AUTOMATIC DETERMINATION OF THE FETAL CARDIAC CYCLE IN ULTRASOUND USING SPATIO-TEMPORAL NEURAL NETWORKS

Lok Hin Lee and J. Alison Noble

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

ABSTRACT

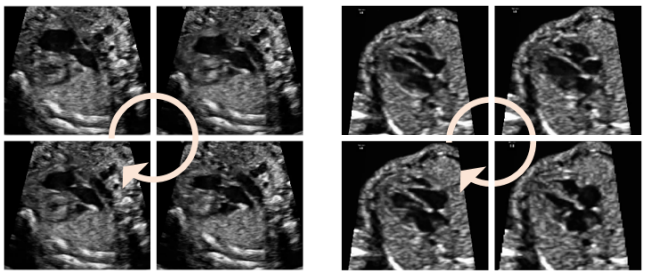
The characterization of the fetal cardiac cycle is an important determination of fetal health and stress. The anomalous appearance of different anatomical structures during different phases of the heart cycle is a key indicator of fetal congenital hearth disease. However, locating the fetal heart using ultrasound is challenging, as the heart is small and indistinct. In this paper, we present a viewpoint agnostic solution that automatically characterizes the cardiac cycle in clinical ultrasound scans of the fetal heart. When estimating the state of the cardiac cycle, our model achieves a mean-squared error of 0.177 between the ground truth cardiac cycle and our prediction. We also show that our network is able to localize the heart, despite the lack of labels indicating the location of the heart in the training process.

Index Terms— Fetal ultrasound, fetal echocardiography, deep learning

1. INTRODUCTION

Congenital heart diseases may be detected in routine monitoring ultrasound scans of the fetus. However, this is difficult, as the sonographer is required to locate the heart and perform diagnosis in real time. Furthermore, ultrasound imagery is subject to significant artefacting in the form of speckle, enhancement and shadowing, which further increase the difficulty of diagnosis. Sonographers therefore rely on standard views in a typical ultrasound screening, where the visibility of certain anatomical structures are fixed. However, the relative location of the fetus and the ultrasound probe is not fixed, due to the mobility of the fetus within the womb. There has therefore been prior work on detecting the overall motion of the fetal heart in video [1] as well as standard view identification using fully convolutional neural networks [2]. Bridge also used a particle-filtering based method to predict heart dynamics, including cardiac phase and view [3].

This work is supported by the Croucher Foundation Croucher Scholarship for Doctoral Study.



(a) (b)

Fig. 1. Two examples of a fetal heart going through a cardiac cycle. As the frame images go clockwise, the heart undergoes systole in the top two frames and diastole in the bottom two frames respectively. (a) shows a heart in the three-vessel view, and (b) shows a heart in the four chamber view.

In this work, we use an end-to-end neural network framework to directly predict the cardiac cycle from the cardiac video for the first time. We use a spatio-temporal model, which takes into account multiple frames of information to estimate the cardiac cycle, and we show that this has superior performance to a spatial convolutional neural network only model. We also show that models trained in this way can be used to weakly localize the heart, despite not having labels of the location of the heart during the training process.

2. METHODS

2.1. Clinical fetal cardiac video data

A dataset of 91 ultrasound videos of the fetal heart was used, drawn from routine clinical scans which were performed using a Voluson E8 ultrasound system. These ultrasound videos were drawn from 12 healthy subjects. Videos varied in length from 2 to 10 seconds, and varied in frame rate from 25 frames per second to 76 frames per second. Each video may contain up to three standard viewpoints, but include significant variations in orientation and magnification as the probe was moved during the clinical scan between standard views. The gestational age of the fetus varied from 20-35 weeks. Frames

Spatial Network / Feature Extractor Network
Input(100x130x3)
CONV(K3, S1, O64), CONV(K3, S2, O64), PL, BN
CONV(K3, S1, O128), CONV(K3, S2, O128), PL, BN
CONV(K3, S1, O256), CONV(K3, S2, O256), BN
FC(2048), FC(1000)
FC(1), Loss

Temporal Aggregation Network
Input(1000x30)
RNN(O1024), Dropout(0.25)
RNN(O512), Dropout(0.25)
RNN(O256), Dropout(0.25)
RNN(O128), Dropout(0.25)
Channel-wise Average, Loss

Table 1. An overview of the network architectures investigated. CONV represents a convolutional layer, BN represents Batch Normalization, PL represents a 2-D Max Pooling operation with a pool size of 2x2, FC represents a fully connected layer, and RNN represents a bidirectional LSTM layer. (K3, S2, O256) implies a layer with a kernel size of 3x3, a stride of 2 and 256 channels. Except for the final layer, all CONV and RNN layers are followed by a ReLU activation.

were resized to be 100 by 130 pixels, and the frame rates of the videos were standardized at 75 frames per second.

The frames containing a heart at maximum contraction and relaxation were then labelled with $y_i = \frac{\pi}{2}$ and $y_i = \frac{3\pi}{2}$ respectively. This was then verified by a clinician experienced in the interpretation of clinical fetal echo videos (Fig. 1). Frames in between were then linearly interpolated between the two labels, and a sinusoidal curve simulating the cardiac cycle was fitted by using $\sin y_i$ and used as the ground truth.

2.2. Network Architecture

We experiment with two network architectures in order to find the architecture that is most suited to this task, an overview of which is provided in Table 1.

We define the first, the purely spatial network as the Spatial Network. This network is purely spatial and frame-based. It does not take into account temporally adjacent frames during the training and inference process.

The second is a spatio-temporal network. The spatial feature extractor network is the same as the spatial network. However, the features from the penultimate fully connected layer FC(1000) are concatenated and further processed in a temporal aggregation network. This temporal aggregation network uses these features and predicts the phase, taking into account features extracted from multiple adjacent frames. In all layers of the temporal aggregation network, we return the hidden state of the network and pass it onto the next layer. The final layer then takes a channel-wise average of the hidden states in order to determine the inferred phase angle.

All networks trained were randomly initialized and trained with the ADAM optimization algorithm with a learning rate of 1e-6 and a batch size of 6. Networks were trained until validation error did not decrease for 10 epochs. K-Fold cross-validation was used during training with folds being set for each patient and the best-performing model was used for evaluation for each fold.

2.3. Data Pre-processing

During the training process, as a method of data augmentation, video clips were subject to random horizontal and vertical flipping, zooming of up to $\pm 20\%$ and rotations of up to ± 20 degrees. We also employ a form of temporal augmentation wherein video clips were sped up or slowed down by up to $\pm 20\%$. The zooming factor is chosen such that the fraction of the image that was taken up by the heart would not exceed half of the frame. The temporal augmentation factor was chosen such that the heart beat frequency that the neural network would encounter during training varied from 100 beats per minute to 190 beats per minute, which would allow the network to be trained on artificial instances of brachycardia and tachycardia, increasing clinical utility [4]. We also precalculate the optical flow between input frames to the neural network and include the calculation as two additional image channels, one for the vertical direction and one for the horizontal.

We investigated the effect of varying the temporal length of the video clip that is used in training the spatio-temporal network. Due to memory limitations, we are limited to 30 input images for the temporal aggregation network; however, we select these thirty frames by selecting {180, 90, 30} consecutive frames from input data and only using every {6, 3, 1} frames during the training and inference process. This represents a clip duration of {2.4s, 1.2s, 0.4s} respectively.

3. RESULTS AND DISCUSSION

3.1. Network Decision Process Visualization

We find that in general, the spatio-temporal neural network performed better than the spatial-only feature extractor network, with an difference in mean squared error of 0.400 between the best performing spatio-temporal network and the spatial-only network.

In order to help explain the difference in performance between the spatial only network and the spatio-

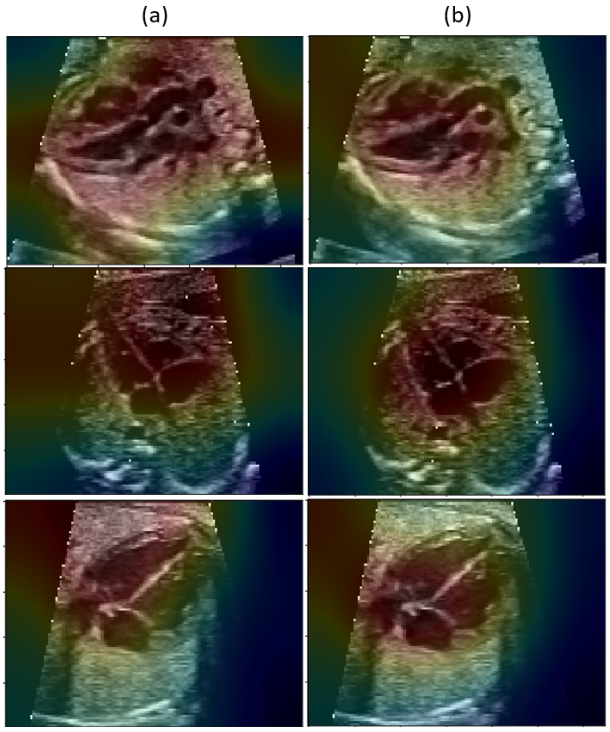


Fig. 2. The figure shows the saliency heatmaps generated by following the gradient flows into the FC(1000) layer for the (a) purely spatial network and (b) the spatio-temporal network (90 frames). We qualitatively find that (b) localizes the heart better during the feature extraction process, despite identical network architectures up to FC(1000), leading to more accurate phase detection.

temporal networks, we use Grad-CAM [5] to generate a heat map to visualize the explanation for the difference in performance. The Grad-CAM algorithm computes the gradients of the final extracted features in the penultimate fully connected layer with respect to the input image, and thereby generates a localization map of the input frame which is important to the final features extracted.

To maintain equality between the heat map comparison between the spatial-only network and the spatio-temporal network, we use the gradients flowing into the penultimate fully connected layer in the spatial extractor network.

Qualitatively, the spatial feature extractor network in the spatio-temporal network appears to perform better at localizing the heart than the spatial-only network (Fig. 2), despite both networks only having access to per-frame level information at this level of the neural network architecture. This may be because of the fact that the spatial feature extractor network is trained in the spatio-temporal network through gradients that have been back-propagated through the temporal aggregation network. This allows gradients to back-propagate with

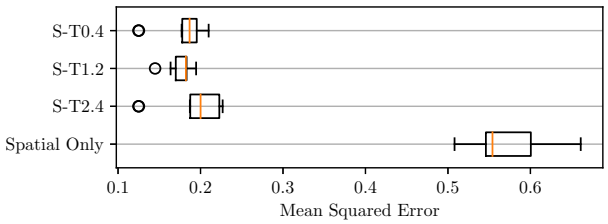


Fig. 3. A comparison of the performance between the spatial model and the spatial-temporal models between all 12 folds. Spatio-Temporal networks had 30 frames of input extracted from a video clip, but the clip duration varied. S-T0.4 denotes a Spatio-Temporal network with a clip duration of 0.4s.

respect to time as well as space, and may therefore aid in the training process. On the other hand, the spatial-only network does not have the temporal aggregation network at the end, and therefore gradients are only back-propagated with respect to each input frame.

3.2. Phase Angle Prediction

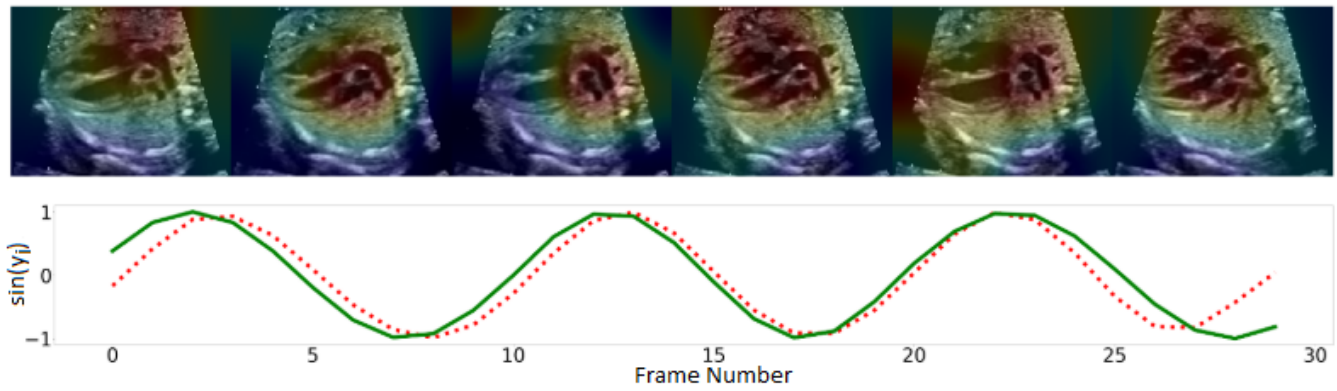
In Table 3, we show the mean squared error with respect to human annotated ground truth from both the spatial-only network and the spatio-temporal networks.

The best results were achieved in S-T1.2 where the network saw a clip duration of 1.2s for each input clip. The outperformance compared to S-T0.4 may be due to the increased number of cardiac cycles seen per input clip. At a fetal heart rate of 100 beats per minute, S-T0.4 would only be able to see approximately 2/3rds of the cardiac cycle per clip, thus hindering performance.

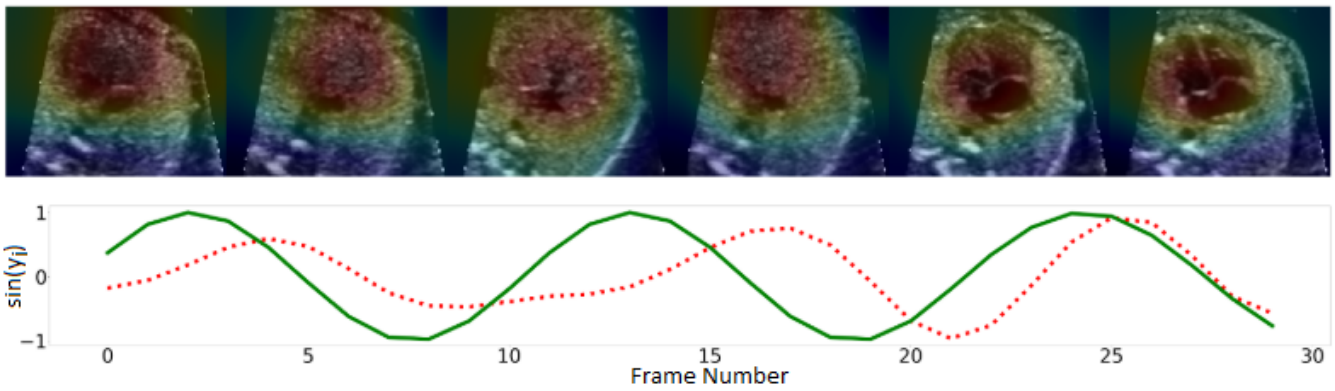
On the other hand, despite seeing more cardiac cycles per input clip than S-T1.2, S-T2.4 would have input images sampled at a rate of 80ms per frame. This may have been too temporally sparse to accurately estimate movement between each frame, which would have been compounded by the freehand probe movement in the routine clinical scan between each frame.

We empirically found that performance on test clips varied, with a majority of clips achieving a low normalized phase error but some clips experiencing high errors. We investigate this difference in the performance of the model by qualitatively evaluating clips where the phase error was high.

We found that in general, clips where the architecture experienced high degrees of phase error were clips where there was significant probe motion relative to the subject, obstructions in the visibility of the heart chambers due to significant ghosting and shadowing, or where the heart chambers were not visible due to probe misplacement. The difference in performance can be seen in Figure 3.2, which is indicative of a type of failure mode in the architecture. When the heart chamber is not clearly



(a)



(b)

Fig. 4. (a) shows an example of a test clip with a mean squared normalized phase error of 0.11. (b) shows an example of a test clip with an mean squared normalized phase error of 0.66. Inferred and ground truth cardiac phase is located below each clip in red (dotted) and green (solid) respectively. In (b), the heart chambers are not visible until frame 24, leading to the failure of the network to infer phase information. A localization heat map is overlaid on each frame, indicating the areas where the spatial feature extractor network is focused on.

visible, phase angle inference fails and the spatial feature extractor network is unable to clearly identify the location of the heart.

4. CONCLUSIONS

This paper describes a novel method of characterizing the cardiac cycle using a spatio-temporal model. We investigated the improvement that the additional temporal element in the architecture brought, and achieved a mean squared error of 0.177. Using a temporal element in the network architecture led to better weakly-supervised training of the localization in the spatial feature extractor network which could be seen in the heart localization heat maps. By localizing the heart and estimating phase, this architecture therefore forms a basis for ultrasound heart characterization. It would be interesting to see if this model could be extended to include the full clinical parameterization of an ultrasound fetal heart scan, including orientation and view, and how it performs on congenital heart cases.

5. REFERENCES

[1] Arijit Patra and J. Alison Noble, "Multi-anatomy localization in fetal echocardiography videos," ISBI, 2019.

[2] Sundaresan et al., "Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks," ISBI, 2017.

[3] Bridge et al., "Automated annotation and quantitative description of ultrasound videos of the fetal heart," MIA, vol. 36, pp. 147–161, 2017.

[4] von Steinburg et al., "What is the "normal" fetal heart rate?," PeerJ, 2013.

[5] Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in ICCV, 2017.