Medical Image Analysis 37 (2017) 22–36

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat

M.A. Maraci[a,*], C.P. Bridge[a], R. Napolitano[b], A. Papageorghiou[b], J.A. Noble[a]

[a] Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[b] Nuffield Department of Obstetrics and Gynaecology, John Radcliffe Hospital, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Received 3 December 2015
Revised 22 December 2016
Accepted 5 January 2017
Available online 10 January 2017

Keywords:
Ultrasound video
Fetal presentation and heartbeat
Machine learning

ABSTRACT

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 on-going pregnancy and the latter is essential for labour management. We propose an automated framework for detection of fetal presentation and heartbeat from a predefined free-hand ultrasound sweep of the maternal abdomen. Our method exploits the presence of key anatomical sonographic image patterns in carefully designed scanning protocols to develop, for the first time, an automated framework allowing novice sonographers to detect fetal breech presentation and heartbeat from an ultrasound sweep. The framework consists of a classification regime for a frame by frame categorization of each 2D slice of the video. The classification scores are then regularized through a conditional random field model, taking into account the temporal relationship between the video frames. Subsequently, if consecutive frames of the fetal heart are detected, a kernelized linear dynamical model is used to identify whether a heartbeat can be detected in the sequence. In a dataset of 323 predefined free-hand videos, covering the mother's abdomen in a straight sweep, the fetal skull, abdomen, and heart were detected with a mean classification accuracy of 83.4%. Furthermore, for the detection of the heartbeat an overall classification accuracy of 93.1% was achieved.

1. Introduction

There have been significant advances in the analysis of ultrasound images in the last decade due in part to increased image quality but also the introduction of modern machine learning into the medical image analysis field (Noble, 2016). Machine learning is arguably very well-suited to recognize sonographic patterns in ultrasound images, which can form the basis of image-based decision-making. By contrast, traditional biomedical image analysis methods can find the dropouts, shadows, and sonographic signatures characteristic of ultrasound images difficult to accommodate, as they are the mapping of anatomy through the ultrasound image formation process. The most successful traditional methods in the literature are model-based methods that use strong geometric models as priors to cope with missing boundaries and artefacts.

Our particular interest is in obstetric ultrasound. The majority of the image analysis literature in this area has focused on automation of fetal biometry measurement for the anomaly scan (taken at 18–22 weeks gestational age). See Challenge US (Rueda et al., 2014) for a recent challenge that looked at a variety of methods and their performances. The anomaly scan is an essential ultrasound screening examination recommended worldwide for the detection of fetal abnormalities and early fetal growth restriction (Tiran, 2005). During a scan, a skilled sonographer acquires and records a number of two dimensional (2D) images of key fetal structures in diagnostic planes, following a standardized clinical protocol (typically a minimum of 6 but often more than 20 images) (Salomon et al., 2011). The goal is to diagnose structural abnormalities and to acquire biometry measurements that are verified against fetal growth charts. Research has looked into automating biometry measurement. For instance, Carneiro et al. (2008) used 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 key structures, where the nodes of the binary tree are strong classifiers trained using AdaBoost. Rahmatullah et al. (2011b); 2011a) used Adaboost for anatomical object detection in 2D fetal abdominal ultrasound images, where their framework was designed to identify whether the correct abdominal landmarks required for a standard plane

* Corresponding author.
E-mail address: mohammad.maraci@eng.ox.ac.uk (M.A. Maraci).

http://dx.doi.org/10.1016/j.media.2017.01.003

were present. Sun (2012) applied a graph-based approach for automatic detection of the fetal skull, where initially the shortest circular path was detected. An ellipse was then fitted to the shape for finding the skull boundary. Ponomarev et al. (2012) applied a multilevel thresholding approach combined with edge detection and shape-based recognition for segmentation of the fetal skull. Imaduddin et al. (2015) used Haar-like feature with AdaBoost to detect fetal skull and femur. They further applied a Randomized Hough Transform for making biometry measurements. Anto et al. (2015) used a Random Forest to segment a head contour in fetal ultrasound scans that were acquired with a low-cost probe. Perhaps the most similar work to our own is the work of Lei et al. (2015), where densely sampled RootSIFT features were extracted and encoded using Fisher vectors for automatic recognition of fetal facial standard planes.

Three dimensional (3D) ultrasound was introduced in the 1990s as a technology designed to improve clinical workflow. It aimed to replace multiple 2D acquisitions by a single 3D acquisition, followed by standard plane finding in the volume. However, manual standard plane finding is quite time-consuming. This has led to a number of methods being proposed for automated plane finding (Chykeyuk et al., 2014; Yaqub et al., 2015) and some commercial systems now have automated plane finding as an option. However, the images from a 3D acquisition have a different appearance to those of a 2D acquisition and hence can contain different diagnostic value. It remains to be seen whether this type of solution will become accepted clinically. Quantification of 3D fetal ultrasound has, however, shown some promising results. For instance, Yaqub et al. (2011) successfully used 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 (Yaqub et al., 2012). Namburete et al. (2015) used Regression Forests to estimate the gestational age of a foetus from sonographic signatures in the brain. In the latter case, the accuracy of the method in the third trimester was shown to be higher than the current clinical standard.

It is important to note, though not often discussed, that in both standard 2D and 3D fetal sonography screening a sonographer follows a standardized clinical protocol, which defines criteria for the plane definition - see for instance the ISUOG guidelines for standard plane criteria (Salomon et al., 2011). Standardized 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 (Sarris et al., 2011) and image quality (Wanyonyi et al., 2014). We refer to this standardized 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 visualization 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 foetus. 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 for uptake of ultrasound in the developing world, the need to acquire constrained scans has to be relaxed in favour of much simpler scanning protocols that a non-expert can readily learn.

[1] In the clinical setting this is referred to as a *standardized scan.*

Encouraging results from observational studies demonstrated that trained and standardized healthcare workers in developing countries can perform as well as qualified sonographers in terms of measurements reproducibility (Rijken et al., 2009). An automatic video acquisition analysis could potentially help in training, standardization 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. 1a. 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 article. The question is then what useful diagnostic information can be automatically analysed from such videos?

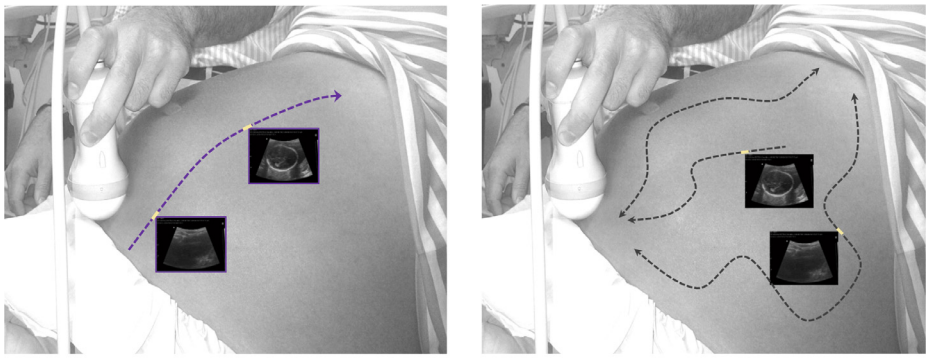
To place our work in perspective, Fig. 2 schematically summarizes how some of the current state-of-the-art literature in fetal ultrasound image analysis maps between the skill needed for acquisition and type of image interpretation and analysis (none, detection & localization, quantification). As can be seen, most image analysis literature is in the lower third of this graph (data acquired by a skilled sonographer). We have included the assisted free-hand works of Kadour and Noble (2009); Kadour et al. (2010); Brown et al. (2013), which use controlled mechanical movement of the probe or subject for elastography on the middle row. These methods generate visualization of ultrasound information and require a small amount of user input to guide probe placement.

In recent years, several methods have been proposed for automatic detection and localization of anatomical fetal structures from ultrasound videos. Linear Dynamical Systems (LDS) were used to localize structures of interest in an ultrasound video obtained from a phantom by Kwitt et al. (2013). In our own work Maraci et al. (2014b), developed independently at around the same time, a method that performed well on clinical ultrasound video sequences was proposed. In that work, the original video is broken into smaller sequences of shorter length, where all sub-sequences have the same length. The 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.

In an attempt to automatically find the image best representing the fetal abdominal standard plane in a video sequence, Kumar and Shriram (2015) used a method based on the spatial configuration of key anatomical landmarks. In previous works on which the current paper builds, we have investigated the bag of visual words approach with feature symmetry filters (Maraci et al., 2014a) as well as improved Fisher vector (IFV) encoding (Maraci et al., 2015) with a support vector machine (SVM) to identify frames of interest in an ultrasound video.

Finally, CNNs are gaining popularity in medical image analysis including analysis of ultrasound images although they are best suited to very large datasets and balanced data (which we do not have in our application). Chen et al. (2015) used a convolutional neural network (CNN) for standard plane localization of the skull and abdomen from an ultrasound video although the details of acquisition were not stipulated. Gao et al. (2016) have recently used a CNN for partitioning ultrasound video and (Baumgartner et al., 2016) for standard plane detection. We discuss CNNs further in the Discussion section.

To the best of our knowledge, the automation of the task of detecting the fetal presentation and heartbeat from a “predefined free-hand” ultrasound video has not been attempted before. We propose a three-step detection framework for characterizing an ultrasound video obtained from a predefined free-hand constrained scan protocol for pregnancies beyond 28 weeks of gestation. The first step in our method automatically identifies the frames corresponding to the fetal skull, abdomen and the heart. This is used



(a) predefined free-hand scan (b) Typical obstetric scan

Fig. 1. A predefined free-hand scan vs. a typical standardized obstetric scan: (a) Sonographer follows a simple scanning protocol for automated analysis to capture some structure of interest. (b) The sonographer scans over multiple paths to locate the best visual representation of the key structures, where they are saved for further analysis.

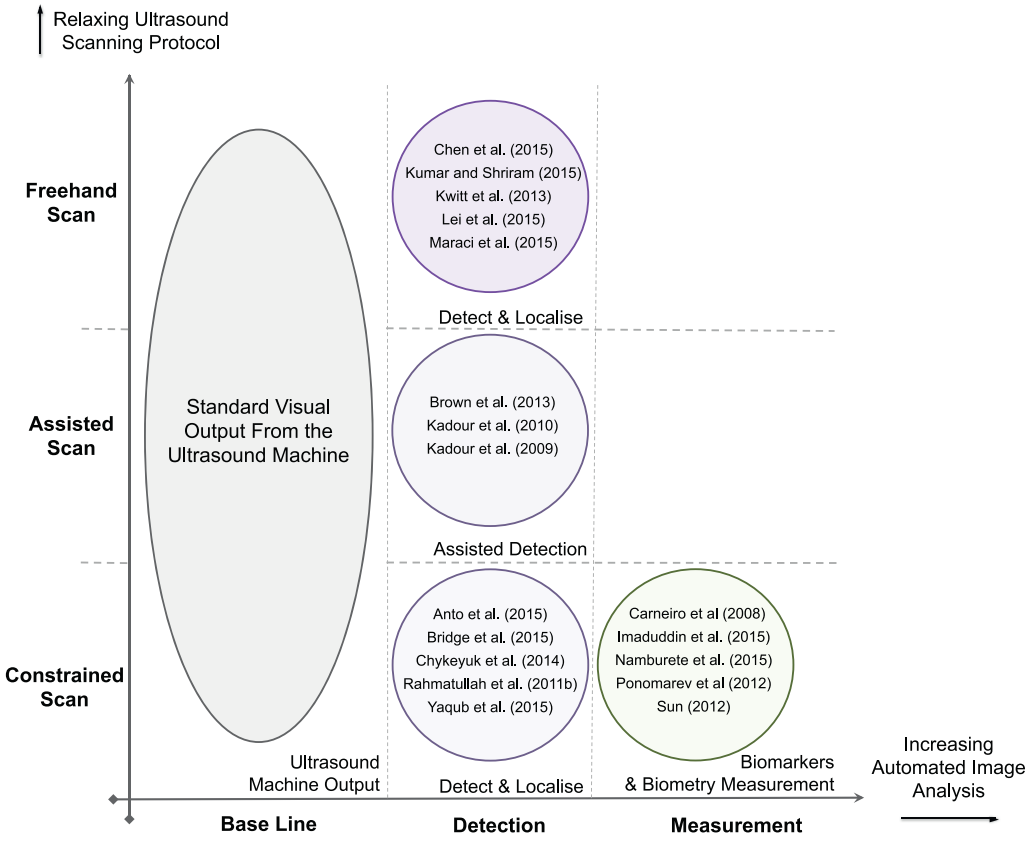


Fig. 2. Ultrasound scan spectrum: Controlled sonographer guidance and automated image analysis increases from left to right, to obtain clinically valid measurements. On the *Y* axis, data acquisition protocol changes from being constrained at the bottom to free-hand on the top.

to infer the fetal presentation as explained in Section 2.2. The second step takes candidate heart frames from the first step to localize the position of the fetal heart as explained in Section 2.3. Finally the dynamics of the fetal heart are modelled from fetal heart frames to identify whether a fetal heart is beating or not. Experiments and results are presented in Section 3, followed by a discussion and conclusion. Earlier versions of some of the component algorithms have been presented in short conference and workshop papers (Maraci et al., 2014b; 2015; Bridge and Noble, 2015). The current paper describes the complete algorithm in detail for the first time and presents substantial experimental evaluation of the complete framework to justify its design.

2. Methods

2.1. Experimental setup

323 videos were acquired from subjects participating in the INTERGROWTH-21ST project (Sarris et al., 2013; Papageorghiou et al., 2014) at the University of Oxford. Data acquisition was carried out using a mid-range ultrasound machine (Philips HD9 with a V7-3 transducer) by a number of experienced obstetricians who were trained for about 10 min to follow the simple scanning protocol. The predefined free-hand ultrasound videos were acquired while moving the transducer from the maternal cervix to the fundus following the longitudinal axis of the uterus as in Fig. 1a. All

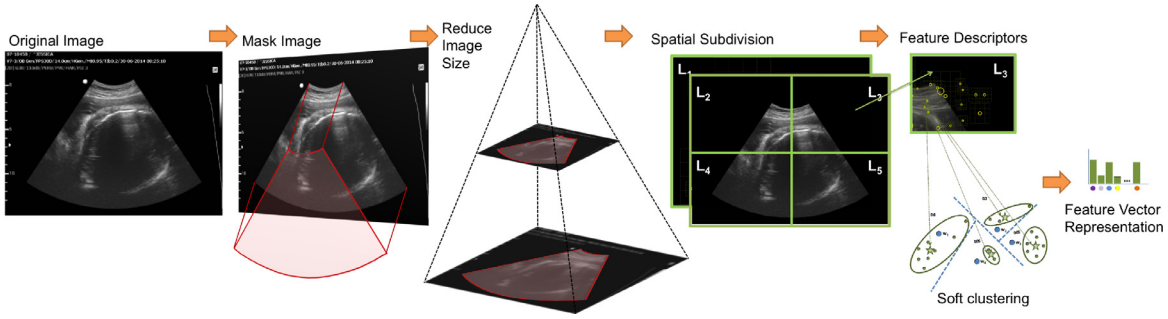


Fig. 3. Steps for feature vector extraction. Preprocessing involves masking each frame and reducing the image size to improve computational cost. Feature extraction (SIFT, rootSIFT, SURF) is then carried out on each image. The extracted features are clustered by a Gaussian mixture model (GMM) and encoded using BoVW, VLAD, or FV encoding.

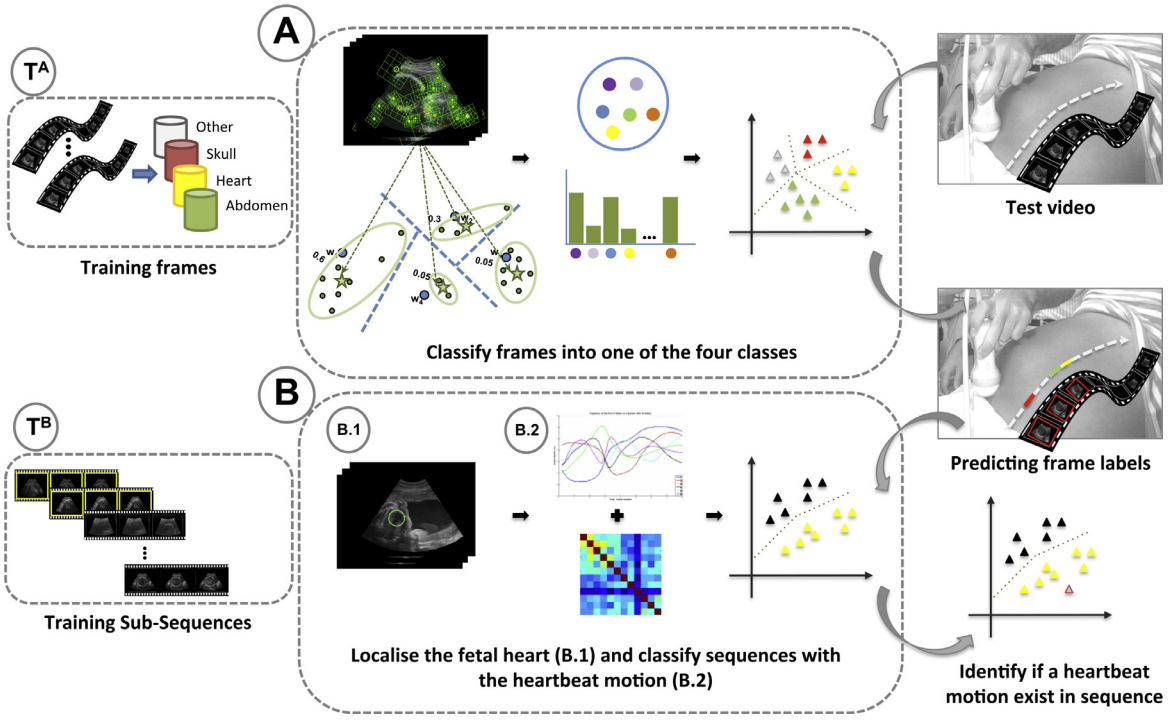


Fig. 4. The main steps of the framework. Given a new training video, all the frames are first classified into "skull", "abdomen", "heart" or "other". If a set of consecutive fetal heart frames are detected in step A, they are further analysed in step B to identify whether a heartbeat can be found. In 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

foetuses had a normal growth according to international standards (Villar et al., 2014).

The feature extraction process is illustrated in Fig. 3 and the full algorithm we have developed is shown schematically in Fig. 4. In step A, a multi-class discriminative classifier - trained using the data in T^A - is deployed to categorize 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 categorize the data into the four classes.

Considering the typical heartbeat frequency of a foetus and scan speed (30 fps) employed in this work, it is assumed that heart motion can be captured in at least 30 frames, if it indeed exists. Therefore, if 30 or more consecutive frames are classified as fetal heart frames 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. In what follows, each of the steps are explained in more detail.

2.2. Step A - Video frame classification

In this subsection we describe the 4-class video frame classification step in more detail. We chose what is sometimes called a hand-crafted feature classification approach rather than deep learning because this class of method is often well-suited to problems defined by relatively small amounts of data (here we had 323 videos), there is significant class imbalance, and the relative richness of features that can represent the problem.

2.2.1. Features

Dense feature extraction, as used in this paper, has become an essential part of many state-of-art image classification methods. In this paper, the speeded up robust feature (SURF) descriptors as described by Bay et al. (2006) and the scale-invariant feature

transform (SIFT) descriptors (Lowe, 2004) were utilized and compared. The SIFT algorithm computes a histogram of local oriented gradients around an interest point and stores the bins in a 128-dimensional vector (8 orientation bins for each of the 4 × 4 location bins). The SURF descriptor describes a distribution of Haar wavelet responses at each interest point neighbourhood and exploits the integral images to estimate Haar features for speed. It results in a 64-dimensional vector and its lower feature dimensions enables a faster detection, at a cost of potentially sacrificing detection accuracy.

In this paper, both features are densely computed over each image with a stride of 4 pixels. Dimensionality reduction of SIFT features using PCA followed by square rooting the feature vectors has been shown to improve classification results (Arandjelović and Zisserman, 2012) in computer vision applications, so we also study its effect on ultrasound images. Additionally, feature vectors are encoded using the traditional bag-of-visual-words (BoVW), vector of locally aggregated descriptors (VLAD) (Jegou et al., 2010), and the improved Fisher vector (FV) (Perronnin et al., 2010) and a comparison between the results of each approach is provided.

The FV encoding approach works by aggregating a large set of feature vectors into a high-dimensional space. A common approach, which we utilize here, is to fit a parametric generative model such as a Gaussian Mixture Model (GMM) to the features and then to encode the derivatives of the log-likelihood of the model with respect to its parameters. First and second order differences between the dense features and each of the GMM centres can then be estimated.

Specifically, given $\mathbf{I} = (\mathbf{x_1}, \ldots, \mathbf{x_N})$ a set of D dimensional SIFT feature vectors extracted from an image, and $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, \ldots, K)$ 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 $\mathbf{x_i}$ to a mode k in the mixture with a strength given by the posterior probability such that,

$$q_{ik} = \frac{exp\Big[-\frac{1}{2}(\mathbf{x_i}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x_i}-\mu_k)\Big]}{\Sigma_{t=1}^{K} exp\Big[-\frac{1}{2}(\mathbf{x_i}-\mu_t)^T\Sigma_k^{-1}(\mathbf{x_i}-\mu_t)\Big]}. \quad (1)$$

Given N SIFT feature vectors, the mean and covariance deviations vectors for each mode k are defined such that,

$$u_{jk} = \frac{1}{N\sqrt{\pi k}}\sum_{i=1}^{N} q_{ik}\frac{x_{ji}-\mu_{jk}}{\sigma_{jk}}, \quad (2)$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi k}}\sum_{i=1}^{N} q_{ik}\Big[\Big(\frac{x_{ji}-\mu_{jk}}{\sigma_{jk}}\Big)^2 - 1\Big], \quad (3)$$

where $j = 1, \ldots, D$ 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,

$$\Phi(\mathbf{I}) = [\mathbf{u_1}^\intercal, \ldots, \mathbf{u_k}^\intercal, \mathbf{v_1}^\intercal, \ldots, \mathbf{v_k}^\intercal]^\intercal. \quad (4)$$

VLAD encoding utilizes a similar approach to Fisher vectors and encodes a set of local feature descriptors, $\mathbf{I} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, extracted from an image using a dictionary built using a clustering method such as Gaussian Mixture Models (GMM) or K-means clustering. More formally, let q_{ik} be the strength of the association of data vector $\mathbf{x}_i$ to cluster μ_k, such that $q_{ik} \geq 0$ and $\sum_{k=1}^{K} q_{ik} = 1$, where the association may be either soft (e.g. obtained as the posterior probabilities of the GMM clusters) or hard (e.g. obtained by vector quantization with K-means). VLAD encodes feature $\mathbf{x}$ by considering the residuals $\mathbf{v}_k = \sum_{i=1}^{N} q_{ik}(\mathbf{x}_i - \mu_k)$. The residuals are stacked together to obtain the vector $\hat{\Phi}(\mathbf{I}) = [\ldots, \mathbf{v_k}^\intercal, \ldots]^\intercal$.

2.2.2. Classification

We use support vector machines (SVM) for classification. One of the advantages of using SVM-based classification is that it allows for an efficient use of kernels. The SVM hyperparameters were tuned based on a small sub-set of the data that was randomly selected. Once the optimal parameters were estimated they were used for training the classifier. For non-linear problems, kernel functions allow the data to be projected to a higher-dimensional *feature space*, where a linear model can then be used to classify the data. Moreover, while linear kernels can be highly efficient (Joachims, 2006), non-linear kernels have shown to produce higher classification accuracy (Zhang et al., 2007). It was shown that square rooting SIFT (*sqrt(SIFT/sum(SIFT))*) is similar to using the non-linear Hellinger's kernel in the original input space, without its computational costs (Arandjelović and Zisserman, 2012).

The classifier is trained to categorize the frames into the four classes of fetal skull, fetal abdomen, fetal heart and "other" structures. As the data used in this study consists of an ordered sequence of frames, temporal information is used to regularize the classification results. In order to utilize this temporal information a conditional random field (CRF) graphical model (Lafferty et al., 2001) is constructed, where each frame of the video is represented as a node in the graph. CRFs have previously been successfully used to regularize machine learning for medical image analysis solutions for example in Bauer et al. (2011); McIntosh et al. (2013); Nowozin et al. (2011). Here the classification scores for each frame are converted into probabilities and used as the node potential in the graph. This setting smooths 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 normalized product of a set of non-negative potential functions,

$$p(f_1, f_2, \ldots, f_N) = 1/Z\prod_{i=1}^{N}\phi_i(f_i)\prod_{e=1}^{E}\phi_e(f_{e_j}, f_{e_k}). \quad (5)$$

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 normalization constant to ensure the distribution sums to one over all possible joint configurations of the variables. Finally, the Viterbi (Forney Jr, 1973) algorithm is used to find the most probable classification result for each frame.

2.3. Step B.1 - Locating the fetal heart

The frame classification procedure described in Section 2.2 is able to identify the frames containing the fetal abdomen. In order to assess fetal viability, it is necessary to detect the location of the heart within these frames. This task is complicated by the fact that, when simple sweeps are used, the orientation of the heart relative to the probe is variable and unknown. We therefore chose to make use of rotation invariant detection methods, first introduced for computer vision applications by Liu et al. (2014) and adapted for fetal echocardiography in Bridge and Noble (2015). An extended version of this work can be found in Bridge et al. (2017).

2.3.1. Rotation invariant features

The method for calculating rotation invariant features 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

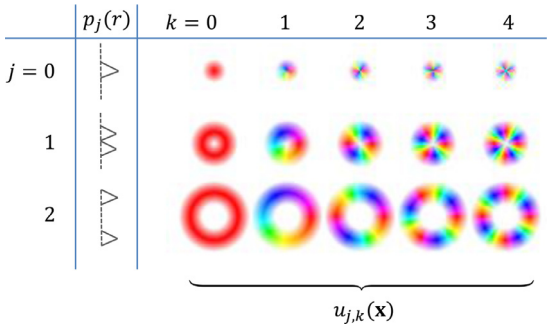


Fig. 5. Set of profiles and basis functions with $J = 3$, $K = 4$ (only $k \geq 0$ displayed). The saturation and hue represent the complex magnitude and argument respectively (Bridge and Noble, 2015).

basis on the angular coordinate, θ (Liu et al., 2014; Bridge and Noble, 2015):

$$b_{j,k}(r,\theta) = p_j(r)e^{ik\theta} \tag{6}$$

for $0 \leq r < R, \quad 0 \leq \theta < 2\pi$, 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. Fig. 5 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 mth coefficient at image position $\mathbf{p}$ is:

$$c_m(\mathbf{p}) = \|v(\mathbf{p})\| e^{-im\arg(v(\mathbf{p}))}, \quad m = 0, 1, \ldots, M. \tag{7}$$

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,

$$D_{j,k,m}(\mathbf{p}) = b_{j,k}(\mathbf{p}) * c_m(\mathbf{p}), \tag{8}$$

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.

2.3.2. Support vector classification

For classification of each window as heart or non-heart we use a linear SVM classifier with the rotation invariant features from Section 2.3.1 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.

Note that the location only needs to be approximate as the next step uses ROIs around the estimated location for heartbeat detection and the accuracy of location is not the critical factor.

2.4. Step B.2 - Detecting the fetal heartbeat

Once a minimum of 30 consecutive video frames of the fetal heart are identified and the heart is localized using the procedures described in Section 2.3, they are compiled together to form a short video sub-sequence. Our goal is to derive a model of a heartbeat in terms of the intensity patterns in this video subsequence. Moreover, we investigate the accuracy of the framework when learning the dynamics on heart ROI compared to the entire image. The positive training examples used are short video sequences of a fetal heartbeat and the negative training sequences are short video sequences that do not contain a heartbeat, randomly extracted from the videos in dataset. Therefore, the classifier is trained to perform a binary classification to identify whether any given sequence, during test time, contains the correct dynamics and motion that corresponds to a fetal heartbeat.

Specifically, 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. (2003) for the system identification of the model, which models pixel intensities in each frame as the output of a LDS. However as opposed to using the raw pixel intensities, we instead use the output of frames filtered by a feature symmetry filter (Rajpoot et al., 2009), which produces a contrast invariant representation of structures on each frame. 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. Therefore, given an ultrasound sequence $\mathbf{S}$ of T video frames, let $\mathbf{S} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$, where $\mathbf{y}_t \in \mathbb{R}^d$ refers to the frame observed at time t. It is assumed that at each time instant t, a noisy version of the image can be measured, $\mathbf{y}(t) = \mathbf{S}(t) + \mathbf{w}(t)$, 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 positive measured sequence $\mathbf{y}(t) \in \mathbb{R}^d$ for $t = 1, \ldots, T$. The evolution of an LDS can be modelled as:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t \\ \mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{w}_t \end{cases} \tag{9}$$

Here $\mathbf{x}_t \in \mathbb{R}^n$ is the state of the LDS and $\mathbf{y}_t \in \mathbb{R}^d$ are 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.

In a linear system such as Eq. 9, the output matrix $\mathbf{C}$ can be estimated via singular value decomposition of the observation matrix $\mathbf{S}$, where $\mathbf{C}$ can be restricted to the N largest eigenvalues. However, here a non-linear model known as a Kernel Dynamic Texture (KDT) (Chan and Vasconcelos, 2007; Kwitt et al., 2013) is used where the evolution of the hidden states of the model are kept linear. In order to capture the dynamics of the video the output matrix $\mathbf{C}$ is replaced by a non-linear observation function $C : \mathbb{R}^n \to \mathbb{R}^d$. Therefore given the same ordered ultrasound sequence $\mathbf{S} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ and a kernel function $k(\mathbf{y}_1, \mathbf{y}_2)$ with associated feature transformation $< \phi(\mathbf{y}_1), \phi(\mathbf{y}_2) >$, the cth eigenvector $\mathbf{kv}_c$ can be used to obtain the cth kernel principal component in the feature space:

$$\mathbf{kv}_c = \sum_{i=1}^{T} \alpha_{i,c}\phi(\mathbf{y}_i) \tag{10}$$

where $\alpha_{i,c}$ represents the ith component of the cth weight vector and $\boldsymbol{\alpha}_c = \frac{1}{\sqrt{\lambda_c}}\mathbf{kv}_c$, assuming the eigenvectors are sorted in descending order of the eigenvalues $\{\lambda_c\}_{c=1}^{T}$. Here λ_c and $\mathbf{kv}_c$ are the cth largest eigenvalue and eigenvector of the kernel matrix $\mathbf{K}$ respectively. Finally the sequence of hidden states $\mathbf{X}$ and the state transition matrix $\mathbf{A}$ can be estimated as

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\alpha}^{\intercal}\mathbf{K} \\ \mathbf{A} &= [\mathbf{x}_1, \ldots, \mathbf{x}_{T-1}][\mathbf{x}_0, \ldots, \mathbf{x}_{T-2}]^{\dagger} \end{aligned} \tag{11}$$

2.4.1. 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. Prior work on comparison metrics of LDSs range from metrics based on subspace angles between the observability subspaces of the systems (De Cock and De Moor, 2000) to metrics based on the Binet–Cauchy kernels (Vishwanathan et al., 2007; Bissacco et al., 2007) and finally metrics based on the KL-divergence of the probability distributions of the stochastic processes (Chan and Vasconcelos, 2005). A full comparison of these classes of metrics is outside the scope of this paper. However, Chaudhry and Vidal (2009) illustrated 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. Furthermore, we have previously shown (Maraci et al., 2014b) that the Binet–Cauchy maximum singular values kernel produced superior results on medical ultrasound data.

The Binet–Cauchy (BC) singular value kernel (Vishwanathan et al., 2007) used in this paper can be explained as an extension of the BC trace kernel. Given two LDS models $\mathbf{M}_1$ and $\mathbf{M}_2$ (represented by their model parameters), with corresponding sequences $\{y_t^{M_i}\}_{t=1}^{T}$ that have the same underlying noise process, the trace kernel for the two non-linear dynamical systems (NLDS) is as follows:

$$K_{NLDS}(\mathbf{M_1}, \mathbf{M_2}) := \mathbb{E}_{v,w}\Big[\sum_{t=0}^{\infty} \lambda_t k(\mathbf{y_t^1}, \mathbf{y_t^2})\Big], \tag{12}$$

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 v_t and w_t. Thus the BC trace kernel for NLDS is defined as

$$K_{NLDS}(\mathbf{M_1}, \mathbf{M_2}) = \mathbf{x}_0^{\intercal}\bar{\mathbf{P}}\mathbf{x}_0^{\intercal} + \frac{\lambda}{1-\lambda} trace(\mathbf{Q}\bar{\mathbf{P}} + \mathbf{R}) \tag{13}$$

where $\mathbf{x}_0$ is the initial state of the system, $\bar{\mathbf{P}} = \sum_{t=0}^{\infty} \lambda_t (\mathbf{A}_t^1)^{\intercal}\mathbf{F}\mathbf{A}_t^2$, $\mathbf{F}$ is the inner product matrix between all the Kernel PCA (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. (2009) 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 NLDS. 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

$$K(\mathbf{M}_1, \mathbf{M}_2) = \frac{K(\mathbf{M_1}, \mathbf{M_2})}{\sqrt{K(\mathbf{M_1}, \mathbf{M_1}), K(\mathbf{M_2}, \mathbf{M_2})}} \tag{14}$$

A distance between two sequences with LDS parameters $\mathbf{M_1}$ and $\mathbf{M_2}$ can now be computed as $d(\mathbf{M_1}, \mathbf{M_2}) = 2(1 - K(\mathbf{M_1}, \mathbf{M_2}))$. This distance is then used as the kernel in an SVM classification framework to identify the presence or absence of a fetal heartbeat in the sequence.

3. Results and discussion

Experiments were designed to evaluate the accuracy of the proposed framework. The first experiment evaluated the accuracy of the frame classification task, including the use of different low-level features and SVM kernels. The second experiment compared detecting heartbeats on full images with first localizing a region of interest (ROI) around the heart and only detecting the heartbeat from analysis of heartbeat ROIs.

Table 1
Mean classification accuracies. The most accurate configurations for the different features and encoding strategies, over the number of words. Breakdown plots are shown in Appendix A.

No. Words ↓	$SIFT_{L1}$	$SIFT_{L5}$	$rootSIFT_{L1}$	$rootSIFT_{L5}$	$SURF_{L5}$
10	74.9	79.5	72	78.4	72.5
20	77.4	80.3	78	79.1	74.8
40	81.5	81.7	80.3	81	77.7
60	82.2	83	81	82.3	77.5
80	80.7	82	81.8	82.8	77.9
100	81.5	82.5	82.7	83	78.5

3.1. Classifying video frames

In order to ensure training and test data are independent, a five-fold cross validation procedure was implemented for training the classifier. At each training step, the model was trained on four fifths of the videos (260 videos) and tested on the unseen one fifth (65 videos). RootSIFT, SIFT, and SURF descriptors were calculated on each 240 × 320 image with a stride of 4 pixels. Moreover, SIFT and rootSIFT descriptors were calculated at 9 different scales with a scaling factor of $\sqrt{2}$. As the ultrasound data is only visible within the ultrasound fan (field of view), all feature descriptors were only computed within the bounding box around this region to avoid calculating redundant information. The number of words (GMM clusters) was varied from 10 clusters to 100 and the three feature encoding techniques (BoVW, VLAD, and FV) were utilized to encode each image before classification. Furthermore, to investigate the effect of SIFT feature dimensionality reduction on classification accuracy, on the experiments in which the number of words exceed 60, SIFT features were decorated and reduced in dimensions from 128D to 40D and 20D, as suggested in Chatfield et al. (2011). The effect of subdividing the data into 1 × 1 and 2 × 2 spatial subdivisions were also investigated. Here, for each tile, the corresponding features were computed and stacked as one. In addition, the effect of using larger SIFT descriptor patches was investigated, by varying the SIFT patch size (8 × 8, 16 × 16, 32 × 32, and 64 × 64 pixels). Finally, the accuracy of using different SVM kernels, namely the linear kernel, Hellinger kernel and the χ^2 kernel was investigated. Fig. 6summarizes the classification accuracies for each of the four classes, where the number of words vary from 10 to 100. The experiments were repeated using the BoVW, illustrated using black colour, VLAD illustrated using blue, and FV encoding illustrated using cyan. For the experiments where PCA is used to reduce feature dimensions, the classification accuracy is illustrated using a single point on the plot, indicated by the same colour and pointer shape. Finally, *L*1 indicates no spatial subdivisions and *L*5 indicates the additional 4 spatial subdivisions. As can be seen, generally, increasing the number of words up to 80, improves classification results but a further increase to 100 does not show any substantial improvement to the classification accuracies. Regardless of the use of spatial subdivision, the skull and "other" classes have the best performance and fetal heart is the class that performs the worst. Moreover, Figs A.11, A.12, and A.13 show the mean classification accuracies where the number of GMM clusters have varied between 10 and 100 utilizing different features (SIFT, rootSIFT, and SURF), feature encoding techniques (BoVW, VLAD, FV), and SVM kernels (linear, Hellinger, χ^2). Similarly, Figs A.14, A.15, and A.16 show the mean average precision for the same experiments. A summary of the most accurate configurations are illustrated in Tables 1 and 2.

As can be seen from Figs A.11 and A.12, the gain in accuracy is only marginal when the number of GMMs is extended beyond eighty clusters. Moreover, when the SIFT and rootSIFT features are used, the χ^2 SVM kernel results in the worst performance com-

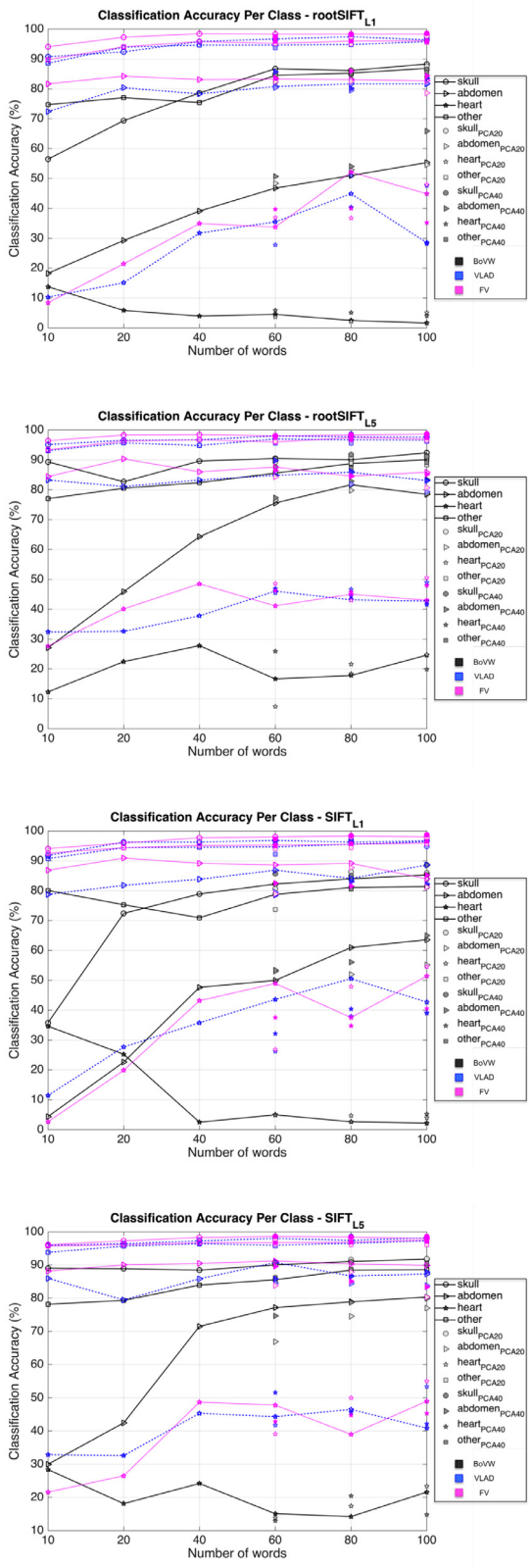


Fig. 6. Classification accuracies for skull, abdomen, heart, and other structures. Individual class accuracies are reported for SIFT and rootSIFT features, while varying the encoder type (BoVW, VLAD, FV) and number of words. A SVM classifier with Hellinger kernel is utilized.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Mean average precision. The most accurate configurations for the different features and encoding strategies, over the number of words. Breakdown plots are shown in Appendix A.

No. Words ↓	SIFT$_{L1}$	SIFT$_{L5}$	rootSIFT$_{L1}$	rootSIFT$_{L5}$	SURF$_{L5}$
10	82.4	88.5	81.2	88.1	81.6
20	87.3	90.3	87.9	90.1	82.8
40	89.6	91.8	89.2	90.9	85.8
60	90.9	92.5	90.4	92.3	86.3
80	91.2	92.8	90.9	92.7	86
100	92	93.3	92.2	93.4	87.2

pared to using the linear or the Hellinger's kernel. Generally the results for the other two kernels are very similar with minor improvements when the Hellinger kernel is used. For the SURF features, the choice of the kernel does not have a dramatic effect on the accuracy.

As for the different feature encodings, FV encoding demonstrates a small gain in accuracy across the experiments compared to using VLAD or BovW. PCA dimensionality reduction can also provide a small boost to the accuracy when 80 or more words are used. It is interesting to note that the use of spatial subdivision boosts the classification results as smaller structures can be better learned when the feature descriptor is augmented by spatial subdivision. Figs. A.11, A.12, A.13, A.14, A.15, and A.16 show the mean classification and mean average precisions results. Moreover, the most accurate cofiguration in these figures are summarized in Tables 1 and 2.

It is worth noting that using PCA to reduce the feature dimensionality to 20 reduces the classification performance results in all experiments. However, for rootSIFT descriptors, using PCA to reduce the feature dimensionality to 40 improves the performance when spatial subdivisions are used, but reduces the performance when spatial subdivisions are not used. This can be explained by the fact that rootSIFT$_{L1}$ descriptors capture less information compared to rootSIFT$_{L5}$, and thus reducing their dimensions even further results in loss of vital discriminative information. It is interesting to note that SIFT$_{L1}$ and SIFT$_{L5}$ features illustrate a similar effect when PCA is applied to reduce feature dimensionality, whereby a decreased classification accuracy is observed. Figs. A.13 and A.16 show plots of the mean classification accuracy and mean average precision that have been obtained using the SURF feature descriptor. Similar to the previous experiments, FV encoding results in better performance compared to the other encoding techniques. In addition, the fetal skull and "other" classes have the best classification performance and the fetal heart is shown to be the most challenging class. In order to better understand the effect of the three encoding techniques, the SVM kernels, and PCA dimensionality reduction on the classification accuracy of each individual class, an experiment was conduction where the number of GMM clusters was fixed to 80. The results are shown in Fig. 7. It is interesting to note that FV and VLAD encoding mainly boost up the classification performance for *skull* and *other* class. Their performance for these two classes are very similar. FV encoding results in slightly better accuracy for the *abdomen* class.

To investigate the effect of various rootSIFT descriptor patch sizes, the number of words was fixed to 80 and PCA was used to reduce the feature dimensions to 40. The mean accuracy and mean average precision (mAP) have been calculated for this experimented and are summarized in Table 3 (bold indicates best results). As can be seen, larger patch sizes improve classification accuracy, especially the results for the fetal heart. This is an intuitive finding as the fetal heart is a small structure and larger descriptors can capture a better representation of structures of interest in relation to other anatomical structures. Moreover, applying

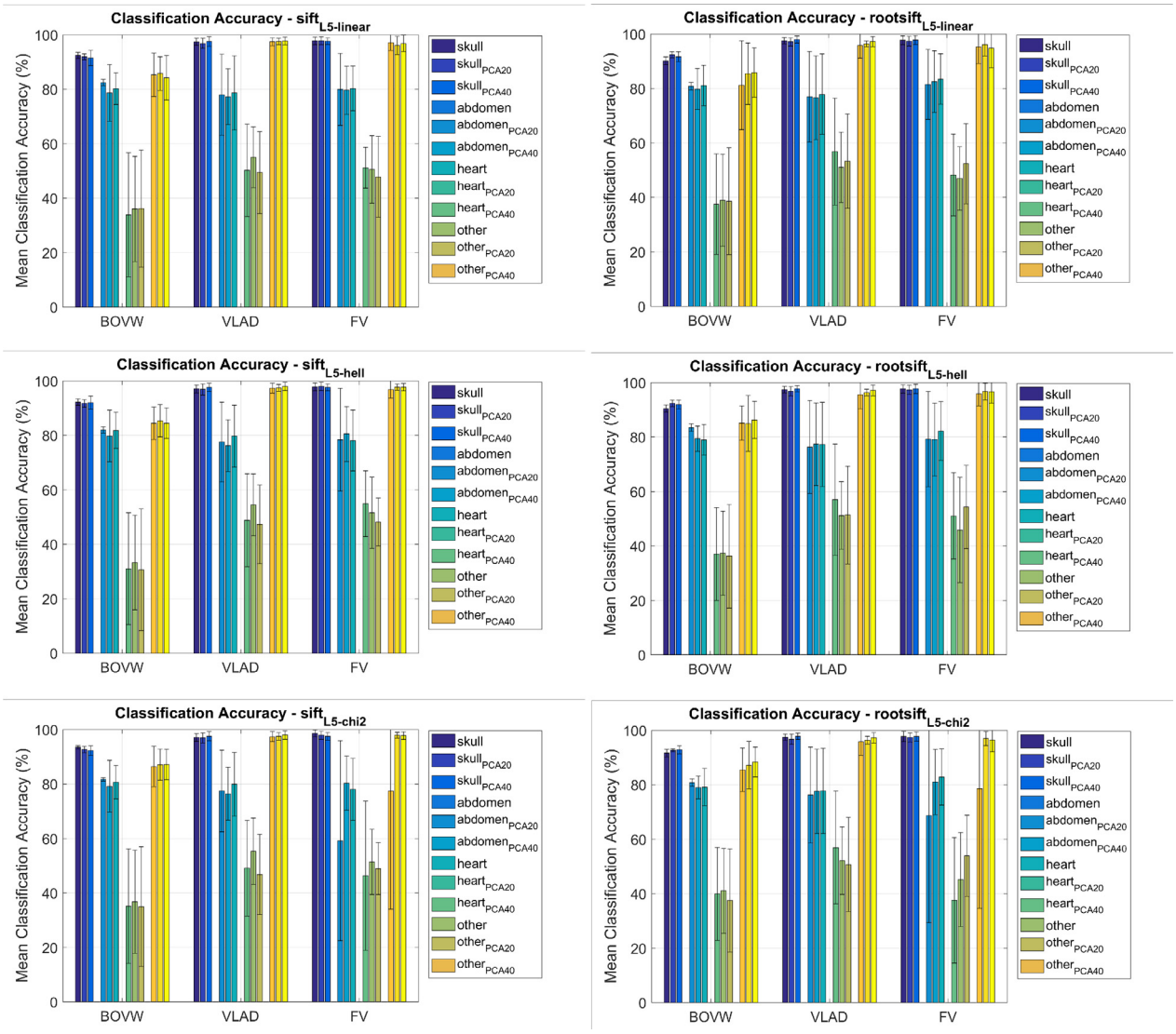


Fig. 7. Mean classification accuracy for all four classes individually. The number of GMM is set to 80 clusters to allow for a performance comparison on each class, using the three encoding techniques, with and without PCA dimensionality reduction.

Table 3
Video frame classification results (Step A). The effect of increasing the rootSIFT 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.

rootSIFT Patch Size	Skull class. Accuracy (%)	Abdomen class. Accuracy (%)	Heart class. Accuracy (%)	Other class. Accuracy (%)	Mean Accuracy (%)	Mean ave. 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

the CRF model to the classification scores regularizes the results and eliminates sudden peaks. This is illustrated in Fig. 8, where the top bar illustrates the raw classification scores. As can be seen, there are a number of frames that have been incorrectly classified as *other* and *abdomen* but applying the CRF regularizes the results as illustrated on the bottom bar. The results show that CRF regularization makes the choice of rootSIFT and SIFT features less significant because it levels their accuracy to a similar level. However, it cannot washout the differences between SURF and rootSIFT or SIFT. This is because the accuracy obtained using the SIFT and rootSIFT features are close, but the SURF features result in a significantly lower accuracy. From a total of 129 unseen videos in the test dataset, 41 videos missed either the skull or abdomen structures as assessed by visual inspection of videos. Unfortunately, keeping the sweeps so simple increases the chance of missing key anatomical structures. Therefore, automatic detection of fetal presentation would not be possible in such scenarios. From the remaining 88 videos, the presentation was correctly identified in 76

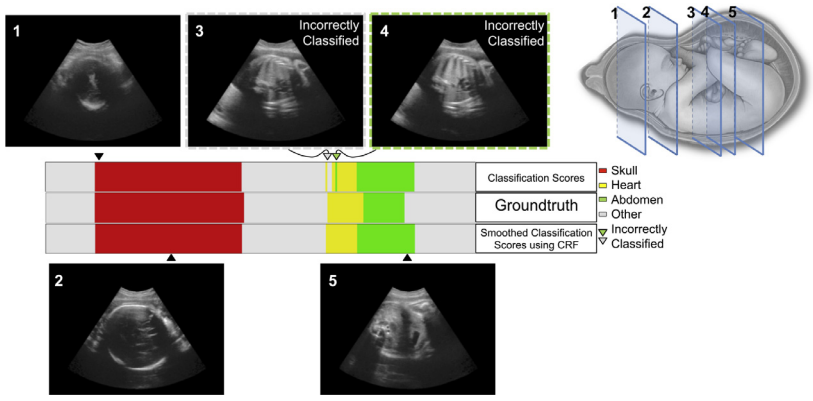


Fig. 8. Classification scores for a test video: Raw classification scores are shown on the top bar and regularized scores on the bottom bar. The red colour represents the frames that have been classified as fetal skull, and similarly yellow, green and grey represent the fetal heart, abdomen and other structures, receptively. As can be seen, the misclassified frames have been relabelled correctly based on their neighbouring frames through the regularization process. Moreover, the slices labelled $1-5$ on the left, correspond to approximate locations of the five sample frames illustrated on the right. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Accuracy of fetal heart localization. The algorithm described in Section 2.3 is used to localize the fetal heart in each sequence and crop the frames around the located heart. Accuracy is reported in terms of the euclidean distance between predicted heart centre point and the groundtruth (GT).

Accuracy indication	%
Within GT diameter	82.4
Within GT radius	65.5
Within half GT radius	55.6

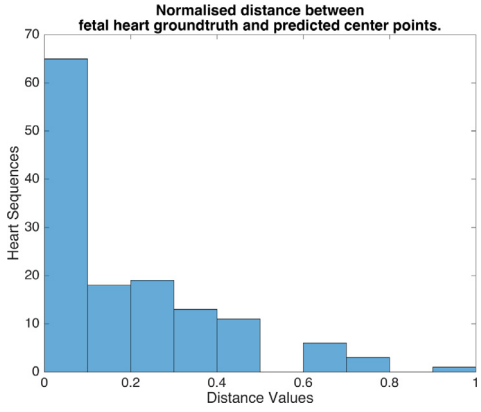


Fig. 9. Normalized Euclidean distance between the centre points of the predicted fetal heart and the groundtruth. A histogram of the normalized euclidean distances between the groundtruth points and the predicted centre points. The histogram is skewed towards lower distance points.

videos sweeps (83.4%). One of the main challenges with free-hand sweeps is to ensure correct anatomical structures are present and displayed appropriately. Inspecting some of the failure cases suggests that *unusual* appearance of the fetal skull or abdomen has contributed to mis-detection or failure to detect the presentation. These include views of the skull or abdomen that had not been seen in the training set and can be addressed in the future studies through larger and more comprehensive datasets.

Generally, the fetal skull and abdomen have significant distinguishing characteristics such as their outer boundaries and inner texture structures. In addition, they both occupy a substantial portion of the image on each frame. However, in our dataset this is not the case for the fetal heart. Due to the simplified scanning protocol, it is easy for fetal heart views to be very similar to those of the fetal abdomen. Moreover, the fetal heart is contained within a very small portion of the image, in comparison to the skull or abdomen. Indeed it may not even be captured as part of sweep at all. Such factors make fetal heart detection and characterization highly challenging in our dataset.

3.2. Localizing the fetal heart

136 short video sequences of a fetal heartbeat each of 30 frames long were extracted from the dataset. The method described in Section 2.3 was applied to find the approximate location of the fetal heart. The Euclidean distance between the predicted centre point of the fetal heart and the ground truth (GT) was calculated. Furthermore, a histogram of the distances is shown in Fig. 9 and the localization accuracy is shown in Table 4. As only the approximate location of the heart is required, the accuracy of this step was evaluated in terms of the Euclidean distance between the predicted and GT centre points of the fetal heart. As can be seen, in 82% of the cases, the distance between the GT and predicted centre point is less than the diameter of the detected fetal heart. This is the maximum permitted distance for an approximate localization of the fetal heart. Moreover, in more than 55% of the sequences the fetal heart has been localized almost perfectly, where the distance between the predicted and GT is less than half the radius of the fetal heart.

3.3. Analysing the fetal heartbeat

136 sequences of the fetal heart were used as positive fetal heartbeat examples. In addition, another 136 short sequences of the same duration were extracted randomly from dataset, where no fetal heart was present. This formed the negative samples. The dataset was split such that 70% was used for training and the remaining sequences were used for evaluating the accuracy of the system. Two experiments were conducted to analyse the dynamics of these subsequences, using the method described in Section 2.4 to identify whether a fetal heartbeat could be detected. The first experiment used the entire ultrasound image, whereas in the second experiment the fetal heart was initially localized following the method described in Section 2.3 and the video frames

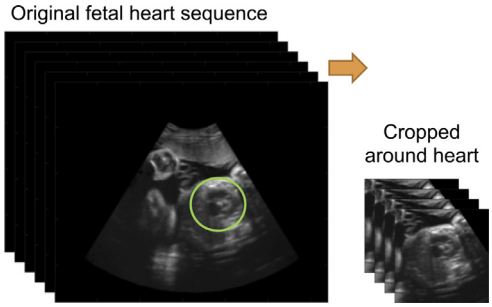


Fig. 10. Cropping around the detected fetal heart. Fetal heart dynamics are analysed once using the original ultrasound sequence (left) and once on the cropped sequence around the detected fetal heart (right).(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Classification results for detecting the fetal heartbeat without localising the fetal heart (Step B). The sequence dimensions are 240 × 320 × 30. Here n indicates the number of KDT model states and σ signifies the filter's centre-wavelength.

	Model n=3	Model n=4	Model n=5	Model n=6	Model n=7
σ =30	80.46	81.61	83.91	83.91	79.31
σ =35	80.46	81.61	83.91	83.91	83.91
σ =40	52.87	82.76	51.72	52.87	45.98
σ =45	83.91	**88.51**	85.06	57.47	68.97
σ =50	**88.51**	86.21	86.21	83.91	52.87
σ =55	55.17	52.87	57.47	52.87	55.17
σ =60	85.06	56.32	56.32	62.07	52.87

Table 6
Classification results for detecting the fetal heartbeat after cropping the frames around the localized fetal heart (Step B). The sequence dimensions are 120 × 120 × 30, cropped around the detected fetal heart. Here n indicates the number of KDT model states and σ signifies the filter's centre-wavelength.

	Model n=3	Model n=4	Model n=5	Model n=6	Model n=7
σ =30	81.61	80.46	80.46	83.91	54.02
σ =35	87.36	88.51	63.22	87.36	86.21
σ =40	49.43	52.87	79.31	78.16	52.87
σ =45	89.66	85.06	52.87	52.87	64.37
σ =50	**93.10**	50.57	89.66	70.11	54.02
σ =55	78.16	51.72	52.87	51.72	74.71
σ =60	63.22	57.47	55.17	64.37	42.53

were cropped around the detected fetal heart as illustrated in Fig. 10.

Recall that this algorithm is only run on frames that have been classified as fetal heart. Fig. 10 shows the heart detection boundary using a green circle. Moreover, as the main application is not accurate heart segmentation, a rectangular area defined by twice the radius of the detection circle plus an offset is empirically set to be the potential area of interest that would contain fetal heart motion. The accuracy values presented in Table 5 are for heartbeat detection without localizing the fetal heart and the accuracy values presented in Table 6 are for the combined localization and heartbeat detection pipeline. The purpose of this step is to assess the accuracy of the motion classification. To elaborate, a dedicated classifier for detecting the fetal hearts was not specifically trained using the sweep data. Instead the best trained model from Bridge and Noble (2015) was applied to the short sequence that have been short-listed in Step A of Fig. 4.

As shown in Table 5 without fetal heart detection, the best results were achieved with a 3-state model and $\sigma_{feat.symm} = 50$ for the signed feature symmetry filter (detection accuracy of 88.5%). In general, the classification accuracy was higher when the heart was first localized and cropped out of the video sequence. This reflects the fact that the full image contains a lot of irrelevant information and motion due to probe movements and fetal movement that can confound the heartbeat detection. When the frames are cropped around the detected fetal heart, a 3-state model and $\sigma_{feat.symm} = 50$ for the signed feature symmetry filter (detection accuracy of 93.1%) produced the highest results. In general, increasing the number of 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 3 or 4 eigenvalues. Additional eigenvalues capture a very small portion of the variation in the feature space, thus resulting in noisier KDT model parameter estimates. Moreover, the duration of the heart sequences in this experiment are considerably short, thus larger increase in the number of states beyond reported does not improve the results.

One of the main challenges in modelling the dynamics of the fetal heart is the quality of the positive and negative samples used to train the dynamical model. Although the positive examples contain motions of beating fetal hearts, our negative dataset does not contain any examples of non-viable (non-beating) fetal heart. Instead the negative dataset consists of short sequences of anything but a fetal heart motion.

4. Conclusions

In this study we have looked at the problem of automatically locating anatomical features in fetal ultrasound video specifically motivated by a real world global health application of low-cost ultrasound for identification of breech presentation and fetal viability. Breech delivery can significantly increase the risk for neonates (Hannah et al., 2000). However, planned vaginal breech deliveries where antenatal ultrasound is available can be associated with a better outcome than reported in randomized trials (Goffinet et al., 2006).

Ultrasound requires a high degree of skill to perform well, and there is a lack of experienced sonographers in many developing world healthcare settings. The image analysis framework we have developed was directly developed to address the need to empower less experienced or well-trained users of ultrasound, or users new to ultrasound to effectively identify structures of interest and interpret the images with high confidence. Further, computer memory requirements for analysis are not large. The solution is amenable to use within a low-cost free-hand ultrasound system platform (where today USB and wireless transducers are of the order of $7k or less).

The implementation reported in the paper was for proof-of-principle and not optimized for real-time use. We have added the processing times but as no attempt was made to optimize them they are not really meaningful from which to infer potential real-time performance. Computation time are shown in Table 7. The experiment was carried out using *Matlab2016a* on a PC with 32GB of RAM, restricting the machine to use only a single core.

This framework assumes that a consecutive sequence of fetal skulls and abdomens are present in any given sweep, in order to identify the fetal presentation. From a total of 129 unseen videos in the test dataset, 41 videos did not contain either the skull or abdomen structures, which are necessary for automatic detection of fetal presentation. From the remaining 88 videos, the presentation was correctly identified in 76 videos sweeps (83.4%). Furthermore,

Table 7
Computation time for encoding SIFT, rootSIFT, and SURF fea- tures using BoVW, VLAD, and FV encoding. The duration for encoding the features for one image, in seconds.

	BoVW	VLAD	FV
SIFT	1.452	0.297	0.165
rootSIFT	1.218	0.286	0.184
SURF	0.267	0.085	0.051

for the detection of the heartbeat an overall classification accuracy of 93.1% was achieved. In the 12 videos where the presentation was not identified correctly, although the fetal and abdomen were present in the ultrasound video sweep, these structures were not captured fully and visually looked different to the majority of those in the training set.

On our choice of classifier, SVM is a classical learning algorithm, which has demonstrated excellent performance in many applications, including our previous work and work of others. For example Lei et al. (2015) recently proposed the use of root-SIFT features with an SVM classification framework for detecting fetal faces in ultrasound scans. We found it gave good results (as evidenced in the article) and did not see the value to move on to consider more sophisticated hand-crafted feature classifiers (for instance, random forests). Convolutional neural networks (CNNs) have very recently become popular in medical image analysis. Popularity of CNNs coincided with the later stages of the work reported here. Current CNN architectures generally require larger datasets than were available for this research, and work best with balanced label datasets (ours is unbalanced). You can use CNNs to partition ultrasound video, as described in recent preliminary research of our group (Gao et al., 2016), and other on-going research in our laboratory. The accuracy is slightly better. Going forward, it will be interesting to investigate whether CNN architectures can be designed to offer significant advantages over other methods for ultrasound video analysis.

In practice obstetricians may repeat an acquisition multiple times before they obtain satisfactory results. In this study, we have used only a single sweep. Initially the aim was to analyse what can be achieved from analysis of an extremely simple linear sweep (a minimal sweep). Given that the results are so promising, a logical next step is to extend the analysis to multiple sweeps which poses interesting research questions about how to fuse information obtained from multiple sweeps for clinical decision-making. This is the subject of some of our on-going work that we hope to report on in a future publication.

The data used in this study was obtained from the INTERGROWTH-21ST project (Sarris et al., 2013; Papageorghiou et al., 2014), which contains mothers at different gestational ages and with diverse body mass indices. Therefore the positive sub-sequences extracted from this data include a variety of representations of the fetal skull with different sizes and shadowing. This provides a set of rich features for the dataset of the positive sequences however a larger dataset of ultrasound sweeps might be required to build a robust classifier for more general populations.

Acknowledgements

M.A. Maraci acknowledges the support of RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation) and the Oxford EPSRC Impact Acceleration Award fund. C.P. Bridge support via a UK EPSRC DTA studentship. A. Noble acknowledgesthe support of the EPSRC Programme Grant Seebibyte. The authors also acknowledges that the data was acquired as part of the INTERGROWTH-21ST project. The authors gratefully acknowledge the help of Dr. Tess Norris, Dr. Sikolia Wanyonyi, Dr. Malid Molloholli, Dr. Christina Aye, and Miss Fenella Roseman, Research Midwife for acquiring the ultrasound data used in this project. All the in-vivo experiments were in accordance with the ethical standards of the institutional and national research committees.

Data statement

The images and image annotations used in this paper cannot be made freely available for reasons of ethical sensitivity. Data related to the tables will be made available from the Oxford University Research Archive (ORA-Data) on paper acceptance.

Appendix A. Breakdown of results

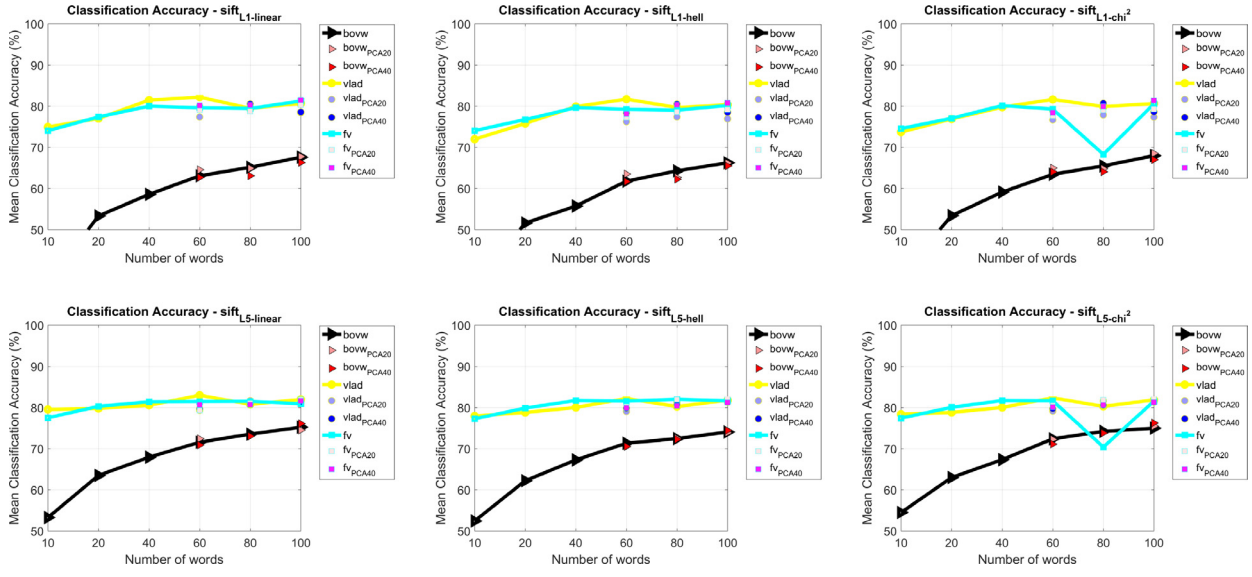


Fig. A.11. Mean classification accuracies for SIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

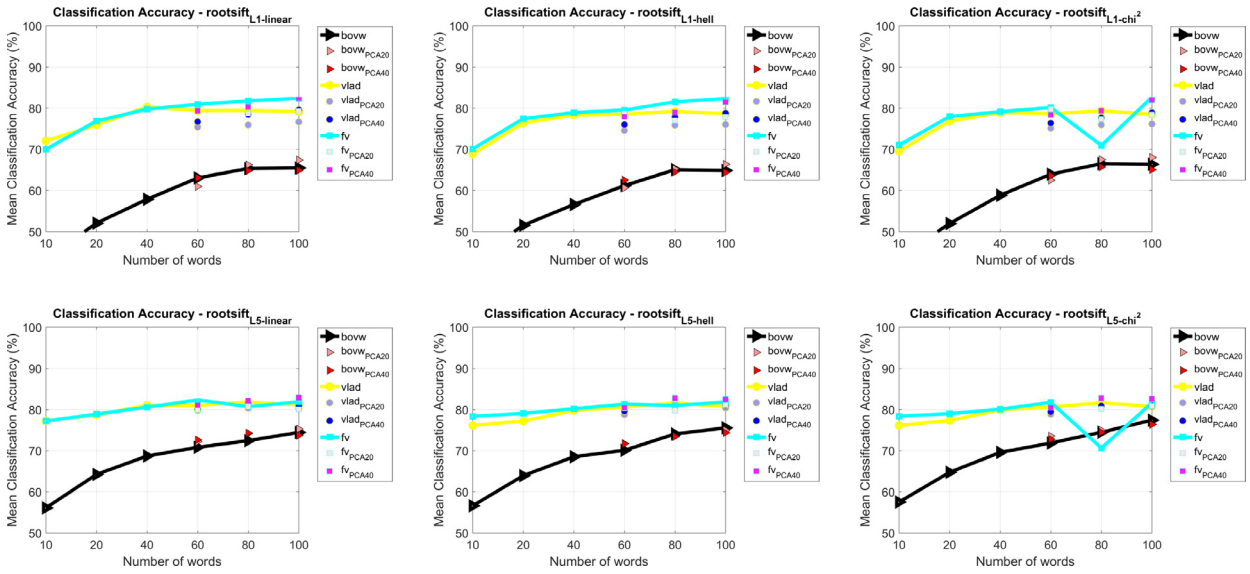


Fig. A.12. Mean classification accuracies for rootSIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

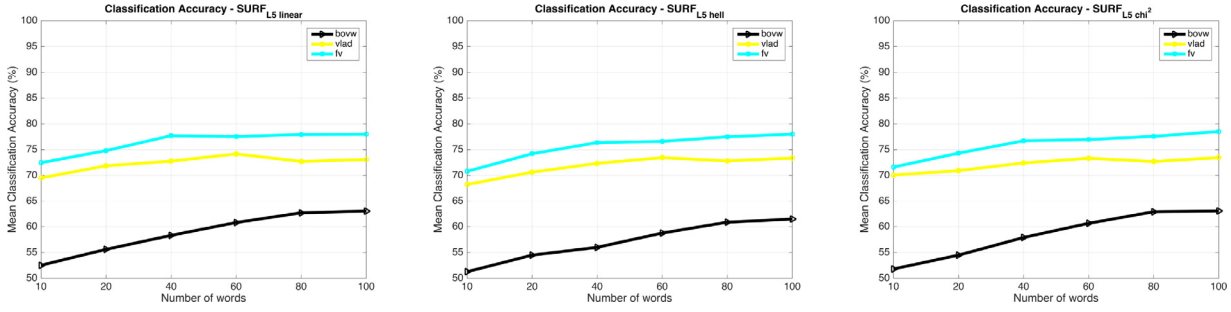


Fig. A.13. Mean classification accuracies for SURF feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

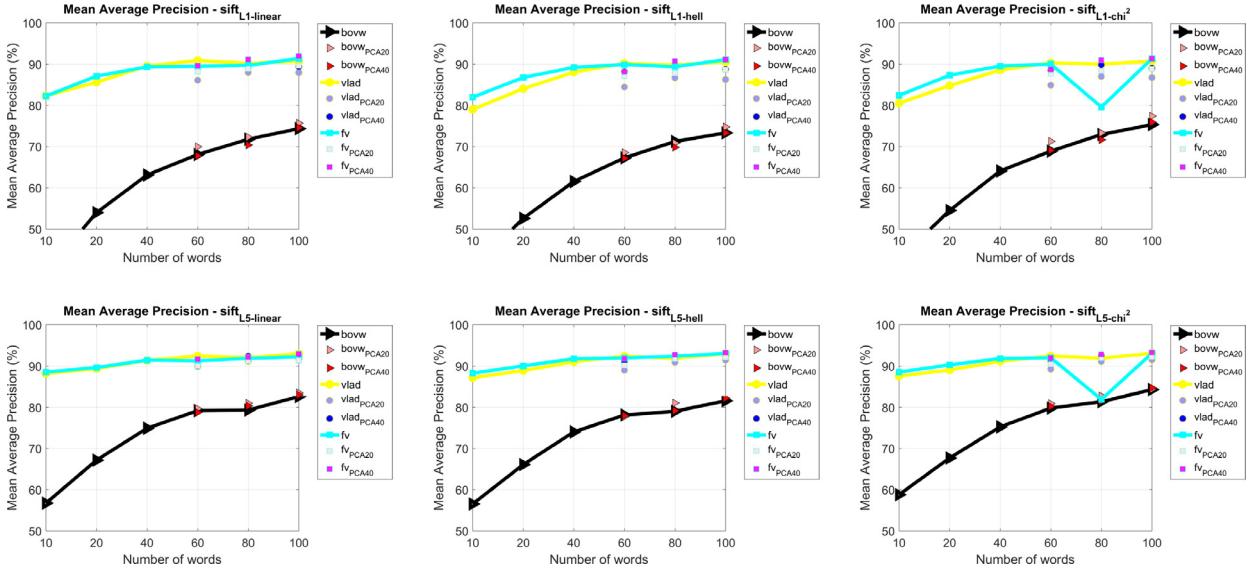


Fig. A.14. Mean average precision for SIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

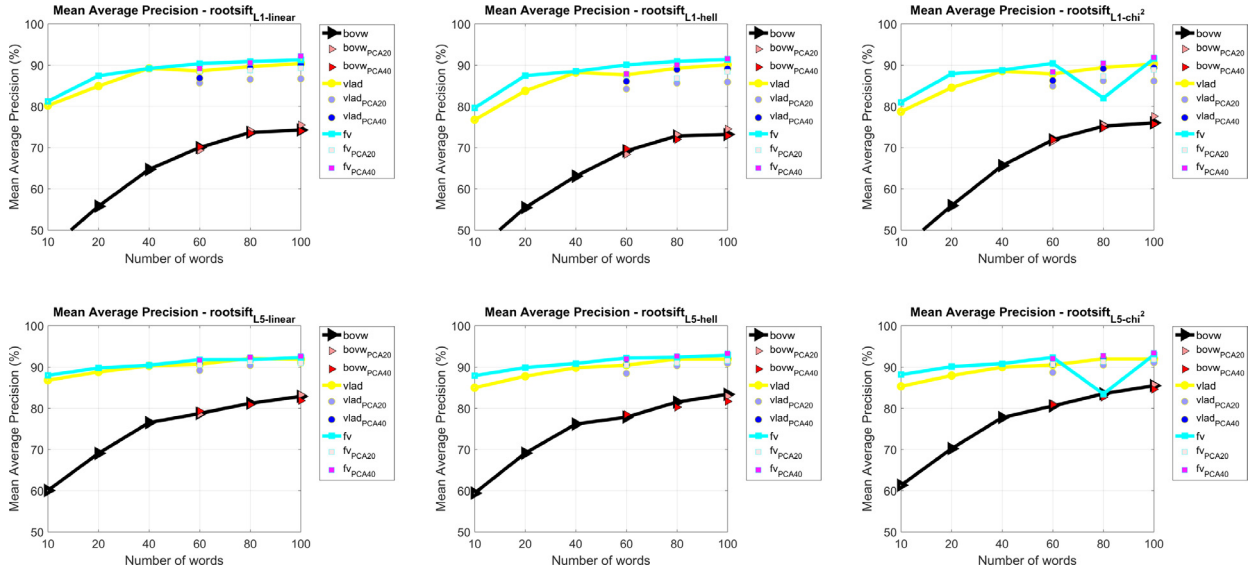


Fig. A.15. Mean average precision for rootSIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

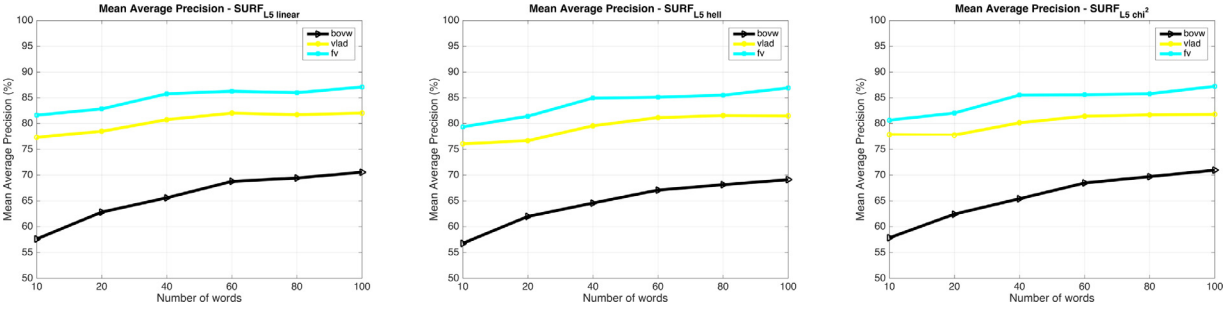


Fig. A.16. Mean average precision for rootSIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

References

Anto, E.A., Amoah, B., Crimi, A., 2015. Segmentation of ultrasound images of fetal anatomic structures using random forest for low-cost settings. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pp. 793–796. doi:10.1109/EMBC.2015.7318481.

Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Washington, DC, USA, pp. 2911–2918.

Bauer, S., Nolte, L.-P., Reyes, M., 2011. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 354–361. chapter Fully Automatic Segmentation of Brain Tumor Images Using Support Vector Machine Classification in Combination with Hierarchical Conditional Random Field Regularization. doi: 10.1007/978-3-642-23626-6_44.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Smith, S., Kainz, B., Rueckert, D., 2016. Real-Time Standard Scan Plane Detection and Localisation in Fetal Ultrasound Using Fully Convolutional Neural Networks. Springer International Publishing, Cham, pp. 203–211. doi:10.1007/978-3-319-46723-8_24.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: speeded up robust features. In: European conference on computer vision. Springer, pp. 404–417.

Bissacco, A., Chiuso, A., Soatto, S., 2007. Classification and recognition of dynamical models: the role of phase, independent components, kernels and optimal transport. Pattern Anal. Mach. Intell. IEEE Trans. 29 (11), 1958–1972.

Bridge, C., Ioannou, C., Noble, J., 2017. Automated annotation and quantitative description of ultrasound videos of the fetal heart. Medical Image Analysis 36, 147–161. doi:10.1016/j.media.2016.11.006.

Bridge, C.P., Noble, J.A., 2015. Object localisation in fetal ultrasound images using invariant features. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, pp. 156–159. doi:10.1109/ISBI.2015.7163839.

Brown, P.G., Alsousou, J., Cooper, A., Thompson, M.S., Noble, J.A., 2013. The autoqual ultrasound elastography method for quantitative assessment of lateral strain in post-rupture achilles tendons. J. Biomech. 46 (15), 2695–2700. doi:10.1016/j.jbiomech.2013.07.044.

Carneiro, G., Georgescu, B., Good, S., Comaniciu, D., 2008. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree.. IEEE Trans. Med. Imaging 27 (9), 1342–1355. doi:10.1109/TMI.2008.928917.

Chan, A.B., Vasconcelos, N., 2005. Probabilistic kernels for the classification of auto-regressive visual processes. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1. IEEE, pp. 846–851.

Chan, A.B., Vasconcelos, N., 2007. Classifying video with kernel dynamic textures. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, pp. 1–6.

Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A., 2011. The devil is in the details: an evaluation of recent feature encoding methods.. In: BMVC, 2, p. 8.

Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R., 2009. 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. IEEE, pp. 1932–1939.

Chaudhry, R., Vidal, R., 2009. Recognition of Visual Dynamical Processes: Theory, Kernels and Experimental Evaluation. Technical Report09–01.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P., 2015. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. Biomed Health Inf. IEEE J. 19 (5), 1627–1636. doi:10.1109/JBHI.2015.2425041.

Chykeyuk, K., Yaqub, M., Alison Noble, J., 2014. Class-specific regression random forest for accurate extraction of standard planes from 3d echocardiography. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Mller, H., Tu, Z. (Eds.), Medical Computer Vision. Large Data in Medical Imaging. In: Lecture Notes in Computer Science, 8331. Springer International Publishing, pp. 53–62. doi:10.1007/978-3-319-05530-5_6.

De Cock, K., De Moor, B., 2000. Subspace angles and distances between arma models. In: Proc. of the Intl. Symp. of Math. Theory of networks and systems, 1. Citeseer.

Doretto, G., Chiuso, A., Wu, Y.N., Soatto, S., 2003. Dynamic textures. Int. J. Comput. Vis. 51 (2), 91–109.

Forney Jr, G.D., 1973. The viterbi algorithm. Proc. IEEE 61 (3), 268–278.

Gao, Y., Maraci, M., Noble, J., 2016. Describing ultrasound video content using deep convolutional neural networks. Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on.

Goffinet, F., Carayol, M., Foidart, J.-M., Alexander, S., Uzan, S., Subtil, D., Brart, G., 2006. Is planned vaginal delivery for breech presentation at term still an option? results of an observational prospective survey in france and belgium. Am. J. Obstet. Gynecol. 194 (4), 1002–1011. doi:10.1016/j.ajog.2005.10.817.

Hannah, M.E., Hannah, W.J., Hewson, S.A., Hodnett, E.D., Saigal, S., Willan, A.R., 2000. 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.

Imaduddin, Z., Akbar, M.A., Tawakal, H., Satwika, P., Saroyo, Y., 2015. Automatic detection and measurement of fetal biometrics to determine the gestational age. In: Information and Communication Technology (ICoICT), 2015 3rd International Conference on, pp. 608–612. doi:10.1109/ICoICT.2015.7231495.

Jegou, H., Douze, M., Schmid, C., Perez, P., 2010. Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 3304–3311. doi:10.1109/CVPR.2010.5540039.

Joachims, T., 2006. Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 217–226.

Kadour, M.J., Adams, R., English, R., Parulekar, V., Christopher, S., Noble, J.A., 2010. Slip imaging: reducing ambiguity in breast lesion assessment. Ultrasound Med. Biol. 36 (12), 2027–2035.

Kadour, M.J., Noble, J.A., 2009. Assisted-freehand ultrasound elasticity imaging. Ultrason. Ferroelectr. Freq. ControlIEEE Trans. 56 (1), 36–43.

Kumar, A., Shriram, K., 2015. Automated scoring of fetal abdomen ultrasound scan–planes for biometry. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. IEEE, pp. 862–865.

Kwitt, R., Vasconcelos, N., Razzaque, S., Aylward, S., 2013. Localizing target structures in ultrasound video - a phantom study. Med. Image Anal. 17 (7), 712–722. doi:10.1016/j.media.2013.05.003. Special Issue on the 2012 Conference on Medical Image Computing and Computer Assisted Intervention

Lafferty, J. D., McCallum, A., Pereira, F. C. N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 282–289.

Lei, B., Tan, E.-L., Chen, S., Zhuo, L., Li, S., Ni, D., Wang, T., 2015. Automatic recognition of fetal facial standard plane in ultrasound image via fisher vector. PLoS ONE 10 (5), e0121838. doi:10.1371/journal.pone.0121838.

Liu, K., Skibbe, H., et al., 2014. Rotation-invariant HOG descriptors using fourier analysis in polar and spherical coordinates. IJCV 106 (3), 342–364.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60 (2), 91–110.

Maraci, M., Napolitano, R., Papageorghiou, A., Noble, J., 2015. Fisher vector encoding for detecting objects of interest in ultrasound videos. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. IEEE, pp. 651–654.

Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A., 2014. Object classification in an ultrasound video using lp-sift features. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Mller, H., Zhang, S., Cai, W.T., Metaxas, D. (Eds.), Medical Computer Vision: Algorithms for Big Data. In: Lecture Notes in Computer Science, 8848. Springer International Publishing, pp. 71–81. doi:10.1007/978-3-319-13972-2_7.

Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A., 2014. Searching for structures of interest in an ultrasound video sequence. In: Wu, G., Zhang, D., Zhou, L. (Eds.), Machine Learning in Medical Imaging. In: Lecture Notes in Computer Science, 8679. Springer International Publishing, pp. 133–140. doi:10.1007/978-3-319-10581-9_17.

McIntosh, C., Svistoun, I., Purdie, T.G., 2013. Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning. IEEE Trans. Med. Imaging 32 (6), 1043–1057. doi:10.1109/TMI.2013.2251421.

Namburete, A.I., Stebbing, R.V., Kemp, B., Yaqub, M., Papageorghiou, A.T., Noble, J.A., 2015. Learning-based prediction of gestational age from ultrasound images of the fetal brain. Med. Image Anal. 21 (1), 72–86.

Noble, J.A., 2016. Reflections on ultrasound image analysis. Med. Image Anal. 33, 33–37. 20th anniversary of the Medical Image Analysis journal (MedIA). doi:10.1016/j.media.2016.06.015

Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., Kohli, P., 2011. Decision tree fields. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 1668–1675.

Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., Noble, J.A., Pang, R., Victora, C.G., Barros, F.C., Carvalho, M., Salomon, L.J., Bhutta, Z.A., Kennedy, S.H., Villar, J., 2014. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. Lancet 384 (9946), 869–879. doi:10.1016/S0140-6736(14)61490-2.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification. In: ECCV. Springer Berlin Heidelberg, pp. 143–156.

Ponomarev, G., Gelfand, M., Kazanov, M., 2012. A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images. In: Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012, pp. 17–19.

Rahmatullah, B., Papageorghiou, A., Noble, J., 2011. Automated selection of standardized planes from ultrasound volume. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (Eds.), Machine Learning in Medical Imaging. In: Lecture Notes in Computer Science, 7009. Springer Berlin/Heidelberg, pp. 35–42. doi: 10.1007/978-3-642-24319-6_5

Rahmatullah, B., Sarris, I., Papageorghiou, A., Noble, J.A., 2011. Quality control of fetal ultrasound images: detection of abdomen anatomical landmarks using adaboost. In: Proc. IEEE Int Biomedical Imaging: From Nano to Macro Symp, pp. 6–9.

Rajpoot, K., Grau, V., Noble, J., 2009. 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, pp. 783–786. doi:10.1109/ISBI.2009.5193166.

Rijken, M.J., Lee, S.J., Boel, M.E., Papageorghiou, A.T., Visser, G.H.A., Dwell, S.L.M., Kennedy, S.H., Singhasivanon, P., White, N.J., Nosten, F., McGready, R., 2009. Obstetric ultrasound scanning by local health workers in a refugee camp on the thaiburmese border. Ultrasound Obstetrics Gynecol. 34 (4), 395–403. doi:10.1002/uog.7350.

Rueda, S., Fathima, S., Knight, C., Yaqub, M., Papageorghiou, A., Rahmatullah, B., Foi, A., Maggioni, M., Pepe, A., Tohka, J., Stebbing, R., McManigle, J., Ciurte, A., Bresson, X., Cuadra, M., Sun, C., Ponomarev, G., Gelfand, M., Kazanov, M., wei Wang, C., Chen, H.-C., Peng, C.-W., Hung, C.-M., Noble, J., 2014. Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: A grand challenge. Med. Imaging IEEE Trans. 33 (4), 797–813. doi:10.1109/TMI.2013.2276943.

Salomon, L.J., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S.L., Kalache, K., Leung, K.-Y., Malinger, G., Munoz, H., Prefumo, F., Toi, A., Lee, W., on behalf of the ISUOG Clinical Standards Committee, 2011. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstetrics Gynecol. 37 (1), 116–126. doi:10.1002/uog.8831.

Sarris, I., Ioannou, C., Dighe, M., Mitidieri, A., Oberto, M., Qingqing, W., Shah, J., Sohoni, S., Al Zidjali, W., Hoch, L., Altman, D.G., Papageorghiou, A.T., I. F., for the 21st Century, N.G.C., 2011. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements.. Ultrasound Obstetrics Gynecol. 38 (6), 681–687.

Sarris, I., Ioannou, C., Ohuma, E., Altman, D., Hoch, L., Cosgrove, C., Fathima, S., Salomon, L., Papageorghiou, A., for the International Fetal, for the 21st Century (INTERGROWTH-21st), N.G.C., 2013. Standardisation and quality control of ultrasound measurements taken in the intergrowth-21st project. BJOG: Int. J. Obstetrics Gynaecol. 120, 33–37. doi:10.1111/1471-0528.12315.

Sun, C., 2012. Automatic fetal head measurements from ultrasound images using circular shortest paths. In: Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012, pp. 13–15.

Tiran, D., 2005. Nice guideline on antenatal care: routine care for the healthy pregnant womanrecommendations on the use of complementary therapies do not promote clinical excellence. Complementary Ther. Clin. Pract. 11 (2), 127–129.

Villar, J., Ismail, L.C., Victora, C.G., Ohuma, E.O., Bertino, E., Altman, D.G., Lambert, A., Papageorghiou, A.T., Carvalho, M., Jaffer, Y.A., Gravett, M.G., Purwar, M., Frederick, I.O., Noble, A.J., Pang, R., Barros, F.C., Chumlea, C., Bhutta, Z.A., Kennedy, S.H., 2014. International standards for newborn weight, length, and head circumference by gestational age and sex: the newborn cross-sectional study of the intergrowth-21st project. Lancet 384 (9946), 857–868. doi:10.1016/S0140-6736(14)60932-6.

Vishwanathan, S.V., Smola, A.J., Vidal, R., 2007. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. Int. J. Comput. Vision 73 (1), 95–119. doi:10.1007/s11263-006-9352-0.

Wanyonyi, S.Z., Napolitano, R., Ohuma, E.O., Salomon, L.J., Papageorghiou, A.T., 2014. Image-scoring system for crownrump length measurement. Ultrasound Obstetrics Gynecol. 44 (6), 649–654. doi:10.1002/uog.13376.

Yaqub, M., Javaid, M.K., Cooper, C., Noble, J.A., 2011. Improving the Classification Accuracy of the Classic Rf Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation. In: Machine Learning in Medical Imaging. Springer, pp. 184–192.

Yaqub, M., Napolitano, R., Ioannou, C., Papageorghiou, A.T., Noble, J.A., 2012. Automatic detection of local fetal brain structures in ultrasound images. In: Proc. 9th IEEE Int Biomedical Imaging (ISBI) Symp, pp. 1555–1558.

Yaqub, M., Rueda, S., Kopuri, A., Melo, P., Papageorghiou, A., Sullivan, P., McCormick, K., Noble, J., 2015. Plane localization in 3d fetal neurosonography for longitudinal analysis of the developing brain. Biomed. Health Inf. IEEE J. PP (99). 1–1. doi: 10.1109/JBHI.2015.2435651.

Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C., 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vision 73 (2), 213–238.