Detection and Characterization of the Fetal Heartbeat in Free-hand Ultrasound Sweeps with Weakly-supervised Two-streams Convolutional Networks

Yuan Gao(✉) and J.Alison Noble

Biomedical Image Analysis Group, Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
Yuan.Gao2@eng.ox.ac.uk

Abstract. Assessment of fetal cardiac activity is essential to confirm pregnancy viability in obstetric ultrasound. However, automated detection and localization of a beating fetal heart, in free-hand ultrasound sweeps, is a very challenging task, due to high variation in heart appearance, scale and position (because of heart deformation, scanning orientations and artefacts). In this paper, we present a two-stream Convolutional Network (ConvNet) -a temporal sequence learning model- that recognizes heart frames and localizes the heart using only weak supervision. Our contribution is three-fold: (i) to the best of our knowledge, this is the first work to use two-stream spatio-temporal ConvNets in analysis of free-hand fetal ultrasound videos. The model is compact, and can be trained end-to-end with only image level labels; (ii) the model enforces rotation invariance, which does not require additional augmentation in the training data, and (iii) the model is particularly robust for heart detection, which is important in our application where there can be additional distracting textures, such as acoustic shadows. Our results demonstrate that the proposed two-stream ConvNet architecture significantly outperforms single stream spatial ConvNets (90.3% versus 74.9%), in terms of heart identification.

Keywords: Two-stream ConvNet, Weakly Supervised Detection, Fetal Heart, Free-hand Ultrasound Video

1 Introduction

Automated detection and characterization of the fetal heart is of great help in diagnosis of Congenital Heart Disease (CHD). However, this task is very challenging in clinical free-hand ultrasound (US) videos, for a number of reasons [1]. The appearance of fetal hearts is highly varied throughout the cardiac cycle, and depends on the pose of the fetus relative to the transducer. Fetal hearts also appear indistinct sometimes, due to variations in imaging contrast, as well as the presence of artefacts (e.g. acoustic shadow). To support the assessment of fetal cardiac abnormalities, there is a clinical need to develop automated recognition tools, which can not only identify and index the different viewing planes of the fetal heart, but can also localize and characterize the fetal heart from the planes in quick succession.

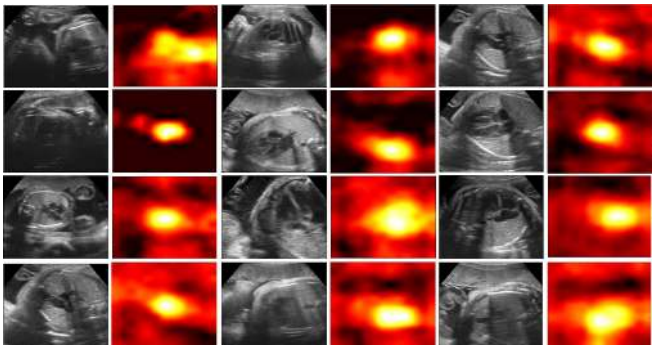


Fig. 1: Localization confidence maps for fetal heart detection of our weakly-supervised two-stream ConvNet. Note our model is robust to variations in fetal heart appearance, scale, position and image contrast.

Contributions: In this paper, we propose an end-to-end trainable two-stream Convolutional Network (ConvNet), inspired by [2] and [3], for fetal heart recognition and characterization in clinical free-hand fetal US videos. The network is fully convolutional, consisting of a temporal stream, extracting motion features, and a spatial stream, extracting appearance features. Two streams are fused to learn spatio-temporal representations as depicted in Fig. 2. We demonstrate that the framework leads to (1) substantial improvement in terms of correct identification of fetal heart frames, compared to spatial ConvNets; and (2) accurate localization (predicts approximate location, as illustrated in Fig. 1) of fetal hearts with only image-level supervision. The resulting output flow maps characterize motion locally and globally in an original way.

Related Work: Most research on fetal heart recognition, e.g.[1] and [7] has been primarily conducted with handcrafted features or referred to as shallow learning. Maraci et al. [7] applied dynamic texture modelling with hand-crafted rotation-invariant feature for detection of a fetal heartbeat. Bridge et al. [1] proposed a framework based on Sequential Bayesian Filtering to predict visibility, position and orientation of fetal heart in consecutive frames. That work is different to our case in two distinct ways; firstly in that scenario the position of the fetal heart was annotated in each frame, whilst our localization of the fetal heart is learnt from image-level labels only; secondly that work used fetal echocardiography video rather than general fetal ultrasound video. Several recent works have used ConvNets for detection of standard planes in fetal US video. Baumgartner et al. [4] employed a fully convolutional network to detect 12 standard planes and localise the respective fetal anatomy. Gao et al. [5] presented a transfer learning based design to study the transferability of features learnt from natural images to ultrasound image object recognition. However, none of these works exploited the spatio-temporal contextual information. In comparison, Chen et al. [6] proposed a hybrid model, composed of ConvNets and recurrent neural network (RNN), to explore spatio-temporal learning from contextual temporal information. However, the performance at fetal heart de-

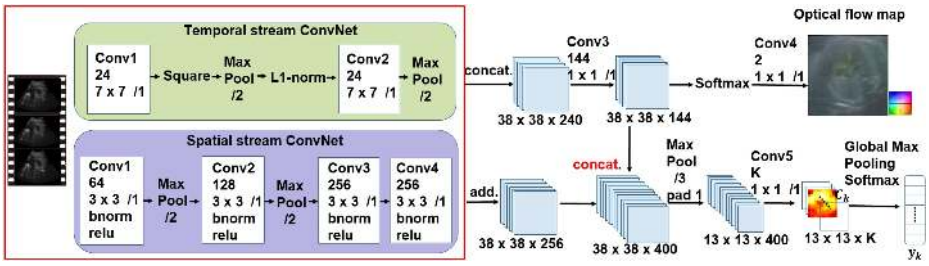
tection was only marginally improved compared to those ConvNets trained with appearance information only. To the best of our knowledge, this is the first paper to investigate using spatio-temporal ConvNets for the detection of the fetal heart in general free-hand fetal US video.

2 Materials and Methods

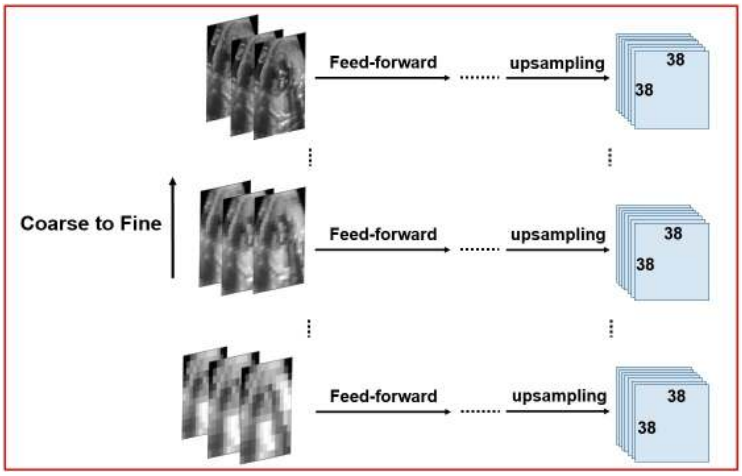
Data and Preprocessing: Our dataset consists of 412 fetal ultrasound videos of healthy volunteers with fetuses of 28 weeks gestation or higher, which have been acquired using a mid-range US machine (Philips HD9 with a V7-3 transducer) by an experienced obstetrician moving the ultrasound probe from the bottom to top of the abdomen in approximately 6-8 seconds. We created video clips with different lengths from each video, which had been annotated with the four class ground-truth labels i.e. fetal abdomen, heart, skull and other structures. Other structures included any video frames that did not fall into the other three classes. The videos clips were sub-sampled to create sequences of three consecutive frames with a step size determined by the length of each video. 15% of the videos were randomly chosen as test data. The rest were used in training (80%) and validation (20%). All video frames had an original size of 240×320 pixels and were cropped at center to produce 152×152 square patches. To increase robustness to changes in brightness and contrast, we preprocessed each video frame by first subtracting a Gaussian smoothed version of the image and then applying standard normalisation with a sliding disk mask (i.e. subtract mean intensity of each frame and divide it by standard deviations obtained locally).

Network Architecture: The architecture of our two-stream ConvNet is summarized in Fig. 2. As depicted in Fig. 2a, the temporal stream implements a shallow Fully Convolutional Networks (FCN) [3] that takes a preprocessed stack of three consecutive frames of a video as input, and extracts high dimensional motion features, which are then projected as 2D optical flow maps (spatially 4x downsampled). **Conv1** in the temporal ConvNet approximates spatio-temporal Gabor filters with a number of orientations that respond to patterns moving at different speeds. A local phase-invariant response is approximated by element-wise squaring rectification and 2 x 2 max-pooling with stride of 2. Channel-wise L1-Norm is then applied to the output to account for intensity variance of patterns moving at different orientations. **Conv2** in the temporal stream is introduced to approximate the smooth regularization, which penalizes high variations (large temporal gradients) by taking neighbouring pixels into account. Decoding is performed by **Conv3** layer using 1×1 kernels, which decodes features into 144 channels representing scores at different speeds and orientations. Then, Softmax is computed at each spatial location, across the 144 feature channels, to constitute a distributed representation of motion. **Conv4** linearly projects the distributed representation to a 2D optical flow map. The spatial stream adopts the configuration of VGG very deep ConvNets (11 weights layer) [8]. All convolutional layers in the spatial ConvNet consist of 3×3 kernels, batch normalization (bnorm) [9] and Rectified Linear Units (ReLUs) non-linearity [10]. The size of feature maps are spatially matched throughout the feed-forward processing between the spatial and temporal stream. Red **concat.** indicates where two

streams are fused by stacking their feature maps (normalized to [0,1]) together. The **Conv5** layer performs a convolutional fusion which projects the concatenated spatio-temporal representation to **K** classes score maps (denoted by **C**). Instead of using fully connected (FC) layers, we use a *Global Max Pooling* layer to obtain the vector for *Softmax* classification.



(a) Network Architecture. The numbers in each layer (e.g. 64 and 3 ×3 in Conv1 refer to the number of kernels and their size, /1 denotes a stride of 1,and the size of max-pooling window is equivalent to its stride (e.g./2 denotes window size 2×2, pooling with a stride of 2). **concat.** denotes stacks the feature maps at the same spatial locations. **add.** denotes the sum of the feature maps at the same spatial locations.



(b) Coarse to Fine Augmentation. The network enclosed by red box applied on multiple down-sampled versions of input frames.

Fig. 2: Overview of our two-stream ConvNet architecture.

Multi-scale Learning: The fully convolutional architecture makes it flexible with regard to the size of input frames. We trained our two-stream ConvNet with a multi-scale scheme as described in 2b. All input frames are first rescaled by a factor $s \in \{0.03, 0.05, 0.07, 0.09, 0.14, 0.2, 0.25, 0.33, 0.5, 1\}$. We apply the network enclosed by the red box (in Fig.2a) on the multiple downsized versions of the input frames. Feature maps of the down-sampled versions are brought back to the common resolution i.e. 38×38 by bilinear up-sampling, and then

concatenated or added together. As both the spatial and temporal ConvNets are shallow, the detection of features is limited by its small effective receptive field. Therefore, the multi-scale processing is used to compensate for this. In addition, the scale of fetal hearts can vary significantly in free-hand scans. The multi-scale scheme includes some scale-invariance in the network.

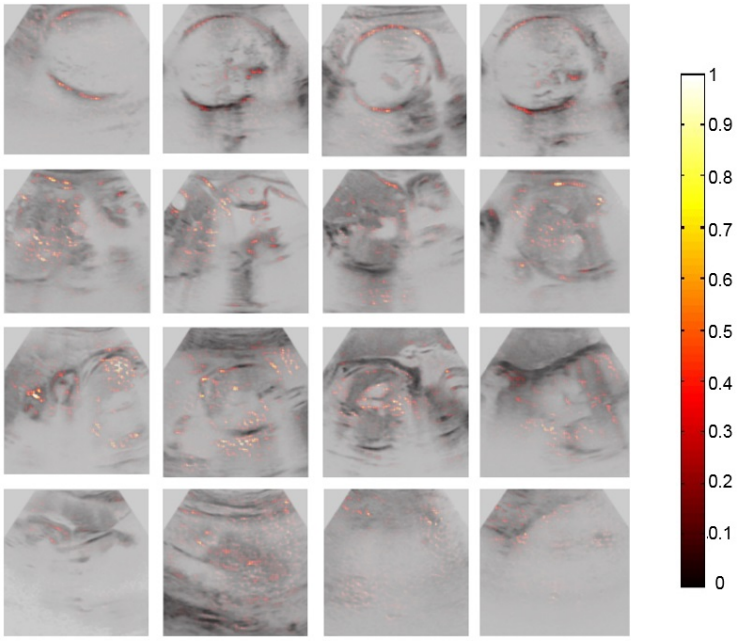
Training: We trained the proposed model for detecting not only fetal hearts, but also the fetal skull and the fetal abdomen (plus background resulting K=4) with a short temporal window of three consecutive frames. Models were trained using min-batch (batch size = 52) Stochastic Gradient Descent (SGD) with momentum (momentum rate 0.9 and weight decay 0.0005). To cope with imbalanced training classes, we reweight the Softmax loss with the reciprocal of class frequency in the training set such that learning gives more attention to minority classes. We also added 50% dropout after the *Conv5* layer to prevent overfitting. The spatial network is initialised with a VGG very deep (11 weights layer) architecture, pre-trained for classification of ultrasound images. The temporal network was pre-trained to extract optical flow with two stages of learning. The network was first trained for classification with a Softmax loss over the 144 decoded feature channels at each spatial location. Ground truth labels for the training are generated from flow vectors ground truth by performing a nearest neighbour clustering. The network is fine-tuned for flow regression by minimizing Euclidean end-points error loss over the network output. We enforced rotation invariance in the temporal network by learning only a subset of weights in Conv1 and Conv2, and rotated them to get a full set of filters. This eliminates the need for rotation augmentation in the training data when learning for fetal heart detection.

3 Experiments and Results

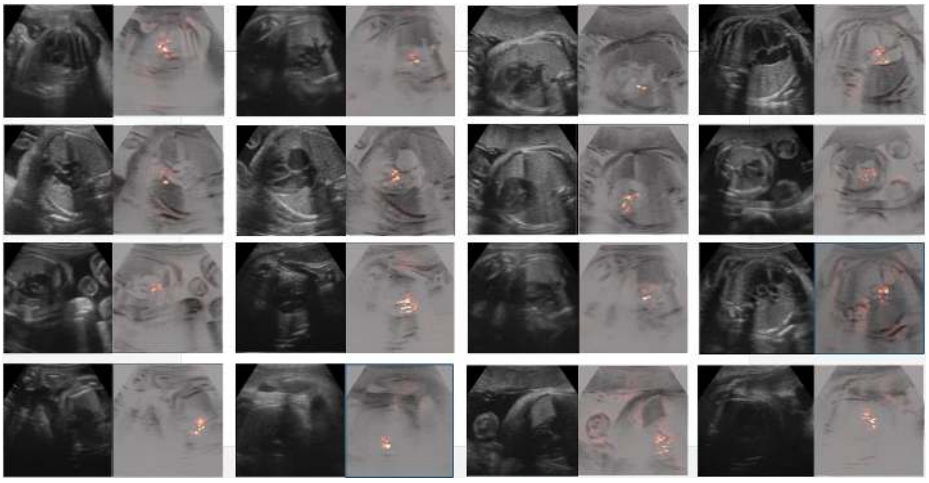
We trained the proposed two-stream ConvNet model for 20 epochs with initial learning rate of 0.01 and continuously reduced to 10^{-4} until training termination. We evaluated the performance of the model in terms of per-class classification accuracy, precision and recall scores respectively (as summarized in Table 1). We report the best performance achieved on the test dataset. In addition, we also trained several state-of-the-art ConvNet architectures for classifying single frames. These single stream architectures were truncated, in which only single FC layer was used with a reduced number of neurons 4096→1024, and the training was carefully regularized with bnorm and dropout to avoid overfitting. We compare the performance of our proposed two-stream model with these modified state-of-the-art models in Table 1. We found that our two-stream ConvNet model achieved better classification performance for most performance metrics compared to single-stream models. Particularly, two-stream Convnet fetal heart detection outperformed those of single-stream architectures by a significant margin, which demonstrated the superiority of exploring temporal context information from consecutive frames by learning spatio-temporal representation.

Table 1: Classification performance for our proposed two-stream ConvNet and its comparison with other single-stream spatial ConvNets. **A, P, R** denote accuracy, precision and recall respectively.

Methods	Fetal skull			Fetal abdomen			Fetal heart (all views)		
	A	P	R	A	P	R	A	P	R
Two-stream ConvNet (as proposed)	0.974	0.911	**0.946**	0.937	**0.906**	**0.893**	**0.903**	**0.853**	**0.892**
AlexNet [10] truncated	0.921	**0.937**	0.913	0.910	0.851	0.877	0.683	0.575	0.651
CNN-M [11] truncated	0.946	0.903	0.930	0.896	0.803	0.825	0.636	0.612	0.534
CNN-S [11] truncated	0.881	0.852	0.841	0.834	0.803	0.816	0.612	0.537	0.604
VGG-VD-A [8] truncated	0.922	0.824	0.831	0.942	0.751	0.862	0.571	0.552	0.513
TCNN [5]	**0.980**	—	—	**0.959**	—	—	0.749	—	—



(a) Spatial ConvNet TCNN[5]. Row 1 to 4 correspond to fetal skull, abdomen, heart and background respectively.



(b) Localised fetal hearts in the two-stream spatial-temporal ConvNet as proposed.

Fig. 3: Localization Saliency Maps of fetal structures.

Localization Saliency: As demonstrated in Fig. 1, we make an approximate prediction of the location of fetal hearts by bilinear up-sampling the class-specific score maps $\mathbf{C}_k$. To have a more precise localisation, we calculate how much each original input pixel $\mathbf{I}_{ij}$ contributes to the activation of the class-specific score in the final output, and obtained localised saliency maps $\mathbf{M}_{ij}$

with the resolution of the original input frames. The maps are obtained by computing the partial derivatives $M_{ij}^k = \frac{\partial y_k}{\partial I_{ij}}$ through backpropagation[12]. Particularly, we adopts the guided backpropagation method, proposed in [13], which backpropagates gradients through ReLU activated neurons (where the inputs are positive), as well as the gradients are positive. This is expressed as $\frac{\partial y_k}{\partial X_n} = \frac{\partial y_k}{\partial X_{n+1}} \cdot \mathbf{1}\left(X_n > 0\right) \cdot \mathbf{1}\left(\frac{\partial y_k}{\partial X_{n+1}} > 0\right)$, where $\mathbf{1}(\cdot)$ is the element-wise indicator function, x_n and x_{n+1} are the input and output of ReLU activation respectively. To evaluate the localisation performance, Fig.3 highlights the localised saliency maps of different fetal structures, particularly the two-stream ConvNet localised fetal hearts in Fig.3b. Hotspots on the maps indicate pixels that contribute to the activation of class-specific score. The brighter the hotspots, the greater the contribution. Of note, the activation hotspots in fetal heart images (Third row in Fig.3a) are highly random in single-stream ConvNet, which implies the network learned lots of noise. This may explain why spatial ConvNets generalize poorly in detection of fetal heart frames. In contrast, by incorporating temporal contextual information, our two-stream ConvNet localizes fetal hearts well, even in some very hard situations. The last row in Fig.3b, for example, shows very low contrast of fetal hearts, in which cardiac structures are almost invisible. However, the model still picks up the location of the fetal heart very well. In addition, the model is robust to the variation of appearance, scale, and location of fetal hearts, and also artefacts like acoustic shadows, as can be seen in Fig.3b.

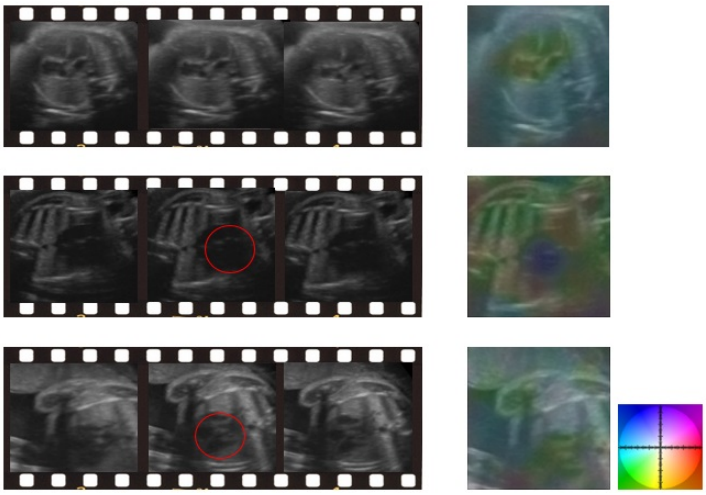


Fig. 4: Characterisation of fetal heart motion.

Characterization of heartbeats: Fig.4 illustrates the optical flow maps extracted from the network, which characterise motions locally and globally (such as probe moving). The flow maps are color-coded to show the direction of motion,

in which darker colors indicate a larger motion speed. The flow maps provide visualisation of fetal heart function locally. Even in low contrast cases (like the last two examples), the model can still sense the motion of the heartbeat.

4 Conclusion

We have demonstrated that temporal information of consecutive sequences in US videos can provide contextual clues for better discrimination in recognition of the fetal heart in general freehand ultrasound sweeps. Particularly, our proposed model enforces rotation-invariance in spatio-temporal representation, which eliminates the requirement of rotation augmentation. Furthermore, the model has been shown to provide robust localisation and characterization of fetal hearts, in the cases, where there are low contrast and distracting artefacts.

Acknowledgments. The authors acknowledge the China Scholarship Council (CSC) for Doctotral Training Award (grant No. 201408060107) and the RCUK CDT in Healthcare Innovation.

References

1. C.P. Bridge, C. Ioannou, J.A. Noble.: Automated annotation and quantitative description of ultrasound videos of the fetal heart. Medical Image Analysis (36) 147-161. 2017.
2. S. Karen, and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.
3. T. Damien, and M. Hebert. Learning to Extract Motion from Videos in Convolutional Neural Networks. In CVPR, 2016.
4. C.F. Baumgartner et al. Real-Time Standard Scan Plane Detection and Localisation in Fetal Ultrasound Using Fully Convolutional Neural Networks. In MICCAI, 2016.
5. Y. Gao, M.A. Maraci, J.A. Noble. Describing ultrasound video content using deep convolutional neural networks. In ISBI,2016.
6. H. Chen et al. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In MICCAI, 2015.
7. M.A. Maraci et al. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. Medical Image Analysis (37) 22-36. 2017.
8. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2014.
9. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR, 2015.
10. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.
11. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Networks. In BMVC, 2014.
12. K. Simonyan, A. Vedaldi, A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ICLR Workshop, 2014.
13. J. Springenberg et al. Striving for simplicity: The all convolutional net. ICLR Workshop, 2015.