaluation of image quality compared with conventional sonography. *AJR. American journal of roentgenology*, 171(5):1203–1206, Nov 1998.
- [186] Min C. Shin, Dmitry B. Goldgof, Kevin W. Bowyer, and Savvas Nikiforou. Comparison of edge detection algorithms using a structure from motion task. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 31(4):589–601, 2001.
- [187] Jamie Shotton, Andrew Blake, and Roberto Cipolla. Multiscale categorical object recognition using contour fragments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1270–1281, 2008.
- [188] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [189] Stephanie Sippel, Krithika Muruganandan, Adam Levine, and Sachita Shah. Review article: Use of ultrasound in the developing world. *International journal of emergency medicine*, 4(1):1–11, 2011.
- [190] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.
- [191] Jacqueline K Spencer and Ronald S Adler. Utility of portable ultrasound in a community in ghana. *Journal of Ultrasound in Medicine*, 27(12):1735–1743, 2008.

- [192] Richard V. Stebbing and John E. McManigle. A boundary-fragment-model for head segmentation in fetal ultrasound. In *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images ISBI 2012*, 2012.
- [193] Richard V. Stebbing, John E. McManigle, and J. Alison Noble. Interpreting edge information for improved endocardium delineation in echocardiograms. In *Proc. 9th IEEE Int Biomedical Imaging (ISBI) Symp*, pages 238–241, 2012.
- [194] W Stein, I Katunda, and C Butoto. A two-level ultrasonographic service in a maternity care unit of a rural district hospital in tanzania. *Tropical doctor*, 38(2):125–126, 2008.
- [195] JP Steinmetz and JP Berger. Ultrasonography as an aid to diagnosis and treatment in a rural african hospital: a prospective study of 1,119 cases. *The American journal of tropical medicine and hygiene*, 60(1):119–123, 1999.
- [196] C Sun. Automatic fetal head measurements from ultrasound images using circular shortest paths. *Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012*, pages 13–15, 2012.
- [197] Kenji Suzuki, Pingkun Yan, Fei Wang, and Dinggang Shen. Editorial: machine learning in medical imaging. *Journal of Biomedical Imaging*, 2012:2, 2012.
- [198] Richard Szeliski and James Coughlan. Spline-based image registration. *International Journal of Computer Vision*, 22(3):199–218, 1997.
- [199] J.G. Thomas, II Peters, R.A., and P. Jeanty. Automatic segmentation of ultrasound images using morphological operators. *Medical Imaging, IEEE Transactions on*, 10(2):180–186, Jun 1991.
- [200] Kinh Tieu and Paul Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, 2004.
- [201] Denise Tiran. Nice guideline on antenatal care: routine care for the healthy pregnant woman - recommendations on the use of complementary therapies do not promote clinical excellence. *Complementary therapies in clinical practice*, 11(2):127–129, 2005.
- [202] Nynke Van Den Broek, Chikonde Ntonya, Edith Kayira, Sarah White, and James P Neilson. Preterm birth in rural malawi: high incidence in ultrasound-dated population. *Human Reproduction*, 20(11):3235–3237, 2005.
- [203] G. A. van Veelen, K. J. Schweitzer, and C. H. van der Vaart. The reliability of three- and four-dimensional pelvic floor ultrasound measurements in primigravid and primiparous women: implications for training. *Ultrasound Obstet Gynecol*, n/a–n/a:n/a–n/a, May 2013.
- [204] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

- [205] Andrea Vedaldi and Brian Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.
- [206] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, 2012.
- [207] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, March 2012.
- [208] Rene Vidal and Avinash Ravichandran. Optical flow estimation & segmentation of multiple moving dynamic textures. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 516–521. IEEE, 2005.
- [209] Rebecca L. Vieira, Deborah Hsu, Joshua Nagler, Lei Chen, Rachel Gallagher, and Jason A. Levy. Pediatric emergency medicine fellow training in ultrasound: consensus educational guidelines. *Academic Emergency Medicine*, 20(3):300–306, Mar 2013.
- [210] J Villar, DG Altman, M Purwar, JA Noble, HE Knight, P Ruyan, L Cheikh Ismail, FC Barros, A Lambert, AT Papageorgiou, M Carvalho, YA Jaffer, E Bertino, MG Gravett, ZA Bhutta, SH Kennedy, the International Fetal, and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). The objectives, design and implementation of the intergrowth-21st project. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120:9–26, 2013.
- [211] J. Villar, H. E. Knight, M. de Onis, E. Bertino, G. Gilli, A. T. Papageorgiou, L Cheikh Ismail, F. C. Barros, Z. A. Bhutta, International Fetal , and Newborn Growth Consortium (INTERGROWTH-21st). Conceptual issues related to the construction of prescriptive standards for the evaluation of postnatal growth of preterm infants. *Archives of disease in childhood*, 95(12):1034–1038, Dec 2010.
- [212] José Villar, Leila Cheikh Ismail, Cesar G Victora, Eric O Ohuma, Enrico Bertino, Doug G Altman, Ann Lambert, Aris T Papageorgiou, Maria Carvalho, Yasmin A Jaffer, et al. International standards for newborn weight, length, and head circumference by gestational age and sex: the newborn cross-sectional study of the intergrowth-21 st project. *The Lancet*, 384(9946):857–868, 2014.
- [213] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.
- [214] S. V. Vishwanathan, Alexander J. Smola, and Rene Vidal. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *Int. J. Comput. Vision*, 73(1):95–119, June 2007.

- [215] S. Z. Wanyonyi, R. Napolitano, E. O. Ohuma, L. J. Salomon, and A. T. Papa-georgiou. Image-scoring system for crown rump length measurement. *Ultrasound in Obstetrics and Gynecology*, 44(6):649–654, 2014.
- [216] Kaizhi Wu, Xi Chen, and Mingyue Ding. Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. *Optik-International Journal for Light and Electron Optics*, 125(15):4057–4063, 2014.
- [217] Mohammad Yaqub, M. Kassim Javaid, Cyrus Cooper, and J. Alison Noble. Improving the classification accuracy of the classic rf method by intelligent feature selection and weighted voting of trees with application to medical image segmentation. In *Proceedings of the Second international conference on Machine learning in medical imaging*, MLMI’11, pages 184–192, Berlin, Heidelberg, 2011. Springer-Verlag.
- [218] Mohammad Yaqub, Raffaele Napolitano, Christos Ioannou, Aris T Papa-georgiou, and J Alison Noble. Automatic detection of local fetal brain structures in ultrasound images. In *Proc. 9th IEEE Int Biomedical Imaging (ISBI) Symp*, pages 1555–1558, 2012.
- [219] Zhao Yi, Antonio Criminisi, Jamie Shotton, and Andrew Blake. Discriminative, semantic segmentation of brain tissue in mr images. In Guang-Zhong Yang, David Hawkes, Daniel Rueckert, Alison Noble, and Chris Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009*, volume 5762 of *Lecture Notes in Computer Science*, pages 558–565. Springer Berlin Heidelberg, 2009.
- [220] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
- [221] Jin-Hua Yu, Yuan-Yuan Wang, Ping Chen, and Hui-Ying Xu. Two-dimensional fuzzy clustering for ultrasound image segmentation. In *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*, pages 599–603, July 2007.
- [222] Jinhua Yu, Yuanyuan Wang, and Ping Chen. Fetal ultrasound image segmentation system and its use in fetal weight estimation. *Medical & biological engineering & computing*, 46(12):1227–1237, 2008.
- [223] Jinhua Yu, Yuanyuan Wang, Ping Chen, and Yuzhong Shen. Fetal abdominal contour extraction and measurement in ultrasound images. *Ultrasound in Medicine & Biology*, 34(2):169 – 182, 2008.
- [224] I. E. Zador, V. Salari, L. Chik, and R. J. Sokol. Ultrasound measurement of the fetal head: computer versus operator. *Ultrasound in Obstetrics and Gynecology*, 1(3):208–211, 1991.

- [225] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007.
- [226] Weiwei Zhang, J.A. Noble, and J.M. Brady. Real time 3d ultrasound to mr cardiovascular image registration using a phase-based approach. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 666–669, 2006.