Automated Annotation and Quantitative Description of Ultrasound Videos of the Fetal Heart

Christopher P. Bridge[a,*], Christos Ioannou[b], J. Alison Noble[a]

[a] *Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom*
[b] *Fetal Medicine Unit, John Radcliffe Hospital, Oxford, United Kingdom*

Abstract

Interpretation of ultrasound videos of the fetal heart is crucial for the antenatal diagnosis of congenital heart disease (CHD). We believe that automated image analysis techniques could make an important contribution towards improving CHD detection rates. However, to our knowledge, no previous work has been done in this area. With this goal in mind, this paper presents a framework for tracking the key variables that describe the content of each frame of freehand 2D ultrasound scanning videos of the healthy fetal heart. This represents an important first step towards developing tools that can assist with CHD detection in abnormal cases. We argue that it is natural to approach this as a *sequential Bayesian filtering* problem, due to the strong prior model we have of the underlying anatomy, and the ambiguity of the appearance of structures in ultrasound images. We train classification and regression forests to predict the visibility, location and orientation of the fetal heart in the image, and the viewing plane label from each frame. We also develop a novel adaptation of regression forests for circular variables to deal with the prediction of cardiac phase. Using a particle-filtering-based method to combine predictions from multiple video frames, we demonstrate how to filter this information to give a temporally consistent output at real-time speeds. We present results on a challenging dataset gathered in a real-world clinical setting and compare to expert annotations, achieving similar levels of accuracy to the levels of inter- and intra-observer variation.

Keywords: ultrasound, fetal, heart, cardiac, view detection, rotation-invariant, random forests, particle filter

1. Introduction

Congenital heart disease (CHD) is one of the most common defects affecting infants at birth and covers a range of specific issues that affect the normal function of the heart. The established method for *in utero* detection of CHD is antenatal ultrasound screening of the fetal heart. Typical screening procedures are conducted at a gestational age of 18-22 weeks and involve the use of a two dimensional (2D) ultrasound transducer to examine visually the development and function of the different structures (Carvalho et al. (2013)). Unfortunately, detection rates of CHD vary widely due to a number of different factors including the training of the sonographer (Pézard et al. (2008); Allan (2000)), the nature of the defect, and the affluence of the region (Hill et al. (2015)).

A recent survey in the United States of America suggested that one of the key factors that limits the diagnosis rate is that many forms of CHD cannot be identified from a four-chamber view alone (Hill et al. (2015)). Recent guidelines (Carvalho et al. (2013)) have also emphasised the importance of using a number of different *viewing planes*, in addition to the common four-chamber view, in order to increase the rate of diagnosis of certain types of CHD.

Analysis of clinical fetal cardiac ultrasound videos is a challenging task, even for humans, for a number of reasons. Firstly, the indistinct appearance of anatomical structures in ultrasound images makes image interpretation difficult. This is compounded by variations in contrast levels and imaging parameters, as well as the presence of imaging artefacts such as speckle, shadowing and enhancement. In fetal cardiac videos (unlike adult echocardiography), the heart may take up only a small fraction of the screen and its location in the image can change due to motion of the probe and/or the fetus during scanning. The orientation of the fetus relative to the direction of the propagation of sound is also unknown and potentially variable. The appearance of the heart changes significantly throughout the cardiac cycle, and there may also be fetal motion in the direction perpendicular to the imaging plane that may cause the appearance to change or cause the heart to disappear altogether. Furthermore, while scanning, a sonographer will often review the different viewing planes of the fetal heart in relatively quick succession.

*Corresponding author
Email addresses: christopher.bridge@eng.ox.ac.uk (Christopher P. Bridge), christos.ioannou@ouh.nhs.uk (Christos Ioannou), alison.noble@eng.ox.ac.uk (J. Alison Noble)

Computer-aided methods have the potential to improve detection rates of CHD, but little previous work has been carried out towards this aim. The focus of this paper is the general problem of automatically estimating key information of interest from videos of the healthy fetal heart during acquisition within a standard screening scan. This represents a critical first step in an image processing pipeline and could support good-quality acquisition and assist an operator in interpretation. Furthermore it provides a basis for further work towards automatic quantification and diagnosis of abnormal hearts.

In order to have a thorough and useful description of the state of the healthy heart at a given point in time, we must estimate the key parameters including its visibility, position and orientation in the image as well as the current viewing plane and the position in the cardiac cycle. Our approach is to pose this problem as an inference problem using *sequential Bayesian filtering*. There are a number of reasons for this choice. Sequential Bayesian filtering techniques allow a probabilistic belief over the 'state' of a 'system' (in our case the heart is the system and the state is its position, orientation, viewing plane and cardiac cycle position) to be updated on-line – and often in real-time – using all the observations that have been made so far. In particular, they naturally account for the uncertainty in individual observations made from the images, and balance them against a prior model of how the 'system' behaves in order to enforce temporal consistency. This is particularly important in this setting, where the information in each frame is often relatively weak or ambiguous due to the difficulty in interpreting ultrasonic reflection patterns, while the temporal model of heart behaviour over a number of frames is comparatively strong.

The outline of the remainder of the paper is as follows. Having reviewed related literature in §2, we formally define our problem in §3 and outline our proposed model in §4, with key components described in §5 and §6. In §7 we describe the evaluation of the model on a dataset of fetal heart videos captured in a clinical setting. We present results in §8 and concluding remarks in §9.

2. Related Work

To the best of our knowledge, this is the first work to attempt to automate analysis of fetal cardiac ultrasound videos. Previous authors have successfully performed view detection in images obtained from *adult* echocardiographic images using a variety of techniques (Agarwal et al. (2013); Wu et al. (2013); Zhou et al. (2006); Park et al. (2007); Qian et al. (2013); Kumar et al. (2009); Ebadollahi et al. (2004)), while others have had success in automatic recognition of other fetal structures in images (Carneiro et al. (2008); Rahmatullah et al. (2012); Namburete et al. (2013); Yaqub et al. (2012)) and, more recently, in videos (Maraci et al. (2014); Chen et al. (2015)). Finally, some work has attempted to estimate a more detailed description of the adult heart in echocardiographic data in the form of boundaries (Nascimento and Marques (2008); Carneiro and Nascimento (2013); Yang et al. (2008)).

2.1. View Detection in Adult Echocardiagraphy

Several approaches to view detection in adult echocardiography make use of global image properties in order to deduce the view label. For example, Agarwal et al. (2013) use a histogram of oriented gradients (HOG) descriptor on the whole image, broken into four non-overlapping blocks. This can distinguish between two very different views (long axis and short axis) with a support vector machine (SVM) classifier. Wu et al. (2013) employ a similar method, using 'GIST' descriptors (Oliva and Torralba (2001)) in 16 image blocks instead of HOG descriptors. Zhou et al. (2006) use a multi-class classifier based on LogitBoost and rectangular filters ('Haar-like' filters) in order to distinguish between apical two-chamber and four-chamber views. Such global methods are not well-suited to fetal echocardiography because they assume a relatively consistent layout of frames, but in fetal imagery the position and orientation of the heart is unknown. Also, in our application, only small areas of the fetal images are relevant to view classification, and the rest of the image is taken up by the fetal abdomen and the womb.

This is overcome, to some extent, in the work of Park et al. (2007), which builds on the work in Zhou et al. (2006) by adding a left ventricle detection stage, which is then used to position the multi-class view classifier in the image. However, this relies upon the appearance of the left ventricle being fairly consistent between views, and there is unfortunately no such guarantee of consistency in the fetal views of interest to us. Furthermore, although it solves the problem of unknown position it does not solve the problem of unknown orientation.

Other methods rely on first detecting keypoints in the frame. Qian et al. (2013) detect space-time interest points in the video stream and describe them using a 3D scale-invariant feature tranform (SIFT) descriptor (in the two spatial dimensions plus time). Similarly, Kumar et al. (2009) detect interest points using the SIFT keypoint detector in the motion magnitude image, and describe them using local histograms of motion magnitude and intensity. In both cases, the extracted descriptors are quantised according to a pre-trained codebook, and an SVM classifier is used on the codebook histogram for classification. Such approaches are also unlikely to be effective in fetal imagery for the same reasons as the global methods. It is also difficult to estimate other information such as position, orientation and cardiac phase information from the frames using this approach.

Ebadollahi et al. (2004) first use the grey-scale symmetric axis transform (GSAT) to detect the "blobs" that are potential heart chambers. They then connect them in a Markov Random Field (MRF) graph structure in order to label the chambers and hence deduce the view label. This approach depends on reliable detection of chambers, and the authors showed that accuracy dropped dramatically when chamber detection was not reliable, as is likely to be the case in fetal imaging where structures other than the heart are visible.

2.2. Structure Detection in Fetal Ultrasound Imagery

Several authors have used ensemble methods that combine weak classifiers based on rectangular block filters to detect particular structures in fetal ultrasound imagery. For example Carneiro et al. (2008) used a Probabilistic Boosting Tree and rectangular filters to detect a number of different structures including the fetal head, abdomen and femur. Rahmatullah et al. (2012) and Namburete et al. (2013) used similar features and an Adaboost classifier to detect abdominal and cerebral landmarks in fetal images, and Yaqub et al. (2012) used random forest classifiers with rectangular filters for cerebral structures. We draw on these works by using random decision forests for detection of and discrimination between the different fetal heart views. However since rectangular block filters do not deal well with unknown orientations, we have instead chosen to use a alternative set of rotation invariant features (see §6.1).

One approach to fetal ultrasound *video* analysis is that of Maraci et al. (2014), who model the frames in short video sequences as the output of a linear dynamical system, and construct a SVM classifier based on kernels between the model parameters in order to detect subsequences containing structures of interest. This method provides a general method for exploiting the information contained within motion patterns for detection. However it detects structures in time but not space and is not well suited to on-line applications as the complete sequence is needed to deduce model parameters.

Perhaps the work with the most similar aims to ours is that of Chen et al. (2015), who use a deep architecture that combines a spatial convolutional neural network and a temporal recurrent neural network to make use of temporal context features for standard viewing plane detection in fetal ultrasound videos. Unfortunately, that approach requires a large amount of training data and the technique has not yet been used for full state tracking in the sense that we are attempting here. Gao et al. (2016) have recently demonstrated that the data requirements for using deep networks with fetal ultrasound can be reduced by using transfer learning from models trained on natural images. However, neither of these papers are specifically dealing with the fetal heart.

2.3. Boundary Tracking in Adult Echocardiagraphy

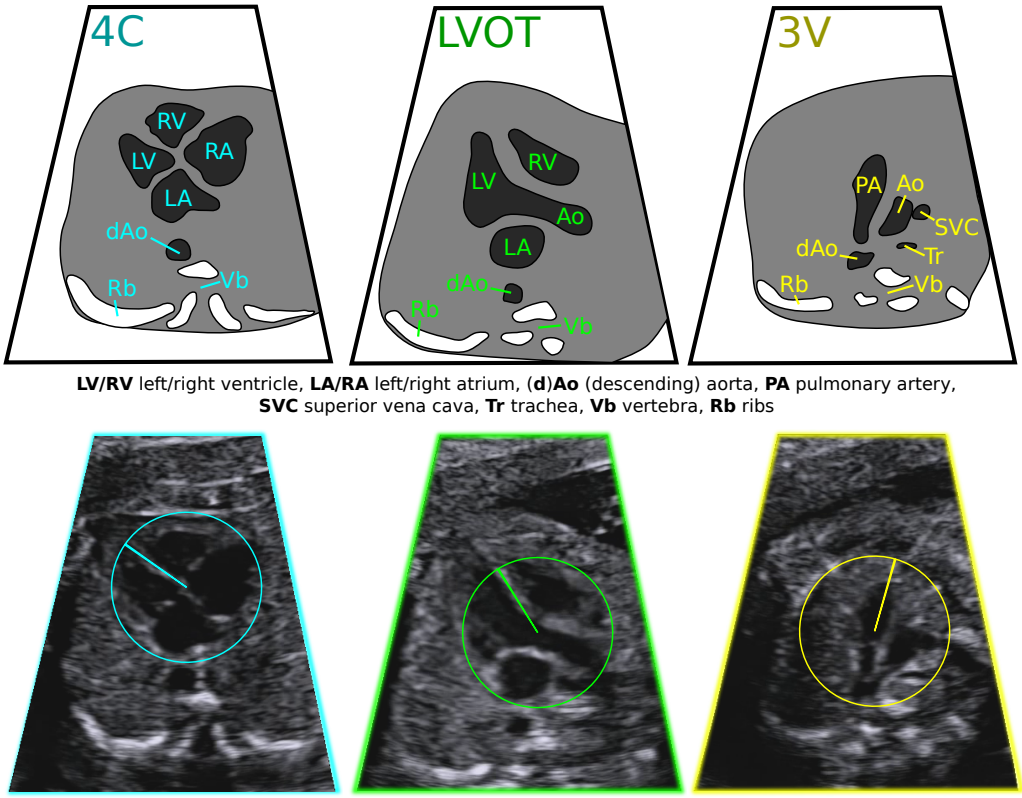
Another area of related work is automatic boundary tracking in (adult) echocardiography using 2D (e.g. Jacob et al. (1998); Nascimento and Marques (2008); Carneiro and Nascimento (2013)) or 3D (e.g. Yang et al. (2008)) video data. Like our work, these algorithms track a high-dimensional representation of the heart as it evolves through video frames, and like in this paper, they tend to use a strong temporal prior model in order to provide robustness to ambiguous image information. For example, in early work Jacob et al. (1998) used a Kalman filter to model the evolution of the left ventricular boundary. Nascimento and Marques (2008) built on this with multiple predictive models and robust data association to eliminate erroneous boundary candidates. Carneiro and Nascimento (2013) track points on the left ventricle endocardium using a robust particle filtering framework that couples a linear transition model (in fact one model for diastole and another for systole) with an observation model built with deep neural networks. Such techniques are also applicable for the higher dimensional problem of 3D boundary tracking, such as the work of Yang et al. (2008), which uses a prediction model based on manifold learning of left ventricle boundary trajectories, and combines it with an observation model using probabilistic boosting trees.

Whilst the methodologies in these papers are related to our work, their aims are somewhat different from ours as they specifically aim to track the ventricle boundary, and assume carefully captured data that reliably contains the boundary of interest and in which there are no changes in viewing plane or significant changes in heart location. Our aim is to provide a more broadly applicable set of measurements and descriptions of fetal heart scans, that could could provide useful information in less constrained scanning sessions.

3. Problem Definition

We formulate our problem within the framework of Bayesian filtering. We therefore have an unobserved *state*, $\mathbf{s}_t$, at time t that contains variables describing the visibility of the heart, the location of the heart centre in the image, the current view category label, the orientation of the heart in the image, and the current cardiac phase. We wish to estimate this state from image data at test time.

To demonstrate our approach we use a slight simplification of the viewing plane taxonomy that is recommended for visualisation during a fetal cardiac assessment (Carvalho et al. (2013)) and define three viewing plane labels: the *four chamber* (4C) view, the *left ventricular outflow tract* (LVOT) view, and the *three vessels* (3V) view. This gives a discrete, categorical view label variable $v_t \in \{4\text{C}, \text{LVOT}, 3\text{V}\}$. The definitions of the location of the heart centre $\mathbf{x}_t \in \mathbb{R}^2$, in



LV/RV left/right ventricle, **LA/RA** left/right atrium, (**d**)**Ao** (descending) aorta, **PA** pulmonary artery, **SVC** superior vena cava, **Tr** trachea, **Vb** vertebra, **Rb** ribs

Figure 1: Definition of the three viewing planes and their annotations. *Top row* schematics showing the anatomic structures visible within the fetal abdomen in each view. *Bottom row* example image and annotation. The colour scheme introduced in this figure will be used throughout the article (*cyan* four-chamber (4C) view, *green* left ventricular outflow tract (LVOT) view, *yellow* three vessels (3V) view).

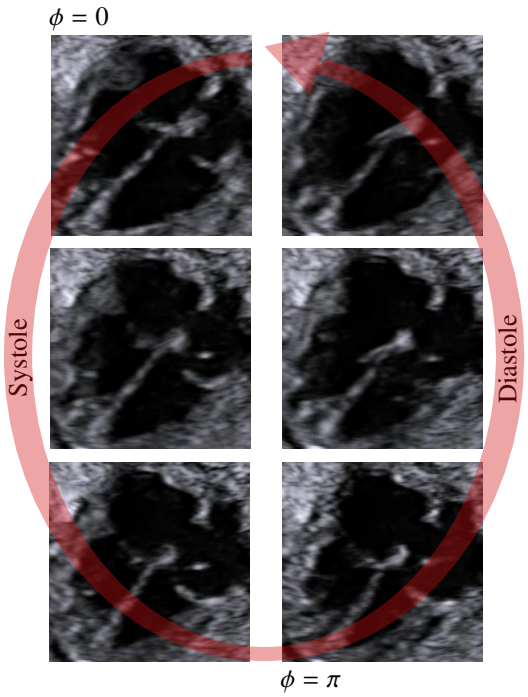


Figure 2: 2D ultrasound images of the fetal heart (four-chamber view). The six images appear at different points in the cardiac cycle. We represent cardiac phase by a circular variable in the range 0 to 2π.

pixels, and the heart orientation $\theta_t \in [0, 2\pi)$, defined anti-clockwise from the increasing x-direction, differ for each of the three views.

Definitions of these views and their coordinate systems are shown in Fig. 1. The *four chamber* view contains all four chambers, with the centre at the crux. The orientation is defined by the orientation of the interventricular septum. The radius is defined in this view as that which encompasses both atria. The *left ventricular* (aortic) *outflow tract* view is defined by the presence of the aorta leaving the left ventricle. The centre is defined by the centre of the aorta where it crosses the interventricular septum and the orientation is again defined by that of the interventricular septum. The *three vessels* view is defined by the simultaneous presence of the pulmonary artery, aorta and superior vena cava. The centre is defined as the centre of the pulmonary artery at the point where it is in line with the other two vessels, and the orientation is defined by that of the right wall of the pulmonary artery. These three planes can be viewed in sequence by sweeping the probe in a cephalad direction from the four-chamber view.

The cardiac cycle is described by a *cardiac phase* variable $\phi_t \in [0, 2\pi)$ in rad, where $\phi_t = 0$ denotes end-diastole and $\phi_t = \pi$ denotes end-systole, and other values are interpolated linearly between these key points (see Fig. 2). Because the heart rate in the videos is unknown but relatively constant, we find that using a second-order model for the cardiac phase is advantageous. Therefore the state vector also contains the current *cardiac phase rate* $\dot{\phi}_t \in [\dot{\phi}_{\min}, \dot{\phi}_{\max}]$ in rad s^{-1} (i.e. the rate of change of the cardiac phase variable with respect to time), where hard limits are placed on the permissible values for the phase rate to avoid temporal aliasing and other unexpected behaviour. By contrast, we find that using second-order models for position and orientation is unnecessary because changes in these values are small.

Finally we track the visibility of the heart with a Boolean variable $h_t \in \{0, 1\}$, which represents whether the heart is currently *visible* (0) or *hidden* (1). The intention is to allow the algorithm to cope with frames where the heart is not visible or heavily obscured due to imaging artefacts or slight misalignment of the probe, rather than gross misalignment of the probe. We therefore assume that when the heart is hidden during the scanning process, it makes sense to continue to track the other state variables because the heart will soon become visible again in a similar state to that in which it was last observed.

If the gestational age of the fetus and magnification factor of the ultrasound system are known, the size of the heart in the image is relatively well-constrained and could be estimated from fetal growth chart, for example (Kim et al. (1992)). For this reason, we choose to assume that the fetal heart size (radius r) is known to the algorithm at test time. However, in principle, the heart size could be incorporated into the state vector as well.

The six state variables are grouped together to form the *state vector*, $\mathbf{s}_t$, of the system, which we estimate on-line from unseen videos.

$$\mathbf{s}_t = \begin{bmatrix} h_t \\ v_t \\ \mathbf{x}_t \\ \theta_t \\ \phi_t \\ \dot{\phi}_t \end{bmatrix} \tag{1}$$

Our aim is to predict the state vector $\mathbf{s}_t$ at time t, using all the *image information* $\mathbf{z}_{0:t}$ available up to this point. This is a *filtering* problem, and the corresponding posterior distribution, $p(\mathbf{s}_t \mid \mathbf{z}_{0:t})$, is known as the *filtering distribution*.

Unfortunately, there is inherent ambiguity in many of the variables that we are trying to estimate, which limits the accuracy it is possible to achieve. For example, the categorisation of the different viewing planes is not clear in some cases, and the cardiac phase is difficult to measure with a high degree of accuracy from video data alone.

4. Proposed Model Definition

In general terms, the sequential Bayesian filtering problem is solved by applying the recursive Bayesian filtering equations (Doucet et al. (2001); Thrun et al. (2005)). However, exact application of these equations is only possible in a few restrictive special cases. One such case is when the variables that make up the state are discrete, in which case the problem reduces to the well-known *hidden Markov model*. It is possible to discretise a continuous state onto a finite grid in order to create a discrete state, and then solve the resulting discrete problem. However, in our case a fine grid over each continuous dimension of the state space would be needed to give useful results. This would result in a very large number of discrete states, meaning that the resulting filter would likely be inefficient.

A second common case where exact inference is possible is where the state variables are continuous, the state transition model is linear, and the distributions over the observed and state variables are all Gaussian. The resulting algorithm is the *Kalman filter*. Unfortunately, the Kalman filter would not be able track our state, since it consists of a combination of real-valued variables (position, cardiac phase rate), discrete-valued variables (visibility, viewing plane label), and circular variables (orientation, cardiac phase). Furthermore, the assumption of Gaussian likelihood distributions is restrictive and more complex models are needed to cope with the challenging task of recognising the patterns found in fetal ultrasound images.

However, it is possible to relax many of these restrictions if one is prepared to accept approximate inference methods in place of exact inference. Fortunately, excellent results can be achieved in practice using approximate methods. We therefore turn to *particle filters*, which have become an established method in computer vision for a number of recursive estimation problems due to being effective, efficient and highly flexible (Doucet et al. (2001)). A particle filter is a stochastic model that approximates the distribution over the state at each time point with a large set of weighted samples ('particles') drawn from it. At each time step, the particles evolve in the *prediction step* according to a prior model of the system's behaviour and are then re-weighted and resampled in an *update step* according to some observation model and the newly observed data.

The standard particle filtering algorithm assumes a *generative* model for the observations $\mathbf{z}_t$ given the current state $\mathbf{s}_t$, and hence the distribution over the unobserved state given the observation is implicitly modelled via Bayes' rule. However, using generative models involves unnecessary modelling of the joint probability distribution and in practice limits the flexibility of the models that can be used. We therefore choose to use the conditional random field filter (CRF-filter) introduced by Limketkai et al. (2007), which is a simple modification of the standard particle filter where general prediction and observation potential functions model the interactions between variables in the cliques of an undirected graphical model. The graphical structure of the CRF-filter is shown in Fig. 3. Intuitively, the *prediction potential*, $\psi_p(\mathbf{s}_t, \mathbf{s}_{t-1})$, measures the compatibility of the current state and previous state, and the *observation potential*, $\psi_o(\mathbf{z}_t, \mathbf{s}_t)$, measures the compatibility of the current state and the current observation. In order to make use of the particle filtering paradigm, it is necessary to be able to sample from the prediction potential. However, the observation potential function can in principle be any non-negative function of its arguments, which affords us far greater modelling flexibility. This allows us to define complex, *discriminative* observation potentials using the random forests algorithm.

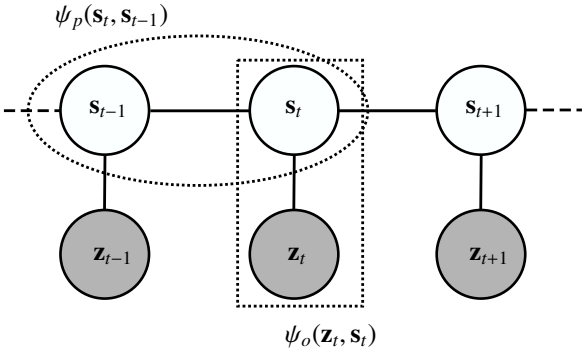


Figure 3: Graphical structure of the CRF-filter model. At each timestep, t, there is a node representing the *state* $\mathbf{s}_t$ and a node representing the *observed image* $\mathbf{z}_t$. A prediction potential $\psi_p(\mathbf{s}_t, \mathbf{s}_{t-1})$ measures the compatibility of successive states and an observation potential $\psi_o(\mathbf{z}_t, \mathbf{s}_t)$ measures the compatibility of a state value and the image evidence.

The particle set in our algorithm is updated at each time step using the procedure summarised in Algorithm 1. In the following two sections we describe the two key remaining parts of our model: §5 describes the prediction model used to update the particles, and §6 describes the specification of the observation potentials used to re-weight the particles.

5. State Evolution Model

5.1. State Update

Recall from §4 that the particle filtering state-update step takes each particle and stochastically updates it according to a prediction potential function at each time step. Because of the need to sample from the prediction potential function, it is realised as a true conditional probability distribution, i.e. $\psi_p(\mathbf{s}_t, \mathbf{s}_{t-1}) = p(\mathbf{s}_t \mid \mathbf{s}_{t-1})$. To simplify the model, we assume that the changes in several (but not all) of the state variables are

Input: a set of N_P particles $\mathbf{s}_{t-1}^{(i)}$ with associated weights $w_{t-1}^{(i)}$, $i = 0, \ldots, N_P - 1$, the observed image $\mathbf{z}_t$
Output: a new set of N_P particles $\mathbf{s}_t^{(i)}$ with associated weights $w_t^{(i)}$, $i = 0, \ldots, N_P - 1$

$N_{eff} \leftarrow \frac{1}{\sum_{i=0}^{N_P-1}\left(w_{t-1}^{(i)}\right)^2}$ {*calculate the effective number of particles*}
if $(N_{eff} < N_{thresh})$ **then** {*if effective sample size is low*}
  for all particles i in the set $i = 0, \ldots, N_P - 1$ **do**
    sample $j_i \sim P(j_i = k) = w_{t-1}^{(k)}$ {*choose new particle index according to the particle weights*}
  end for
  for all particles i in the set $i = 0, \ldots, N_P - 1$ **do**
    $\mathbf{s}_{t-1}^{(i)} \leftarrow \mathbf{s}_{t-1}^{(j_i)}$ {update resampled particle}
    $w_{t-1}^{(i)} \leftarrow \frac{1}{N_p}$ {reset weights}
  end for
end if
for all particles i in the set $i = 0, \ldots, N_P - 1$ **do**
  sample $\mathbf{s}_t^{(i)} \sim \psi_p(\mathbf{s}_t^{(i)}, \mathbf{s}_{t-1}^{(i)})$ {*state update according to Algorithm 2*}
  $w_t^{(i)} \leftarrow w_{t-1}^{(i)} \cdot \psi_o(\mathbf{z}_t^{(i)}, \mathbf{s}_t^{(i)})$ {*re-weight the particles (see §6)*}
end for
for all particles i in the set $i = 0, \ldots, N_P - 1$ **do**
  $w_t^{(i)} \leftarrow \frac{w_t^{(i)}}{\sum_{j=0}^{N_P-1}\left(w_{t-1}^{(j)}\right)}$ {*re-normalise the particle weights*}
end for

Algorithm 1: A single step of the particle filtering algorithm

independent of each other so that the distribution can be decomposed as follows:

$$\begin{aligned} p(\mathbf{s}_t \mid \mathbf{s}_{t-1}) = {} & p(h_t \mid h_{t-1}) \times \\ & p(v_t \mid v_{t-1}) \times \\ & p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \theta_{t-1}, v_t, v_{t-1}) \times \\ & p(\theta_t \mid \theta_{t-1}, v_t, v_{t-1}) \times \\ & p(\phi_t \mid \phi_{t-1}, \dot{\phi}_{t-1}) \times \\ & p(\dot{\phi}_t \mid \dot{\phi}_{t-1}) \end{aligned} \tag{2}$$

We now describe each of these terms in turn.

5.1.1. Visibility Update

At each time step, a hidden particle becomes visible with a fixed probability $p_{\mathrm{h}\to\mathrm{v}}$, and a visible particle becomes hidden with a fixed probability $p_{\mathrm{v}\to\mathrm{h}}$, i.e.

$$p(h_t \mid h_{t-1}) = \begin{cases} p_{\mathrm{h}\to\mathrm{v}}, & h_t = 0, h_{t-1} = 1 \\ 1 - p_{\mathrm{h}\to\mathrm{v}}, & h_t = 1, h_{t-1} = 1 \\ p_{\mathrm{v}\to\mathrm{h}}, & h_t = 1, h_{t-1} = 0 \\ 1 - p_{\mathrm{v}\to\mathrm{h}}, & h_t = 0, h_{t-1} = 0 \end{cases} \tag{3}$$

These probabilities are chosen carefully to give a desired equilibrium fraction of hidden particles, i.e. the fraction of particles that are hidden when the stationary distribution of the resulting Markov chain is reached assuming that all particles are re-weighted equally. If we specify a desired hidden fraction at equilibrium of q_h, then we must choose

$$p_{\mathrm{v}\to\mathrm{h}} = p_{\mathrm{h}\to\mathrm{v}} \frac{q_h}{1 - q_h} \tag{4}$$

to ensure that this equilibrium is achieved.

5.1.2. Viewing Plane Update

The probability of a transition between the different viewing planes is implemented simply as a discrete distribution with a constant probability of moving to each new state:

$$p(v_t \mid v_{t-1}) = \begin{cases} p_{\text{same}}, & v_t = v_{t-1} \\ p_{\text{change}}, & v_t \neq v_{t-1} \end{cases} \tag{5}$$

where generally $p_{\text{same}} >> p_{\text{change}}$. However, it is often helpful to slightly overestimate the probability of transition to allow the filter to recover from mistakes.

5.1.3. Location Update

Because the heart centre, $\mathbf{x}_t$, is defined differently in each view (see Fig. 1) it is necessary to model the position change that occurs when the view changes. We use a 2D Gaussian distribution to model each offset. The distributions are learnt at training time relative to a heart at orientation zero and with unit radius, giving *relative* offset distributions with means $\hat{\boldsymbol{\mu}}_{v_1 \to v_2}$ and covariances $\hat{\boldsymbol{\Sigma}}_{v_1 \to v_2}$, where the '$\hat{\cdot}$' is used to distinguish the *relative* distribution parameters. At test time, these are then scaled by the radius r and rotated by the orientation θ_{t-1} to give the *absolute* mean and covariance of the offset. Furthermore, we attempt to track the likely

changes in heart centre position using a simple off-the-shelf optical flow estimator (Farnebäck (2003)), giving a dense estimate of the displacement field $\mathbf{m}(\mathbf{x})$ between the previous frame and the current frame.

Specifically we have:

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \theta_{t-1}, v_t, v_{t-1}) = \mathcal{N}_{2D}(\mathbf{x}_t\,;\, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \quad (6)$$

where the (absolute) mean and covariance are given by:

$$\boldsymbol{\mu}_t = \mathbf{x}_{t-1} + \mathbf{m}(\mathbf{x}_{t-1}) + r\mathbf{R}_{[\theta_{t-1}]}\hat{\boldsymbol{\mu}}_{v_{t-1}\rightarrow v_t} \quad (7)$$

$$\Sigma_t = r\mathbf{R}_{[\theta_{t-1}]}\hat{\boldsymbol{\Sigma}}_{v_{t-1}\rightarrow v_t}\mathbf{R}^T_{[\theta_{t-1}]} \quad (8)$$

Here, $\mathcal{N}_{2D}(\,\cdot\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the probability density function (PDF) of a 2D Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\mathbf{R}_{[\theta]}$ is the 2×2 rotation matrix representing a rotation through angle θ. Note that we constrain the mean of the relative offset distribution, $\hat{\boldsymbol{\mu}}_{v_1,v_1}$, to be zero when the view does not change. However the covariance, $\hat{\boldsymbol{\Sigma}}_{v_1,v_1}$, is non-zero to represent random motion. In practice, sampling from the 2D Gaussian is achieved using the precomputed Cholesky decomposition of the covariance matrix (see Algorithm 2).

5.1.4. Orientation Update

The change in orientation accompanying each view transition is modelled by a *wrapped normal distribution* (Jammalamadaka and SenGupta (2001)) as this gives rise to a simple sampling method. Each view transition uses its own mean, $\hat{\xi}_{v_1\rightarrow v_2}$ and covariance $\tau_{v_1\rightarrow v_2}$ for the orientation offset, which are learnt at training time:

$$p(\theta_t \mid \theta_{t-1}, v_t, v_{t-1}) = \mathcal{WN}(\theta_t\,;\,\xi_t, \tau_{v_1\rightarrow v_2}) \quad (9)$$

where

$$\xi_t = \theta_{t-1} + \hat{\xi}_{v_{t-1}\rightarrow v_t} \quad (10)$$

and $\mathcal{WN}(\,\cdot\,;\xi,\tau)$ is the PDF of the wrapped normal distribution. Again we assume zero mean but non-zero variance when no view transition has occurred.

5.1.5. Cardiac Phase Update

The second order cardiac phase model applies a deterministic cardiac phase update according to the current cardiac phase rate (in rad s^{-1})

$$\phi_t = \phi_{t-1} + \frac{\dot{\phi}_{t-1}}{\Delta t} \quad (11)$$

where Δt is the (constant) time elapsed between video frames. The purpose of dividing by Δt here is to ensure that the state evolution model is not sensitive to the frame rate of the video being analysed.

5.1.6. Cardiac Phase Rate Update

Finally, to model the uncertain and variable cardiac phase rate, it is updated according to simple Gaussian noise with standard deviation υ:

$$p(\dot{\phi}_t \mid \dot{\phi}_{t-1}) = \mathcal{N}_{1D}(\dot{\phi}_t\,;\,\dot{\phi}_{t-1}, \upsilon) \quad (12)$$

This choice of state evolution model leads to a straightforward and efficient sampling algorithm, as outlined in Algorithm 2.

5.2. Initialisation

Before the first video frame, the set of particles $\mathbf{s}_0^{(i)}, i = 0, \ldots, N_p - 1$ is randomly initialised by drawing samples from initial distributions independently for each of the state variables. The initial distributions are: a discrete distribution representing the intended equilibrium hidden fraction q_h for the hidden/visible variables h, a discrete uniform distribution for the class label variables v, a continuous uniform distribution within the ultrasound fan area for the location variables $\mathbf{x}$, a circular uniform distribution for the orientation and phase variables θ and ϕ, and a gamma distribution fitted from the training set for the phase rate variables $\dot{\phi}$. The particle weights are initialised to a uniform value $w_0^{(i)} = \frac{1}{N}$.

6. Observation Model

The purpose of the observation potential function is to model the compatibility of a hypothesis about the current state, $\mathbf{s}_t$, with measurements from the observed image, $\mathbf{z}_t$ (see Fig. 3). This process is performed differently for hidden and visible particles. The overall form of the observation potential is:

$$\psi_o(\mathbf{s}_t, \mathbf{z}_t) = \begin{cases} \psi_a(v_t, \mathbf{x}_t \mid \mathbf{z}_t) \times & \\ \psi_b(\phi_t \mid v_t, \mathbf{x}_t, \mathbf{z}_t) \times & h_t = 0 \\ \psi_c(\theta_t \mid v_t, \mathbf{x}_t, \phi_t, \mathbf{z}_t), & \\ & \\ w_{\text{hidden}}, & h_t = 1 \end{cases} \quad (13)$$

The observation potential for a hidden particle is a constant value, w_{hidden} (see §6.5). We choose to decompose the observation potential function for non-hidden particles into three terms relating to the different variables that form the state. The first term, $\psi_a(\cdot)$, acts as a detector for a given heart view (in any orientation and phase) at a given position in the image. The second term, $\psi_b(\cdot)$, is a cardiac phase prediction term given the view classification and position. The final term, $\psi_c(\cdot)$, predicts the orientation given the predicted

Input: a particle $\mathbf{s}_{t-1}$ at time $t-1$ described by associated variables v_{t-1}, $\mathbf{x}_{t-1}$, θ_{t-1}, ϕ_{t-1} and $\dot{\phi}_{t-1}$, and a motion field estimate $\mathbf{m}(\mathbf{x})$

Output: an updated particle $\mathbf{s}_t$ at time t described by associated variables v_t, $\mathbf{x}_t$, θ_t, ϕ_t and $\dot{\phi}_t$

sample h_t according to the discrete distribution in Equation 3
sample v_t according to the discrete distribution in Equation 5
sample $\mathbf{n} \sim \mathcal{N}_{2D}(\mathbf{n}\,;\,\mathbf{0}, \mathbf{I}_{2\times 2})$ {*standard i.i.d. 2D Gaussian noise*}
$\mathbf{R} \leftarrow r \times \texttt{rotation\_matrix}(\theta_{t-1})$ {*calculate rotation and scaling matrix*}
$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \mathbf{m}(\mathbf{x}) + \mathbf{R}\left(\hat{\boldsymbol{\mu}}_{\mathbf{x},v_{t-1}\rightarrow v_t} + \texttt{cholesky}(\hat{\boldsymbol{\Sigma}}_{\mathbf{x},v_{t-1}\rightarrow v_t})\mathbf{n}\right)$ {*find position update*}
sample $\zeta \sim \mathcal{N}_{1D}(\zeta\,;\,0,1)$ {*standard Gaussian noise*}
$\theta_t \leftarrow \theta_{t-1} + \hat{\mu}_{\theta,v_{t-1}\rightarrow v_t} + \hat{\sigma}_{\theta,v_{t-1}\rightarrow v_t}\zeta$ {*update the orientation*}
$\phi_t \leftarrow \phi_{t-1} + \frac{\dot{\phi}_{t-1}}{\Delta t}$ {*update the cardiac phase*}
sample $\zeta \sim \mathcal{N}_{1D}(\zeta\,;\,0,1)$ {*standard Gaussian noise*}
$\dot{\phi}_t \leftarrow \dot{\phi}_{t-1} + \sigma_{\dot{\phi}}\zeta$ {*update the cardiac phase rate*}

Algorithm 2: Algorithm for sampling a single particle from the prediction potential

view classification, position and cardiac phase. Note that the cardiac phase rate is not observed explicitly, but rather observed implicitly by successive updates of the cardiac phase variable.

In order to construct models for the first two terms, we make use of random decision forests (Breiman (1999, 2001)) for three key reasons. Firstly, they are flexible and a similar algorithm can be applied to a variety of tasks, including classification and regression. Secondly, they are usually highly accurate discriminative classifiers/regressors and can naturally manage complex data without the tendency to overfit. Finally they can be highly efficient, particularly as only a subset of the available features need to be evaluated in order to make a decision. This is particularly important when evaluating each feature is relatively expensive.

In §6.1 we describe the image features we use for these three terms, and then in §6.2, §6.3, and §6.4 we describe the three terms in turn. In §6.5 we explain the choice of observation potential for hidden particles.

6.1. Rotation Invariant Image Features

In this work we choose to use rotation invariant features (RIFs) to describe circular regions of the image (as first introduced by Liu et al. (2014) and used in our earlier work, Bridge and Noble (2015)). This allows us to test the image at an arbitrary number of orientations without having to rotate the image before conducting each test. We will give only a brief overview of the method here and refer the reader to Liu et al. (2014) for more details. Underpinning our use of these features is an assumption that the acoustic reflection patterns from the tissue do not depend upon the insonification angle. While in general there are appearance variations with insonification angle in ultrasound imaging, particularly with highly reflective structures such as bone, we have found that RIFs work well in practice in our application.

RIFs are extracted from a circular region of the image by convolving it with a set of complex-valued rotation invariant basis functions. Each such convolution yields a complex number, and together these numbers describe the circular region. Taking the magnitude of these complex numbers gives a description of the region that is analytically invariant to the orientation of the underlying image region. 2D vector-valued image representations, such as a gradient or motion field, can also be described in this framework by first representing each vector as a magnitude-weighted delta function in a continuous orientation histogram and expressing this continuous histogram in terms of a truncated set of Fourier series coefficients. Then, the same basis functions can be used on these Fourier coefficients to yield a set of complex numbers whose magnitudes are invariant to the orientation of the underlying image. A set of basis functions can be described by its number of radial divisions J and its maximum rotation order K, and the number of Fourier coefficients M is a further parameter of the feature extraction stage.

In this work, we experimented with using *intensity*, intensity *gradient* and *motion* representations of the frames, as well as combinations of these where features from either set may be chosen by the split nodes in the forests. We denote the set of complex-valued RIFs that may be calculated from the image $\mathbf{z}_t$ at image location $\mathbf{x}_t$ by the vector $\mathbf{f}(\mathbf{z}_t, \mathbf{x}_t)$. The split functions in our random forests design are comparisons of the magnitude of a single RIF from this set with a threshold, or a comparison of the result of coupling two RIFs of the same rotation order with a threshold.

Where applicable, the motion estimate used is the one obtained for the state evolution model (§5). In order to apply an approximate correction for the fact that the motion patterns will depend upon the video's frame rate, we normalise the magnitude of the motion field by the frame period before extracting features. Despite being a crude approximation, we have found that this

works well in practice. We believe that this is because the random forests learn to look only for features that encode very coarse motion patterns (i.e. roughly which areas of the image patch are moving in roughly what direction) rather than fine detail.

In order to make the feature extraction process as efficient as possible, we have implemented the algorithm of Liu et al. (2014) with the following alterations[1]:

- Individual features are only calculated as they are required by the random forests algorithm. Furthermore, once calculated, individual feature results are stored so that they may be efficiently used again if required later in the processing of the same frame.
- Where the same feature is required for a large number of image locations at once (typically at nodes near the root of the trees), it becomes far more efficient to implement the convolutions as Fourier domain multiplications (via the 2D fast Fourier transform). This requires the Fourier-domain representations of the basis functions, which we derive and present in the supplementary materials.
- Images are scaled at training time and test time such that the radius of the detection window is a constant value, r_{RIF}, across all samples. A small window size results in faster calculations but may result in loss of detail from the image patches before the RIFs are extracted. In practice we have found that a relatively small value (around 30 pixels) can be chosen without significant loss of accuracy.

6.2. Classification Forests

In order to detect and distinguish between the three different views of the fetal heart given appearance features from the image, we use a four-class classification forest. In this case, the view label is a discrete class identifier from the set $\mathcal{V} = \{\mathrm{BG}, 4\mathrm{C}, \mathrm{LVOT}, 3\mathrm{V}\}$, representing the background, four-chamber view, left ventricular outflow tract view and three vessel view respectively. Accordingly, each leaf node consists of an empirical discrete distribution over these labels. The training objective function that governs which split functions are chosen at the nodes during training is the *information gain* (e.g. Criminisi et al. (2011)), which measures the change in entropy between the label sets before and after the split:

$$I_v(\mathcal{S}_n, \mathcal{S}_L, \mathcal{S}_R) = H(\mathcal{S}_n) - \sum_{i \in \{L,R\}} \frac{|\mathcal{S}_i|}{|\mathcal{S}_n|} H(\mathcal{S}_i) \qquad (14)$$

[1] our C++ implementation using OpenCV is available at `https://github.com/CPBridge/RIFeatures`

where $\mathcal{S}_n$ is the set of labels in the n^{th} node (being trained), and $\mathcal{S}_L$ and $\mathcal{S}_R$ are respectively the sets of labels in the left and right nodes after the proposed split. $H(\cdot)$ represents the entropy of a set of discrete labels:

$$H(\mathcal{S}) = -\sum_{v \in \mathcal{V}} p(v) \log p(v) \qquad (15)$$

A forest consists of N_{trees} trees, and training is stopped after a maximum tree depth (d_{max}), or when the number of training data in a node goes below a threshold ($N_{nodemin}$), or when the information gain from splitting goes below a threshold $I_{v,min}$. After the classification forest has been trained, the resulting probability density function (PDF) is used straightforwardly as the first term of the observation potential from Equation 13.

$$\psi_a(v_t, \mathbf{x}_t \mid \mathbf{z}_t) = p\left(v = v_t \mid \mathbf{f}(\mathbf{z}_t, \mathbf{x}_t)\right) \qquad (16)$$

6.3. Circular Regression Forests

We use a circular regression forest to predict the cardiac phase of the heart given appearance features from the image. In this case the label is a real number ϕ in the range $[0, 2\pi)$. Because of the wrapped nature of circular variables, it would be incorrect to treat this task as a standard regression problem. We therefore adapt the random forests algorithm to deal with angular variables correctly.

Firstly we define a *circular mean* for a set of N angular labels (Jammalamadaka and SenGupta (2001)):

$$\bar{\phi} = \mathrm{atan2}\left(\frac{1}{N}\sum_{i=1}^{N} \sin\phi_i, \frac{1}{N}\sum_{i=1}^{N} \cos\phi_i\right) \qquad (17)$$

where $\mathrm{atan2}(\cdot)$ is the four quadrant arctangent function.

We then use an approximate measure of circular information gain found by substituting the notion of a circular distance from (Jammalamadaka and SenGupta (2001)) in place of linear distance in the commonly-used regression objective function.

$$I_\phi(\mathcal{S}_n, \mathcal{S}_L, \mathcal{S}_R) = \sum_{i \in \mathcal{S}_n} \frac{1}{2}\left(1 - \cos(\phi_i - \bar{\phi}_{\mathcal{S}_n})\right)^2 - \sum_{j \in \{L,R\}} \left(\sum_{i \in \mathcal{S}_j} \frac{1}{2}\left(1 - \cos(\phi_i - \bar{\phi}_{\mathcal{S}_j})\right)^2 \right) \qquad (18)$$

where $\bar{\phi}_{\mathcal{S}_j}$ is the mean of the angular labels in set $\mathcal{S}_j$. This cost function measures the difference between the sum of squared distances from the mean in the node before and after splitting, and therefore favours splits that cluster similar angular labels together.

The leaf distribution in the case of the circular regression forest is a *von Mises* distribution (also known as the *circular normal* distribution, Jammalamadaka

and SenGupta (2001)). This is a commonly used distribution when working with circular data, as it has a convenient form and is the maximum entropy distribution for circular variables with a given circular mean and circular variance. The distribution has two parameters: the mean angle μ and the concentration κ, where μ and $1/\kappa$ are analogous to the μ and σ^2 parameters from the Gaussian distribution. The probability density function of the von Mises distribution is:

$$p(\phi \mid \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\phi - \mu)} \tag{19}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero. See Fig. 4 for examples of von Mises PDFs. We refer the reader to (Jammalamadaka and SenGupta (2001)) for further details on fitting a von Mises distribution using maximum likelihood estimation.

We train one circular regression forest with N_{trees} trees for phase regression in each of the non-background classes separately. The stopping criteria were the same as for the classification forests, with an appropriate information gain threshold $I_{\phi,min}$. At test time, the data point is passed down each tree in the relevant forest until it reaches a leaf node, and the PDF is then calculated for the cardiac phase value ϕ_t given the distribution parameters (μ_b, κ_b) at that leaf node. The second part, $\psi_b(\cdot)$, of the observation potential from Equation 13 is then given by the averaged PDF across the trees in the forest.

6.4. Orientation Regression Model

The orientation prediction step takes advantage of the fact that the complex-valued image features (that is, the raw feature values before the magnitude is taken to give rotation invariance) are in fact equivariant under rotation of the underlying image window. We can therefore use the complex arguments of RIFs with a rotation order of 1 (or -1) as an indication of the orientation of the heart.

We find that it is not necessary to build new decision forests for the task of orientation prediction. Rather, the clustering that results at the end of the phase prediction term is sufficient to give good results for orientation prediction, even though it is not optimised for this purpose. Therefore, after we have trained a phase prediction forest, we simply fit an individual orientation prediction model to the data in each leaf node. For each data point we calculate the *offset angle* between the orientation label θ_i and j^{th} complex feature $f_j(\mathbf{x}_i)$ calculated at the image patch with centre $\mathbf{x}_i$ to be

$$\delta_{ij} = \arg\left(f_j(\mathbf{x}_i)\right) - \theta_i \tag{20}$$

We then fit a von Mises distribution (μ_c,κ_c) of this offset angle across all the datapoints i in the leaf node for each feature j of rotation order one, and choose the feature j^* that has the largest concentration parameter κ. Then, at test time, the PDF at each leaf node, n, is calculated using this chosen feature and its von Mises distribution

$$p(\theta_t \mid v_t, \mathbf{f}(\mathbf{z}_t, \mathbf{x}_t)) = p\left(\arg\left(f_{j^*}(\mathbf{x}_t)\right) - \theta_t \mid \mu_{c,n}, \kappa_{c,n}\right) \tag{21}$$

Again, the final part, $\psi_c(\cdot)$, of the observation potential from Equation 13 is found by averaging the PDF predictions from the individual trees.

6.5. Observation Potential for Hidden Variables

Recall that at each resampling step in the particle filter, the particle weights are normalised to sum to unity. This means that the values of the observation functions only matter relative to the other particles in the set, and the absolute values make no difference to the behaviour of the overall filter. If there were no hidden particles, this would mean that when the heart becomes hidden, the filter would continue to track whichever area of the image results in the largest observation potential, regardless of the absolute value of those observation function evaluations. As a result there would be no easy way to decide whether the heart is visible in the image or not.

For this reason, we re-weight the hidden particles (those with $h_t = 1$) with a small constant weight value, w_{hidden}, that does not depend on the other state variables or image information (Equation 13). When the majority of non-hidden particles receive a large weight from the random forests, indicating that the random forests are confident about the presence of a heart, the weights of the hidden particles are relatively insignificant and most will not survive the next resampling step. However, when most of the non-hidden particles are given a small weight by the random forests, the fixed weights of the hidden particles become relatively more significant and may come to dominate the particle set.

The value of w_{hidden} controls the sensitivity of the filter, and must be selected carefully to give the desired behaviour.

7. Experiments

7.1. Experimental Data

We acquired a diverse dataset of 91 short ultrasound videos of the fetal heart drawn from 12 subjects during routine clinical scans using a GE Voluson E8 ultrasound device. Each video had a length of between 2 and 10 seconds and a frame rate between 25 and 76 frames per second, and contained one or more of the three views of the fetal heart defined in §3. The videos captured the healthy fetal heart in a range of magnifications and orientations, though with the heart taking up approximately 30% or more of the image. There was

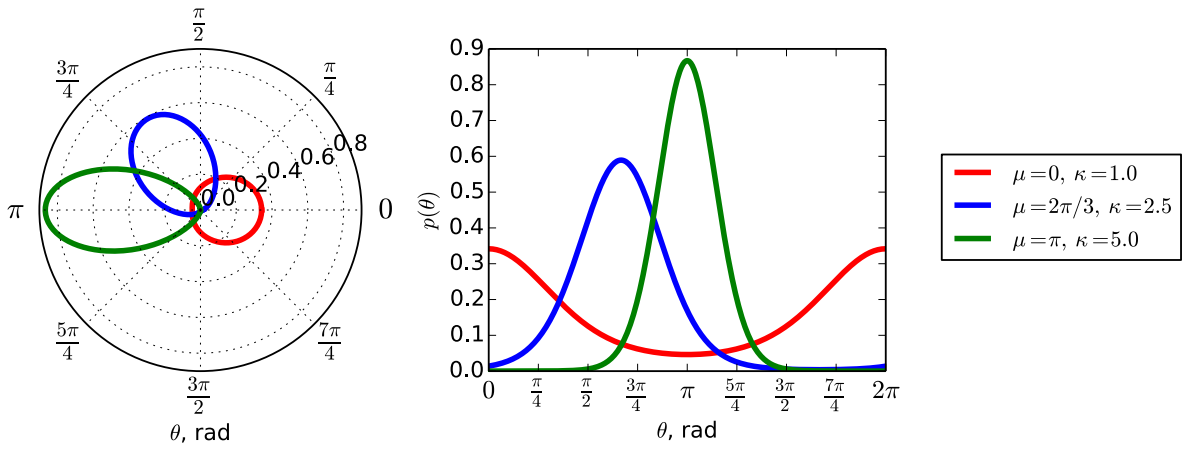


Figure 4: PDFs of three von Mises distributions defined over the interval 0 to 2π shown in both polar (*left*) and Cartesian (*right*) form. The μ parameter represents the mean angle, and the κ parameter describes the concentration of probability mass around this mean.

a range of gestational ages from 20 to 35 weeks. All videos were gathered such that the fetal head would be towards the viewer when viewed on a screen, or were flipped horizontally for consistency before being used for training and testing.

Each frame of each video was manually annotated according to the criteria shown in Fig. 1 in order to provide labels for training and validating the model. These annotations were approved by a clinician experienced in interpreting ultrasound videos of the fetal heart (C. Ioannou).

7.2. Implementation Details

Our framework was implemented in the C++ programming language using the OpenCV 3.1.0 and Eigen 3.2.5 software libraries. Several processes were parallelised using OpenMP compiler extensions of the G++ 4.8.4 compiler. All timings were obtained on a desktop computer (8-core Intel Core i7-3770 3.4 GHz running a 64-bit OS).

7.3. Training

Due to the relatively small number of subjects, we used a leave-one-out cross-validation procedure across each subject. Specifically we tested every video in the dataset with a model trained on the data from all other subjects in the dataset. The training procedure for each partition involved randomly selecting 5000 positive example windows for each of the three cardiac views, and an equivalent number of background examples from random image locations in the same videos, but at least $0.3r$ from the labelled heart centre, where r is the heart radius. The selected examples were used to train the view classification forest, $\psi_a(\cdot)$, and then the positive examples from the relevant class were used to train a phase regression model, $\psi_b(\cdot)$, and an orientation regression model, $\psi_c(\cdot)$, for each class individually. Furthermore, a simple maximum likelihood model was fitted for the various phase transition distributions in §5 by finding view transitions in the labelled videos. Other parameters for training the random forest models and for the state evolution model were chosen empirically and are shown in Tables 1 and 2. In particular, previous experiments (unreported) have shown that increasing the number of trees in the random forests does not significantly improve the accuracy of the models, but does increase the execution time due to the increased number of feature evaluations.

Parameter	Value
r_{RIF}	30 pixels
N_{trees}	8
d_{max}	10
$N_{nodemin}$	50
$I_{v,min}$	0.5 bits
$I_{\phi,min}$	0.01

Table 1: Random Forest Training Parameters

Parameter	Value
N_P	1000
N_{thresh}	$0.3N_P$
p_{same}	0.9
$\hat{\mathbf{\Sigma}}_{v_1 \to v_2}$ for $v_1 = v_2$	$1.0 \times \mathbf{I}$ pixels^2
$\tau_{v_1 \to v_2}$ for $v_1 = v_2$	0.05 rad
υ	0.2 rads^{-1}
$\dot{\phi}_{\min}, \dot{\phi}_{\max}$	100, 200 bpm
q_h	0.3
$p_{h \to v}$	0.4

Table 2: Particle Filter Parameters

7.4. Validation Methodology

We evaluated the performance of our framework in two variants: the first used just the observation potentials at each frame individually with no particle filtering, whilst the second used full temporal filtering.

For the variant in which just the observation potentials were used, all image patches in each image were passed through the classification forest and the predicted position was chosen to be the patch with the highest probability of having any of the non-background labels. The predicted view label was chosen to be the label giving this highest probability. The predicted cardiac phase and orientation were then determined using the phase/orientation forest at only that image location and taking the mean of the resulting von Mises distribution.

For the particle filtered variant, the particle filtering model described in §5 was used with the parameter values in Table 2, which were chosen by empirically following previous experiments (unreported). A single state prediction was determined at each time step using the mean-shift algorithm on the particles. All reported accuracy values were averaged over all videos, with each video given equal weight regardless of its length. In the particle filtered variant, accuracy values from five test runs were averaged due to the inherently stochastic nature of the filter's output.

For the purposes of reporting accuracy, we considered the heart view to be correctly detected if the predicted view label, v, matched the annotation and the predicted heart centre, $\mathbf{x}$, was within $0.25r$ of the annotated centre. This corresponds to approximately 2.5 mm to 4.5 mm at the gestational ages we are considering. Error between the true and predicted values of the angular variables (orientation θ and cardiac phase ϕ) was assessed using the following normalised circular distance metric between two angles θ_1 and θ_2:

$$\frac{1}{2}\left(1 - \cos\left(\theta_1 - \theta_2\right)\right) \tag{22}$$

giving a value in the range 0 (meaning precisely correct) to 1 (meaning an error of 180°or π rad). Error values for orientation and cardiac phase were only averaged over frames where the view classification and position were correctly determined.

The two variants deal with the possibility of the heart being hidden in very different ways. When no filtering is used the posterior detection probability of the maximum class can be simply thresholded to determine whether the heart is hidden. When filtering is used, the prediction is instead based on whether the total weight of 'hidden' particles is greater than the total weight of 'visible' particles. These methods are both sensitive to the relevant parameters, which are the threshold value for the former case and the 'hidden' weight w_{hidden} in the latter. In order to give a fair comparison between the two algorithms, we conducted initial experiments with the most sensitive setting (threshold of 0.0 and $w_{hidden} = q_h = 0.0$), and then performed a second experiment in which the threshold was varied.

7.5. Analysis of Inter- and Intra-Observer Variation

In order to place our results in context, we analysed the inter- and intra-observer variation of our annotations. Due to time contraints, a subset of the full dataset consisting of 12 videos (one video from each subject) was used for this purpose, however this was deemed sufficient to quantify approximately the degree of variation in the annotations. The annotations were repeated on these videos by the same annotator approximately 10 months after the initial annotations to estimate the intra-observer variation. Additionally, a second annotator was trained to annotate the videos using the same guidelines and provide a third set of annotations on the smaller dataset to estimate the inter-observer variation. These new annotations were compared to the ground truth annotations in exactly the same way as the predictions from the automatic algorithm.

8. Results and Discussion

We present results of the leave-one-out cross-validation experiments in the two variants (with and without filtering) and using a number of different sets of features. Results are shown in Figs. 5, 6 and 8.

Figure 5 shows a comparison of results obtained using the full filtering framework (right hand plots 5b, 5d and 5f) and using the observation potentials alone (left hand plots 5a, 5c and 5e). These plots show computation speed per frame on the y-axis and prediction error (in the relevant sense) with respect to the manual ground truth annotations on the x-axis. Results for a number of different feature extraction methods are shown, and highlight that there is generally a trade-off between speed and accuracy when choosing the feature extraction method. All the results in Fig. 5 where obtained using a threshold of posterior detection threshold of 0.0 (for the unfiltered case) and $w_{hidden} = q_h = 0.0$ (for the filtered case).

The first pair of plots (Fig. 5a and 5b) show the per-frame conmbined detection and classification error rate averaged over every video in the cross-validation regime, where this error rate is defined as the fraction of 'positive' frames (those labelled as containing a view of the heart in the ground truth) in which the heart was either detected in the incorrect location (i.e. more than $0.25r$ from the labelled centre location) *and/or* the predicted view label was incorrect. The best feature extraction methods are able to achieve under 20% error rate on this challenging imagery. The second and third pairs of plots respectively show the orientation (Fig. 5c

and 5d) and cardiac phase (Fig. 5e and 5f) error rates, defined using the normalised angular distance metric (Equation 22) *over only the frames with correct classification and detection*. Again these are averaged over all videos in the cross-validation regime. An ideal feature extraction method would be fast and accurate and therefore appear close to the bottom left of all plots.

By comparing the plots in the left and right columns, we can see that, for a given feature extraction method, the addition of the particle filtering framework to the random forest observations significantly reduces all three of the error rates (for some feature extraction methods the classification error rate can be reduced by 10 percentage points or more) at the expense of making the algorithm slower by 5-10 ms per frame. The speed reduction is primarily due to the need to calculate the phase and orientation output values at a large number of image locations, rather than a single location (as is required by the observation potentials only regime), and not due to the overhead of the filter implementation itself.

We also see that the choice of features is another important consideration, and all plots show a similar trend here. Using simple intensity features (diamond markers) gives a very fast (<20 ms per frame) but less accurate prediction over classification, orientation and cardiac phase (e.g. with classification error rates of around 50% or more), whereas using gradient features (triangle markers) gives a lower error rate (e.g. around 30% classification error rate) and can run only slightly slower at around 20 ms per frame. The inclusion of motion features (circular markers) further reduces the classification error rate to under 20% and greatly improves the cardiac phase prediction (as might be expected), but at the expense of a significant reduction in speed, resulting in speeds of around 25 ms per frame. However even this lower frame rate (around 40 frames per second) is fast enough to process the majority of ultrasound videos in real time. We also tried using the monogenic odd filter as an alternative image representation, as in our previous work (Bridge and Noble (2015)), but found that this did not have any advantages over the image gradient. Intensity features tend to perform much better when more features are available (high J, and K values). By contrast gradient features only require a smaller number (around $J = 3, K = 3, M = 2$) to give a sufficiently rich description, and there is little or no increase in performance above this but a large reduction in speed.

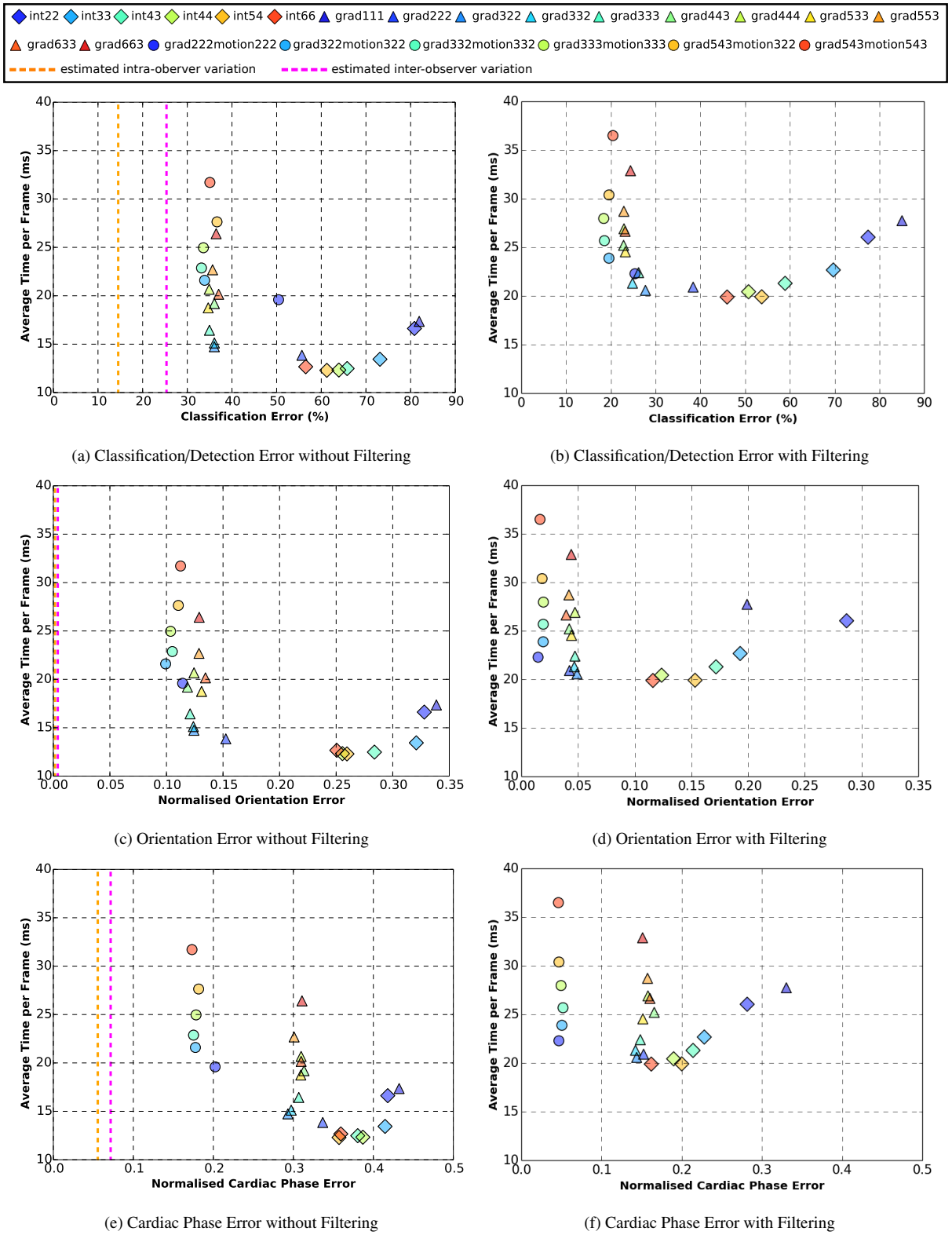
The dashed vertical lines in Figure 5 show the 'error rate' (i.e. disagreement) between the ground truth and the annotations performed for a second time by the first annotator (orange line) and by the second annotator (magenta line), evaluated in the same way as the automatic predictions. This gives us a target region for the performance of our automatic method. We can see that there is a significant disagreement between the different sets of annotations, reflecting the highly ambiguous nature of the annotation task. The performance of the best automatic methods is in approximately the same region as this agreement for the classification/detection and cardiac phase prediction tasks, and slightly worse for the orientation prediction tasks.

Figure 6 shows average confusion matrices for a few representative parameter sets. It is clear from these confusion matrices that the three vessel (3V) view is the most commonly missed view, which is unsurprising given that its appearance is less distinctive than the other anatomical two views. Furthermore, we see that the majority of inter-class confusion arises between the four chamber (4C) and left ventricular outflow tract (LVOT) views, which is again unsurprising given the sometimes ambiguous distinction between the views when the probe is physically located in the space between the two.

In order to investigate the performance of the proposed framework at detecting when the heart is hidden, we performed a further experiment using a subset of the feature extraction methods. By increasing the relevant parameters (the detection threshold in the unfiltered case and the hidden particle weight in the filtered case) it is possible to eliminate many false positive detections, but there is inevitably a trade-off here with the true positive classification rate. This is illustrated in Fig. 8. For the filtered case we determined good values for the other parameters of the hidden particles via pilot experiment ($q_h = 0.3, p_{h\to v} = 0.4$) and kept them constant during the experiment.

However, these figures are somewhat misleading, as many of the frames that were labelled negative in the training set in fact contained an obscured view heart set. This can happen if, for example, the heart appears indistinct due to motion blurring or shadowing artefacts. If the algorithm detects that such a frame contains the heart, this is classified as a false positive and leads to a reduced performance. We therefore also evaluated the false positive rate using a different set of "generous" labels, in which these borderline cases were not considered incorrect if the position and class was correct. The results are shown in Fig. 8 with dashed lines, where it can be seen to reduce the false positive rate.

Again, we compared the performance of the automatic method to the agreement between the human annotators. The plotted points in Fig. 8 show the true positive and false positive rates of the alternative sets of annotations with respect to the ground truth set. Whilst the algorithm approaches the performance of the inter-observer variation, it is significantly worse than the intra-observer variation when considering the false-positive rate. Partly this reflects the fact that different annotators have different thresholds for when the heart is 'visible', as it is very difficult to establish an objective threshold.



(a) Classification/Detection Error without Filtering

(b) Classification/Detection Error with Filtering

(c) Orientation Error without Filtering

(d) Orientation Error with Filtering

(e) Cardiac Phase Error without Filtering

(f) Cardiac Phase Error with Filtering

Figure 5: (a)(b) Classification/detection error, (c)(d) orientation error, and (e)(f) phase error versus time for a selection of feature extraction methods. Each feature extraction method appears as a separate marker which appears in the legend with the name of the image representation (*int* intensity, denoted by a diamond marker; *grad* gradient, denoted by a triangle marker; or a combination of gradient and *motion*, denoted by a circle marker). After each name the parameters of the basis function set are listed (number of radial profiles J, number of rotation orders K and, where relevant, number of Fourier coefficients M). So, for example, *int43* refers to using features from an intensity representation with parameters $J = 4$ and $K = 3$, and *grad543motion322* refers to a combination of features from a gradient representation (parameters $J = 5$, $K = 4$, $M = 3$) with those from a motion field representation (parameters $J = 3$, $K = 2$, $M = 2$).

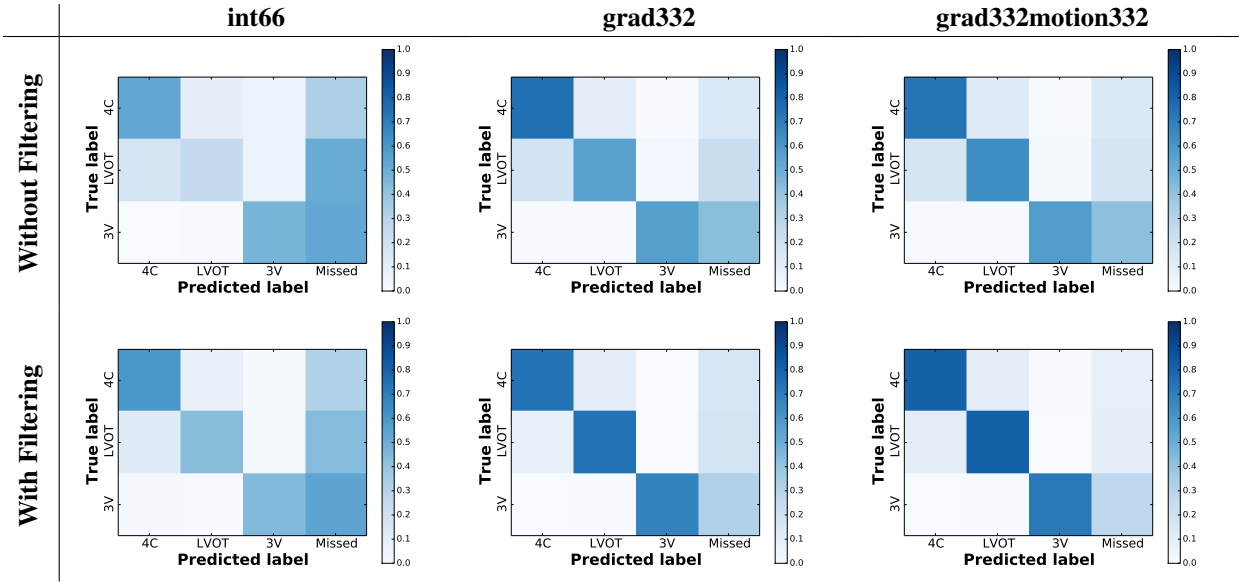


Figure 6: Average confusion matrices over videos for some parameter sets. Matrices are normalised such that each row sums to one. *Top row* without filtering, *bottom row* with filtering. The heart was 'missed' if the location of the detected centre was greater than $0.25r$ from the true centre.

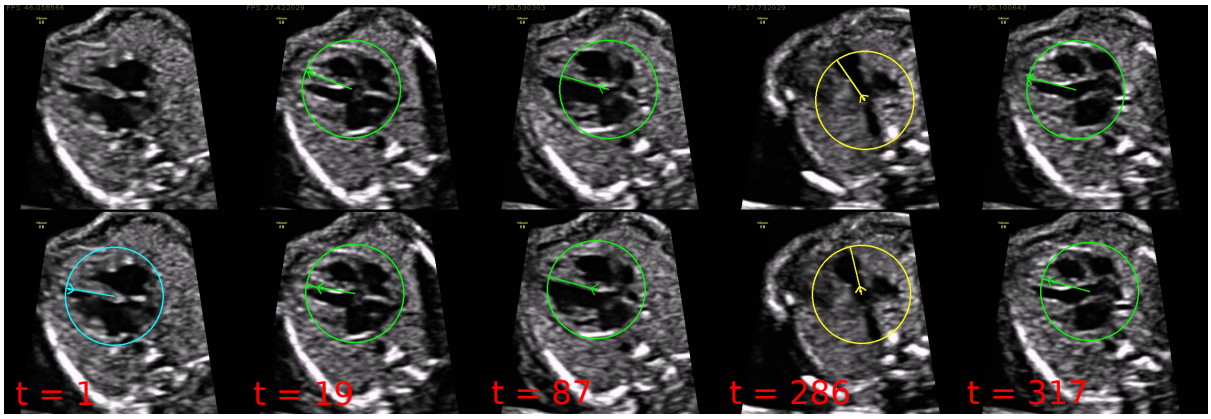
Figure 7 shows the performance on two example test sequences. The full sequences may be viewed in the supplementary materials.

9. Conclusions and Future Work

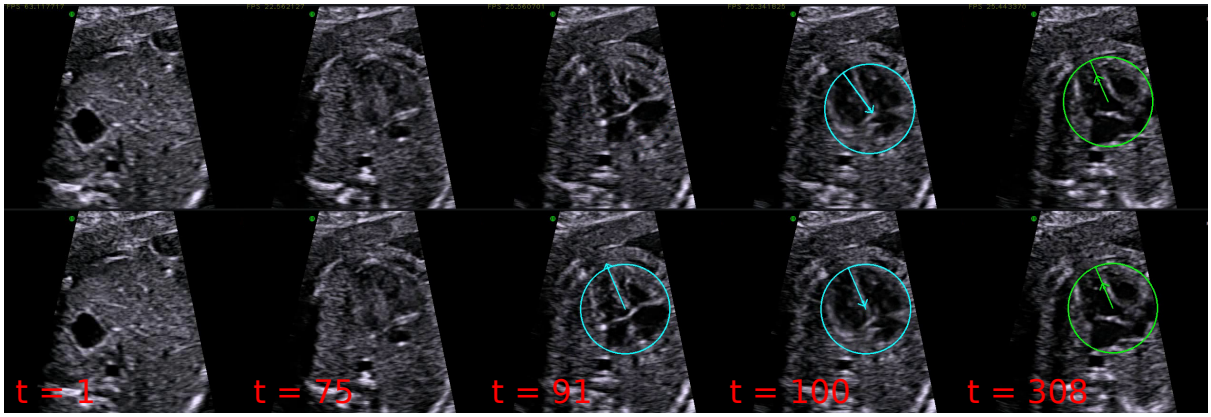
In this paper we have presented a method for extracting key information from 2D ultrasound videos of the fetal heart at high frame rates (average of about 40 frames per second for the best parameter sets). We chose to use a particle-filtering based method to overcome the intractability of the recursive state estimation problem with our state definition, and employed random forest based predictors as effective, discriminative observation potentials. The use of a relatively strong model of heart dynamics was found to significantly improve upon prediction on a frame-by-frame basis. We validated our model on real data gathered in a clinical setting, with promising results. Future work should investigate ways to optimise systematically the various parameters of the particle filter.

We hope that this paper will inform and inspire further work towards providing automated tools for the diagnosis of CHD from cardiac ultrasound videos. In particular there are a number of open questions raised by this work. Firstly it remains to be shown how the information that we are currently able to extract can be used to maximal benefit in clinical practice. We envisage three particular possible uses, the first is that the information could be fed back live to the sonographer via on-screen graphical cues. This would be particularly useful for trainee sonographers and would enable them to confirm their own assessments of the images and ensure that they have visualised all the correct views. Secondly, information about several scanning sessions could be stored and sonographers' scanning habits analysed to be ensure that scans are being conducted consistently. Thirdly, if the scan video is stored for later review, the extracted information could be stored alongside as metadata in order to enable easy retreival of relevant parts of the video.

Further it remains to be shown how the model can be extended to cope with, and identify cases of, CHD. CHD represents a large variety of interacting abnormalities, and therefore this will likely entail a variety of approaches. Some indicators of CHD may be deduced from the information the existing method provides with little or no extension. For example abnormal heart rate could be detected from our algorithm, and abnormal cardiac situs or abnormal axis (orientation with repect to the abdomen) could be assessed by coupling this work with an abdomen detector. Other subtle or highly-localised problems, such as small septal defects, abnormal alignment of valves or vessel coarctation, are unlikely to affect the functioning of the algorithm and could be detected by further learning-based processes that make use of the coordinate system our framework can provide. More significant problems, such as ventricular hypertrophy and conotruncal anomalies, significantly alter the appearance of the heart and will therefore require more substantial changes to the framework. This would entail training the observation models and state update models with abnormal data, and then distinguishing abormal cases either by extending the state vector to contain these variables, or through secondary processes

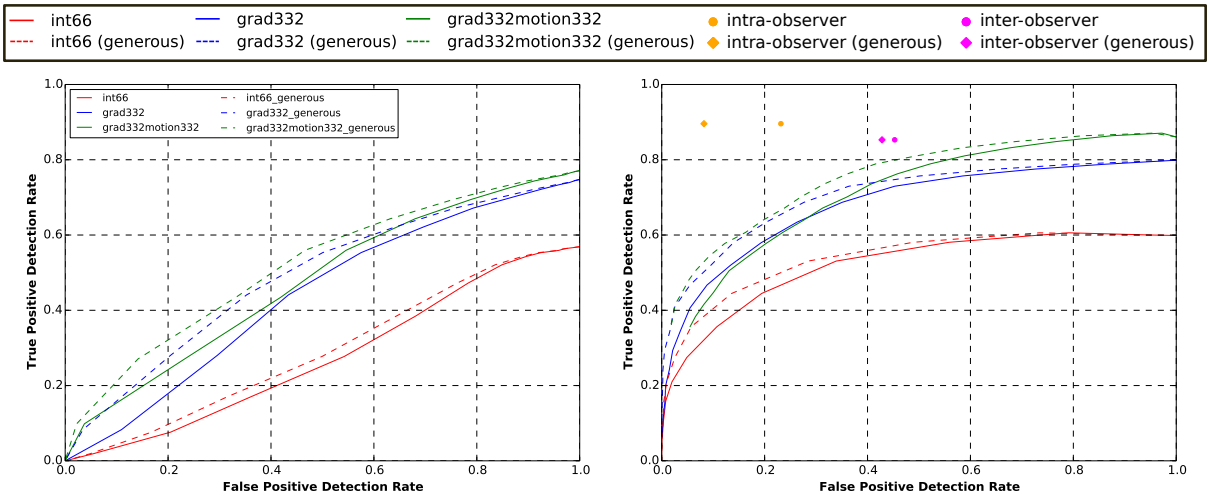


(a) The algorithm finds and tracks the view label, position and orientation very quickly, and then tracks approximately the correct cardiac phase after about 10 frames, including through changes in view label.



(b) The filter correctly detects that the heart is not visible at the beginning of this sequence (the stomach is instead visible on the left of the abdomen). When the heart appears, the filter is slow to pick it up but then begins to track correctly.

Figure 7: Results of the algorithm on two example sequences (in each sequence the *top row* shows the prediction, and the *bottom row* shows the ground truth). See Figure 1 for the meaning of the annotations including the view label colour scheme. Times shown are frame numbers. Parameters were as listed in §7 with $w_{hidden} = 0.025$, and combined gradient and motion features with $J = 3, K = 3, M = 2$ were used. Additionally, the position of the arrow head represents the position in the cardiac cycle (pointing outwards represents systole and pointing inwards represents diastole). See the online supplementary materials for full video.



(a) Without Filtering

(b) With Filtering

Figure 8: True positive and false positive detection rates as the relevant parameter is varied for some typical parameter sets. a) The unfiltered case where the detection threshold parameter is varied from 0 to 1. b) The filtered case where w_{hidden} is varied from 0.00 to 0.01.

that operate on the estaimted heart location. Fortunately however problems in this final class are typically the easiest to detect without computer assistance.

Acknowledgements

We are grateful to Vaanathi Sundaresan for providing a second set of annotations. Christopher Bridge acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award (ref. 1337676). Alison Noble acknowledges the support of EPSRC grant EP/M013774/1 (the SeeBiByte Project).

Data Statement

The videos used in this paper cannot be made freely available for reasons of ethical sensitivity. Data related to the results have been made available at the Oxford University Research Archive (ORA-Data) at DOI 10.5287/bodleian:VYxxRo55b.

References

Agarwal, D., Shriram, K., Subramanian, N., 2013. Automatic view classification of echocardiograms using histogram of oriented gradients, in: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, pp. 1368–1371.

Allan, L., 2000. Antenatal diagnosis of heart disease. Heart 83, 367.

Breiman, L., 1999. Random Forests. Technical Report TR567. U. C. Berkeley.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Bridge, C., Noble, J., 2015. Object Localisation in Fetal Ultrasound Images Using Invariant Features, in: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, pp. 156–159.

Carneiro, G., Georgescu, B., Good, S., Comaniciu, D., 2008. Detection and Measurement of Fetal Anatomies from Ultrasound Images using a Constrained Probabilistic Boosting Tree. Medical Imaging, IEEE Transactions on 27, 1342–1355.

Carneiro, G., Nascimento, J.C., 2013. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2592–2607.

Carvalho, J., Allan, L., Chaoui, R., Copel, J., DeVore, G., Hecher, K., Lee, W., Munoz, H., Paladini, D., Tutschek, B., Yagel, S., 2013. ISUOG practice guidelines (updated): sonographic screening examination of the fetal heart. Ultrasound Obstet Gynecol 41, 348–359.

Chen, H., Dou, Q., Ni, D., Cheng, J.Z., Qin, J., Li, S., Heng, P.A., 2015. Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent Neural Networks, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing. volume 9349 of *Lecture Notes in Computer Science*, pp. 507–514.

Criminisi, A., Shotton, J., Konukoglu, E., 2011. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Technical Report MSR-TR-2011-114. Microsoft Research.

Doucet, A., de Freitas, N., Gordon, N., 2001. Sequential Monte Carlo Methods in Practice. Springer.

Ebadollahi, S., Chang, S.F., Wu, H., 2004. Automatic view recognition in echocardiogram videos using parts-based representation, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA. pp. 2–9.

Farnebäck, G., 2003. Two-frame motion estimation based on polynomial expansion, in: Bigun, J., Gustavsson, T. (Eds.), Image Analysis. Springer Berlin Heidelberg. volume 2749 of *Lecture Notes in Computer Science*, pp. 363–370.

Gao, Y., Maraci, M., Noble, J., 2016. Describing Ultrasound Video Content Using Deep Convolutional Neural Networks, in: Biomedical Imaging, 2016 IEEE International Symposium on, pp. 787–790.

Hill, G., Block, J., Tanem, J., Frommelt, M., 2015. Health Disparities in the Prenatal Detection of Critical Congenital Heart Disease. Presented at the Pediatric Academic Societies Annual Meeting, San Diego.

Jacob, G., Noble, J., Blake, A., 1998. Robust contour tracking in

echocardiographic sequences, in: Computer Vision, 1998. Sixth International Conference on, pp. 408–413.

Jammalamadaka, S.R., SenGupta, A., 2001. Topics in Circular Statistics. World Scientific Pub Co Inc.

Kim, H.D., Kim, D.J., Lee, I.J., Rah, B.J., Sawa, Y., Schaper, J., 1992. Human fetal heart development after mid-term: Morphometry and ultrastructural study. Journal of Molecular and Cellular Cardiology 24, 949–965.

Kumar, R., Wang, F., Beymer, D., Syeda-Mahmood, T., 2009. Echocardiogram view classification using edge filtered scale-invariant motion features, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 723–730.

Limketkai, B., Fox, D., Liao, L., 2007. CRF-filters: Discriminative particle filters for sequential state estimation, in: Robotics and Automation, 2007 IEEE International Conference on, pp. 3142–3147.

Liu, K., Skibbe, H., Schmidt, T., Blein, T., Palme, K., Brox, T., Ronneberger, O., 2014. Rotation-invariant HOG descriptors using Fourier analysis in polar and spherical coordinates. International Journal of Computer Vision 106, 342–364.

Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A., 2014. Searching for Structures of Interest in an Ultrasound Video Sequence, in: Machine Learning in Medical Imaging. Springer International Publishing. volume 8679 of *Lecture Notes in Computer Science*, pp. 133–140.

Namburete, A., Rahmatullah, B., Noble, J., 2013. Nakagami-based Adaboost learning framework for detection of anatomical landmarks in 2D fetal neurosonograms. Annals of the BMVA 2, 1–16.

Nascimento, J.C., Marques, J.S., 2008. Robust Shape Tracking With Multiple Models in Ultrasound Images. IEEE Transactions on Image Processing 17, 392–406.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175.

Park, J., Zhou, S., Simopoulos, C., Otsuki, J., Comaniciu, D., 2007. Automatic cardiac view classification of echocardiogram, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE. pp. 1–8.

Pézard, P., Bonnemains, L., Boussion, F., Sentilhes, L., Allory, P., Lépinard, C., Guichet, A., Triau, S., Biquard, F., Leblanc, M., Bonneau, D., Descamps, P., 2008. Influence of ultrasonographers' training on prenatal diagnosis of congenital heart diseases: a 12-year population-based study. Prenatal Diagnosis 28, 1016–1022.

Qian, Y., Wang, L., Wang, C., Gao, X., 2013. The synergy of 3D SIFT and sparse codes for classification of viewpoints from echocardiogram videos, in: Medical Content-Based Retrieval for Clinical Decision Support. Springer Berlin Heidelberg. volume 7723 of *Lecture Notes in Computer Science*, pp. 68–79.

Rahmatullah, B., Papageorghiou, A., Noble, J., 2012. Integration of local and global features for anatomical object detection in ultrasound, in: Medical Image Computing and Computer-Assisted Intervention. Springer Berlin Heidelberg. volume 7512 of *Lecture Notes in Computer Science*, pp. 402–409.

Thrun, S., Burgard, W., Fox, D., 2005. Probabilistic Robotics. The MIT Press.

Wu, H., Bowers, D., Huynh, T., Souvenir, R., 2013. Echocardiogram view classification using low-level features, in: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, pp. 752–755.

Yang, L., Georgescu, B., Zheng, Y., Meer, P., Comaniciu, D., 2008. 3D ultrasound tracking of the left ventricle using one-step forward prediction and data fusion of collaborative trackers, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8.

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

Zhou, S., Park, J.H., Georgescu, B., Comaniciu, D., Simopoulos, C., Otsuki, J., 2006. Image-based multiclass boosting and echocardiographic view classification, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, pp. 1559–1565.