

Learning Spatio-Temporal Aggregation for Fetal Heart Analysis in Ultrasound Video

Arijit Patra, Weilin Huang, and J. Alison Noble

Institute of Biomedical Engineering, University of Oxford, UK,
`arijit.patra@eng.ox.ac.uk`

Abstract. We investigate recent deep convolutional architectures for automatically describing multiple clinically relevant properties of the fetal heart in Ultrasound (US) videos, with the goal of learning spatio-temporal aggregation of deep representations. We examine multiple temporal encoding models that combine both spatial and temporal features tailored for US video representation. We cast our task into a multi-task learning problem within a hierarchical convolutional model that jointly predicts the visibility, view plane and localization of the fetal heart at the frame level. We study deep convolutional networks developed for video classification, and analyse them for our task by looking at the architectures and the multi-task loss in the specific modality of US videos. We experimentally verify that the developed hierarchical convolutional model that progressively encodes temporal information throughout the network is powerful to retain both spatial details and rich temporal features, which leads to high performance on a real-world clinical dataset.

1 Introduction

Fetal ultrasound (US) is a standard part of pre-natal care primarily due to its non-invasive nature and the unsuitability of other imaging modalities through the length of pregnancy. Understanding fetal cardiac screening US videos is important to diagnose congenital heart disease. Fetal heart conditions are often missed because of factors like the shortage of trained clinicians and the requirement of expertise and equipment for fetal cardiac screening which causes it to be excluded from compulsory screening requirements in the 20-week abnormality scans. Precisely analysing fetal cardiac US videos is a challenging task even for human experts due to the complex appearance of different anatomical structures in a small region. Furthermore, there exist speckle, shadowing, enhancement and variations in contrast levels in US images. In addition, the clinician has to perform multiple activities during a typical US scan such as viewing plane identification, anomaly detection, gender identification and so on. In this work, we analyse the scope for using deep learning techniques with US videos for describing key parameters of the fetal heart, with a focus on aggregating spatio-temporal features within deep convolutional architectures.

Convolutional neural networks (CNNs) have been successfully applied to many computer vision tasks. Various CNN models have been developed recently

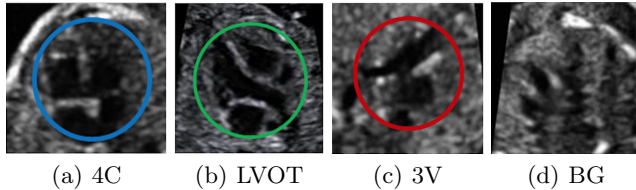


Fig. 1: Three different view planes of fetal heart: the four chamber (4C), the left ventricular outflow tract (LVOT), the three vessels (3V), and the background (BG).

for action recognition, with the goal of learning spatio-temporal features from videos [5, 9, 8]. These models were mainly designed for video-level classification with a single action category assigned to a whole video. High-level global context information is important for the task of video classification, while our task is more challenging by jointly considering both classification and localization at the frame level, where detailed local information is vital for precise prediction of multiple parameters of the heart. Furthermore, action recognition is mostly built on RGB domain where a single image can include strong visual information for identifying an action category. By contrast, object structure in an US image is often weakly defined, and local temporal dynamics are critically important for producing a reliable prediction.

We approach the problem of automated analysis of the fetal heart by formulating it as a multi-task prediction, where the visibility, view plane and localization of the heart are jointly predicted at the frame level. We study the problem of aggregating spatio-temporal information in a short US video clip by investigating various temporal connectivities within the convolutional architecture. We examine the impact of temporal information on the domain-specific task of fetal heart analysis in US videos. Our main contributions are summarized below:-

- We cast the problem of fetal heart analysis as a multi-task prediction in a hierarchical convolutional architecture that progressively encodes temporal information throughout the network. This is vital to retaining both spatial details and meaningful temporal patterns, which are key to accurate and automatic heart description in fetal US videos.
- We investigate multiple temporal encoding architectures that learn a strong spatio-temporal representation tailored for US video representation. We study these approaches by analysing the architecture, the specific US image modality, and the loss functions designed for joint classification and localization.
- We conducted experiments on a real-world clinical dataset. The results suggest the ability to encode temporal information, which is of particular importance for our task, where image-level information is relatively weak and can be ambiguous.

1.1 Related Work

Deep learning approaches have been recently applied for action recognition in videos, with the focus on learning spatio-temporal information efficiently and

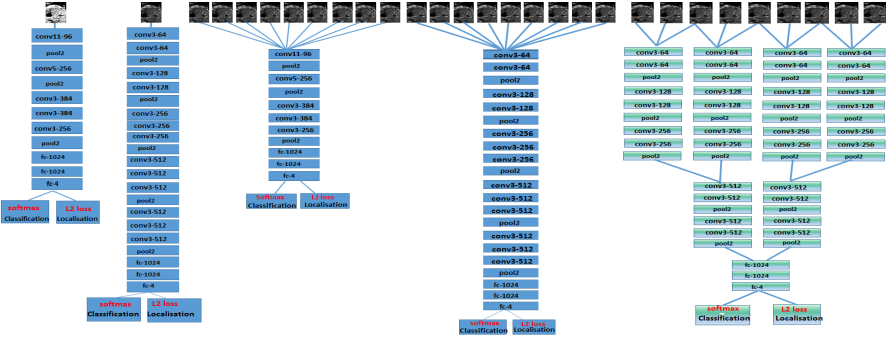


Fig. 2: CNN architectures for multi-task fetal heart prediction in US videos. Left: spatial baseline models based on AlexNet (SBM_{alex}) and VGGnet (SBM_{vgg}). Middle: Direct Temporal Encoding (DTE) models modified from AlexNet and VGGnet. Right: the developed Hierarchical Temporal Encoding (HTE) that progressively incorporates spatio-temporal information throughout the network.

accurately. Karpathy *et. al.* [5] developed a number of deep fusion methods that encode temporal features within CNN architectures. Similarly, various pooling methods that aggregate spatio-temporal CNN features, were investigated in [8] for action recognition. Two-stream CNNs were introduced in [9], where short-term temporal information is captured by using a separate CNN processing on optical flow. These methods were developed for action recognition in video clips where the CNNs were designed to compute global high-level context from sequences of RGB images. Our work focuses on aggregating spatio-temporal information for our US video analysis, and our task requires local detail in both the spatial and the temporal domains for joint classification and localisation.

Recent work on automatic US video analysis mainly focuses on image-level classification of anatomical structures, e.g., [7]. CNNs have also been applied to this task with transfer learning [3] and recurrent models [2]. Fetal heart description is a more complicated application that jointly predicts multiple parameters of the fetal heart. Our work is closely related to recent work in [1] and [4], where multiple properties of the fetal heart were estimated. In [1], a CRF-filter along with hand-crafted features were developed for predicting the defined parameters of the fetal heart. We build on advanced deep learning methods, which allow our model to learn stronger deep representations in an end-to-end fashion. In [4], a Temporal HeartNet was proposed by using a recurrent model to encode temporal information. In the current paper, we focus on aggregating spatio-temporal features directly within a CNN architecture, setting it apart from [4].

2 Fetal Heart Description with Spatio-Temporal CNNs

Our task is to jointly predict the visibility, view plane, location of the fetal heart in US videos, which are similar to [1]. The visibility relates to the presence or absence of the heart in each frame (with partial visibility of less than 50% being deemed as not visible). When the heart is present, the view plane can be either the four chamber (4C) view, the left ventricular outflow tract (LVOT), or the

three vessels (3V) view, as shown in Fig. 1. By jointly considering the visibility and the view-plane identification, we include a background class to define a 4-class classification problem. The location of the heart is defined by its center and radius, which can be cast as a 3-parameter regression problem.

In this work, both classification and localisation are formulated jointly as a multi-task learning problem within the convolutional architecture. The outputs of the two tasks are connected to the last fully-connected (FC) layer of a CNN, such as AlexNet [6] or VGGnet [10], as shown in Fig. 2. A cross-entropy loss with a softmax function is used for the classification task, while a l_2 loss is applied for regressing the values of heart centre and radius. These configurations customize a regular CNN to our task, and form our baseline spatial architecture where the input is a single US image. Our temporal models are extended from this baseline architecture by allowing input of a sequence of continuous frames. To investigate the model capability for learning spatio-temporal details, we extend recent temporal encoding methods, originally introduced to action recognition in [5], to multi-task prediction of the fetal heart in US videos.

2.1 Spatial Baseline Models

For the first step of creating spatial baseline models, we leverage the classical AlexNet [6] and the 16-layer VGGnet [10], both of which are modified to accept a single channel gray-scale image. The two spatial baseline models are referred to as SBM_{alex} and SBM_{vgg} respectively. The networks accept as inputs an US image with size of 224×224 pixels. Details of the architectures are presented in Fig. 2 (left). Notice that the number of neurons in the FC layers is changed from the original 4096 to 1024. This modification was made to avoid overfitting, as the number of predicted parameters in our task is significantly smaller than the 1000 categories of object classification in ImageNet. The last FC layer is connected to the output layer which computes the softmax probabilities of 4 classes (3 view planes and the background), with the predicted heart center and radius.

2.2 Direct Temporal Encoding

To compute temporal information, a straightforward approach is to modify the CNN architecture to allow for an input of multiple contiguous frames. Such a multi-frame input naturally includes temporal context which is important to produce a more reliable prediction. As the size of the CNN architecture is fixed by design, and the number of CNN parameters should be scalable, the input of the CNNs is configured and fixed to a short, fixed-length US video clip cropped from temporally contiguous frames. In experiments, we use an 8-frame video clip as input for all temporal models where the labels of the 4th frame are applied.

For model configurations, the main idea is to extend network connectivities in the time domain, allowing these connectivities to automatically learn temporal dynamics between contiguous US images. To this end, we extend our spatial baseline models to Direct Temporal Encoding (DTE) by introducing direct temporal filters with kernel size of $k \times k \times T$ in the first convolutional layer, where

T is the temporal dimension which is set to $T = 8$ in our experiments. $k \times k$ is the spatial size of convolutional kernels which is identical to those of the spatial baseline models, e.g., 11×11 and 3×3 for the AlexNet and VGGnet respectively. These temporal filters are able to directly encode both spatial details and temporal connections of the continuous US frames. This enhances the low-level representation in the first convolutional layer, which in turn leads to stronger deep high-level representations by propagating them throughout the networks.

2.3 Hierarchical Temporal Encoding

The DTE approach directly computes local temporal features at the pixel level, and is powerful enough to capture detailed low-level motion characteristics. However, CNN predictions are built on high-level deep representation (e.g., the FC features). It is critically important to investigate: (i) how this low-level temporal information is propagated efficiently and accurately to the final deep representation; and (ii) which level of the temporal information is crucial to the final prediction. Since US images often contain relatively weak visual information at the pixel level, temporal representation immediately computed at the first layer by the DTE may not be robust, and multi-layer propagations throughout the whole network may lead to a certain degree of information loss. The need is to develop a new temporal encoding approach that allows the temporal details of US videos to be computed more accurately and robustly, and to be propagated more efficiently throughout the network without significant information loss.

With these considerations, we have developed a new Hierarchical Temporal Encoding (HTE) approach, which is extended from the slow fusion model developed for action recognition in [5]. We present a number of technical improvements that elegantly tailor it towards our task of fetal US description: (i) we incorporate multi-task prediction into this hierarchical architecture for jointly estimating the view plane and localization of the fetal heart. (ii) and, more convolutional layers using small kernel size of 3×3 are applied, resulting in a deeper model able to capture more spatio-temporal details which are particularly important to our task. These design features aim to capture key characteristics of US video interpretation to encode the temporal domain and the spatial domain, and a finer feature extraction from the previous layer is more desirable to propagate the features throughout the network.

Details of the HTE model are presented in Fig. 2 (right), where frames from the video clip are gathered in groups of 3, and then fed into the first layers of multiple sets of CNNs which share the same parameters. These serve as quasi-independent CNNs until the 3^{rd} pooling layer, where the outputs of the 3^{rd} pooling layer are taken in groups of 2, and further fed into separate sets of 2 CNN structures also sharing the same parameters. This is maintained until the final encoding happens just prior to the first FC layer, and subsequent FC layers effectively process the overall information from the video. Therefore, a slow and progressive combination of the learned features enables the propagation of the diversity of information encoded in the temporal space.

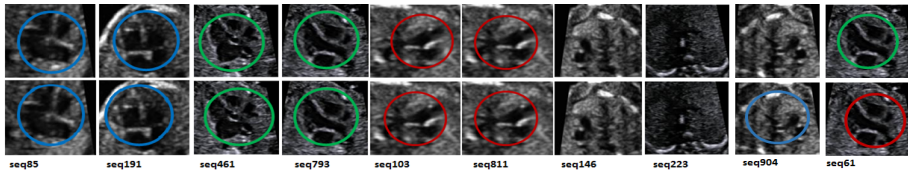


Fig. 3: The predicted results by the HTE (bottom) with GT (top). Color for different view planes.

3 Implementation Details

Real-world clinical dataset. Our dataset consisted of 91 routine clinical cardiac screening videos from 12 subjects at gestational ages ranging from 20 and 35 weeks. Each video had a duration of between 2 and 10 seconds and a frame rate between 25 and 76 frames per second (39556 frames in total). It contained one or more of the three views of the fetal heart. Videos from 10 subjects were used for training, and the remaining 2 for test.

Data pre-processing. For the training step, we split available videos into frames and applied data augmentation by performing an up-down and a top-bottom flipping. Then, we chose 8 frames at a time in the sequence of their original occurrence in the videos and stitched them back into 8-frame sequences and repeated the same for the augmented data. This enabled us to obtain 9337 video clips with roughly equal samples from different view classes. The current frame is the 4th frame in a 8-frame video clip. All frames in the dataset of size 430 x 510 were cropped into 224 x 224 centered about the heart center.

Training details. All new layers across the networks used are initialized by using random weights. Our spatial baseline models are trained by using a 25-frame mini-batch with a learning rate of 0.01. The DTE models are trained with 8-frame chunks of videos using a batch size of 25 with a learning rate of 0.01. A batch size of 20 and a learning rate of 0.001 are used for the HTE.

4 Experimental Results and Comparisons

We performed multiple sets of experiments for both the classification and regression tasks. We start with the generation of a strong feature-based baseline using single frames of video sequences, followed by temporal encoding models, which are then augmented by a hierarchical transfer of spatio-temporal details through the network. We analyse our models for multi-task learning from shared features, achieved by focussing on requisite regions of interest for each task.

Method	Classification				Overall	Localization
	4 Chambers	LVOT	3 Vessels	Background		
<i>SBM_{alex}</i>	89.22	67.60	59.11	90.36	78.95	69.71
<i>SBM_{vgg}</i>	90.84	69.83	62.07	91.37	80.73	72.29
<i>DTE_{alex}</i>	88.68	73.18	57.63	88.83	79.16	72.21
<i>DTE_{vgg}</i>	91.11	70.95	67.00	89.34	81.78	77.26
<i>HTE</i>	81.67	88.27	79.80	86.80	83.58	79.68

Table 1: Performances on real-world fetal cardiac screening videos, in accuracy rates (%). A correct localization is defined with an Intersection-over-Union ratio over 0.25.

Our models are evaluated on real-word fetal cardiac screening videos. Results on several exemplar frames are shown in Fig. 3, and full results are compared in Table 1, where temporal models obtain better performance, particularly on classification. Importantly, the improvements on the challenging classes are remarkable: $67.60\% \rightarrow 88.27\%$ for the LVOT, and $59.11\% \rightarrow 79.80\%$ for the 3 Vessels. The HTE outperforms DTE models by over 15% in both classes, which suggests that the HTE provides a more principled method for encoding detailed and meaningful spatio-temporal features for fetal heart description. This can be further verified by the confusion matrices shown in Fig. 4. The advantage of the HTE is also affirmed by gains seen in the localisation task. We compare our models with that in [1] which uses hand-crafted features and a CRF-filter. The IoU implementations of HTE achieve localisation errors of 20.32% compared to the best result of 34% in [1]. Our models explore stronger deep spatio-temporal representations learned end-to-end with features shared over multiple tasks. Furthermore, our model is able to predict both the centre and radius for heart localisation, while the radius is used as a strong prior information and only the centre is estimated in [1]. Finally, we present further analysis of spatio-temporal aggregation of US representations throughout the CNN architecture. The performance of the direct encoding architecture, DTE_{alex} , can approach that of the deeper baseline model SBM_{vgg} with VGGnet. This is indicative of the dynamical interplay between competing factors governing the ability of deep networks to optimise learning of spatial and temporal information. While an information fusion at the level of extracted features at different time-steps or frames could capture significantly more contextual detail than single frame approaches, the aggregation may cause an overlap between varying spatial features which tends to compromise the ability to learn distinct spatial features. This effect is pertinent to the ultrasound modality because of the non-trivial influence of imaging artefacts, shadows and speckle. Particularly, a pixel-level combination before the first convolution layer, as described in the DTE structures, tends to undermine the frame-wise variations in spatial information by directly fusing along the temporal direction. This trade-off is indirectly addressed in the HTE model, by feeding a relatively smaller temporal extent per convolutional network, and by aggregating multi-level spatio-temporal information progressively throughout the CNN architecture. This provides a more principled approach for spatio-temporal encoding that enables the final deep representation to essentially capture the entire temporal regime of the video segment. Thus, our HTE networks are able to achieve significant gains in accuracy for both the classification and localisation tasks. It is to be noted that the optimal number of frames to achieve the best gains in this trade-off would exhibit significant variations as a function of the modality, the size of datasets and the complexity of features in spatial and temporal domains.

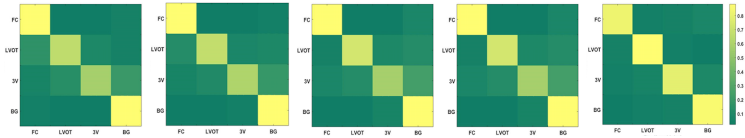


Fig. 4: Confusion matrices for classification. L to R: SBM_{alex} , SBM_{vgg} , DTE_{alex} , DTE_{vgg} , HTE .

5 Conclusions

We have presented and compared multiple CNN models for jointly predicting the visibility, view plane and localisation of the fetal heart in US videos. We developed a Hierarchical Temporal Encoding (HTE) with a multi-task CNN that allows spatio-temporal information to be aggregated progressively through the network. This approach is shown to efficiently retain meaningful spatial and temporal details which are critical to implementing multi-task prediction in US video. We also investigated different temporal encoding models on a real-world clinical dataset, where the proposed HTE achieved significant performance gains.

Acknowledgments. This work was supported by the EPSRC Programme Grant Seebibyte (EP/M013774/1). Arijit Patra is supported by the Rhodes Trust.

References

1. Bridge, C.P., Ioannou, C., Noble, J.A.: Automated annotation and quantitative description of ultrasound videos of the fetal heart. *Medical Image Analysis* 36, 147–161 (2017)
2. Chen, H., Dou, Q., Ni, D., Cheng, J.Z., Qin, J., Li, S., Heng, P.A.: Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks (2015), in *MICCAI*
3. Gao, Y., Maraci, M.A., Noble, J.A.: Describing ultrasound video content using deep convolutional neural networks (2016), in *ISBI*
4. Huang, W., Bridge, C.P., Noble, J.A., Zisserman, A.: Temporal heartnet: Towards human-level automatic analysis of fetal cardiac screening video (2017), in *MICCAI*
5. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks (2014), in *CVPR*
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks (2012), in *NIPS*
7. Maraci, M.A., Napolitano, R., Papageorgiou, A., Noble, J.A.: Searching for structures of interest in an ultrasound video sequence. *Medical Image Analysis* 37, 22–36 (2017)
8. Ng, J.Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification (2015), in *CVPR*
9. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos (2014), in *NIPS*
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), in *ICLR*