Temporal HeartNet: Towards Human-Level Automatic Analysis of Fetal Cardiac Screening Video

Weilin Huang, Christopher P. Bridge, J. Alison Noble, and Andrew Zisserman

Department of Engineering Science, University of Oxford

Summary

- Automated analysis of fetal cardiac ultrasound screening videos
- Multi-task deep architecture jointly identifies heart presence, location, orientation, view plane
- Intersection-over-union loss (IoU) found to give superior localisation compared to anchor mechanisms
- Recurrent bi-directional LSTM layers capture *region-level* temporal context

Background

2D ultrasound screening is the clinical standard for antenatal detection of congenital heart disease (CHD), which encompasses a large range of abnormalities of the developing heart. Screening is performed during routine scans, typically by a non-specialist in fetal cardiology.

However, it is challenging due to the need to find multiple anatomical views and check for multiple anomalies in a time constrained setting. This can be further complicated by other factors such as poor image quality, imaging artefacts, fetal motion, and/or unfavourable fetal lie.

Aim

The aim is to track key variables automatically to put a 'global coordinate system' on the video:

- Heart Visibility, $h_t \in \{0, 1\}$
- Heart Centre Position, $\mathbf{x}_t \in \mathbb{R}^2$
- View Label, $v_t \in \{4C, LVOT, 3V\}$ (Fig. 1)
- Heart Orientation, $\theta_t \in [0, 2\pi)$
- Heart Radius, $r_t \in \mathbb{R}^+$

This information can be fed back to sonographers, used for quality control and audit (e.g. to ensure that all views have been observed), and represents a crucial first step in automatic and diagnosis of CHD.

View Plane Definitions

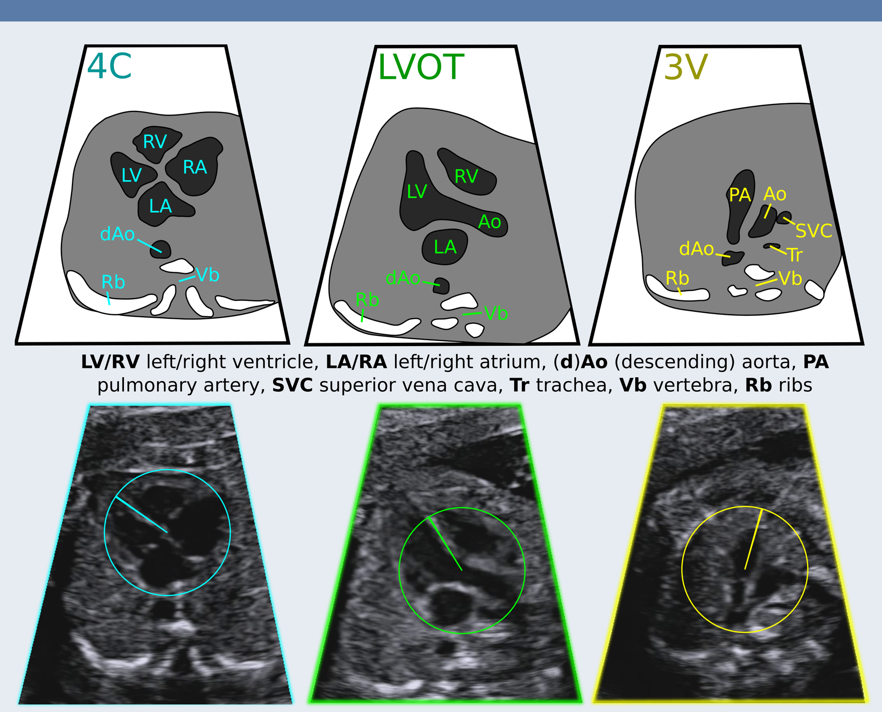

Figure 1: Definition of viewing planes

Architecture Overview

The Heartnet architecture (Fig. 2) consists of:

- **Convolutional layers**: Standard VGG-16 for feature extraction [1]
- **Recurrent layer**: Bi-directional long short-term memory cells (BLSTM) for each 3×3 spatial location in the final convolutional layer capture local temporal context (Fig. 4).
- **Multi-task output layer**: Jointly predicts:
  - Location
  - View plane category
  - Orientation

for each 3×3 spatial location. We experiment with two alternative architectures for this task: a **circular anchor** architecture and an **IoU** architecture.

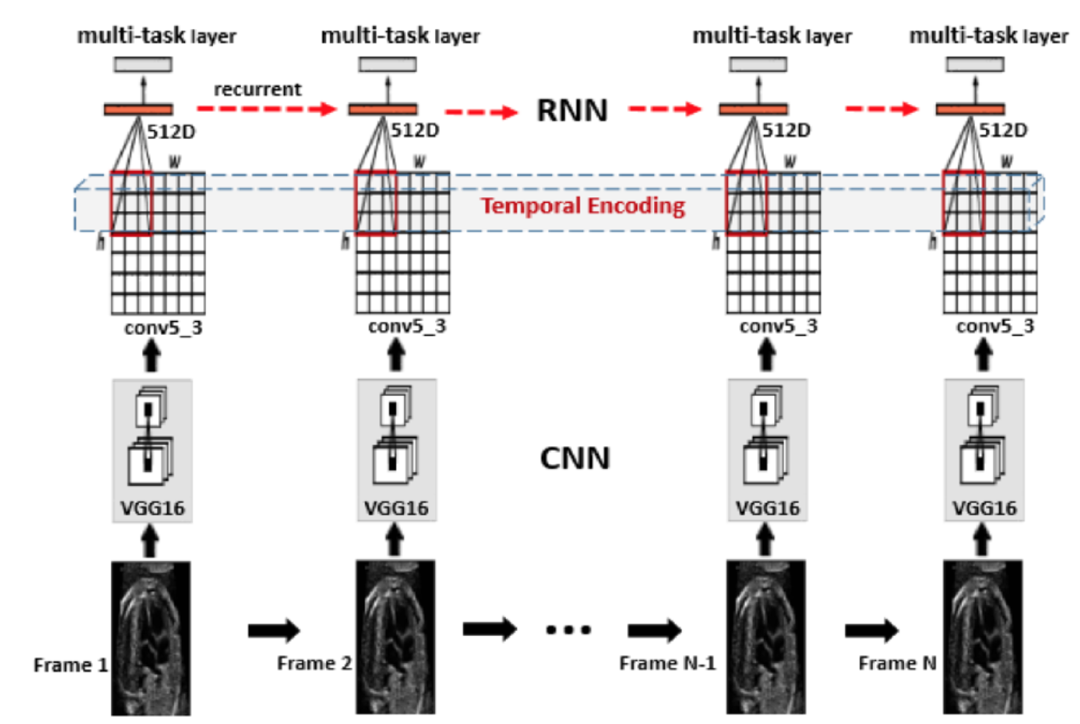

Figure 2: Full architecture

Circular Anchor Architecture

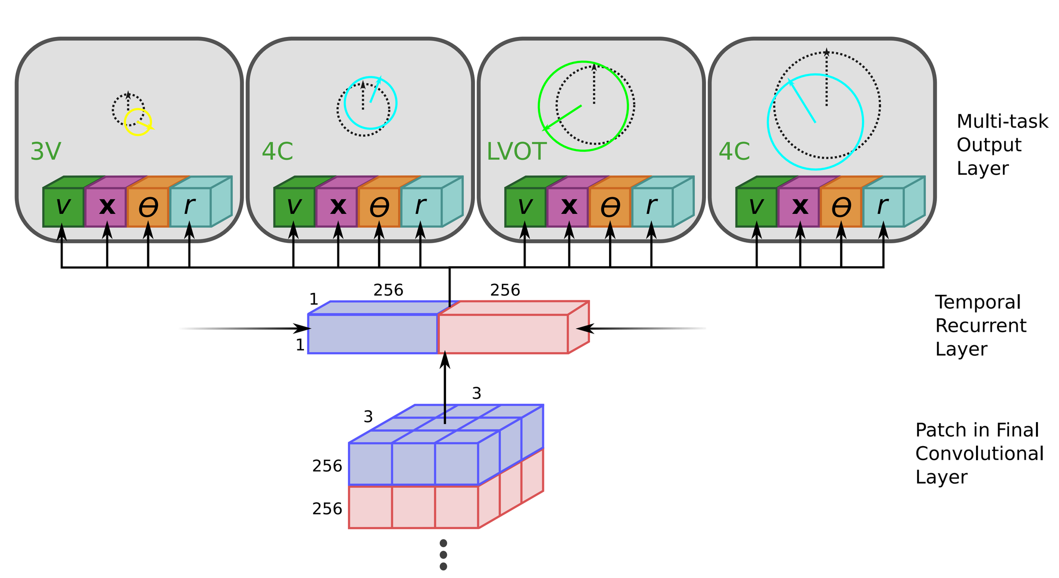

Figure 3: Multi-task and circular anchor architecture

Within each spatial location, predictions are made independently for each of four 'circular anchors' of different radii [2]. Each uses the same input patch, but gradients are only applied for positive anchors. The loss functions are:

- L_{cls} **Classification** (v): Softmax
- L_{loc} **Localisation** ($\mathbf{x}, \theta, r$): Smooth-l_1 loss
- **Total:**

$$L = L_{cls} + \lambda_1 L_{loc}$$

Temporal Recurrent Layer

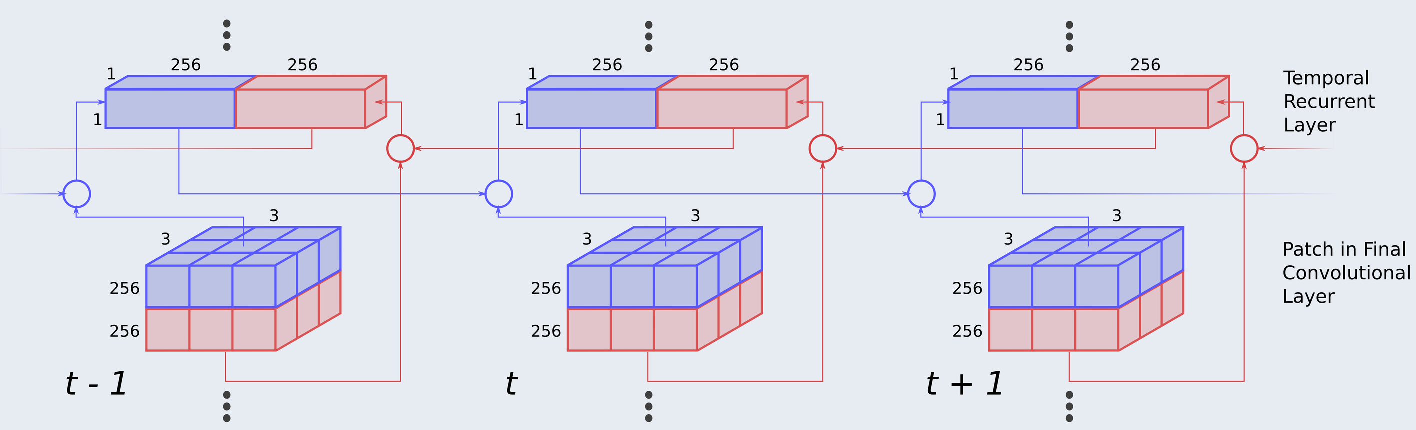

Figure 4: The recurrent layer connects a 3×3 patch in one frame to the corresponding patch in neighbouring frames. There are two separate 256D long short term memory (LSTM) RNNs, one that propagates forward through the video frames (blue) and another that propagates backwards (red) [3].

Intersection-over-Union (IoU) Architecture

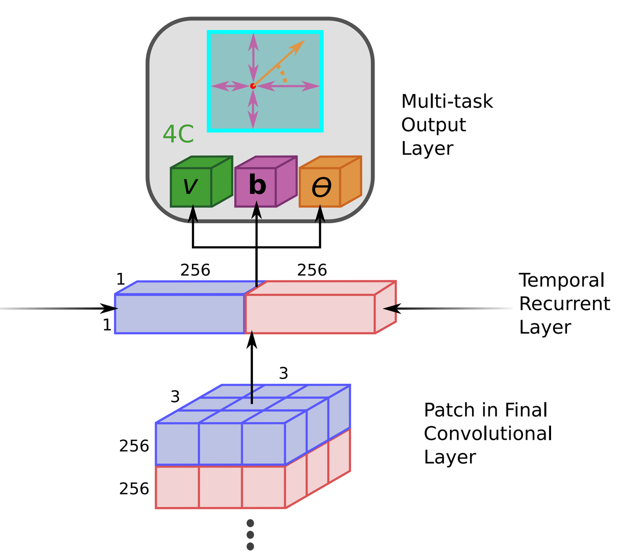

Figure 5: IoU region prediction architecture

The top, bottom, left, and right (**b**) of the bounding box are regressed directly with an IoU loss. Orientation is regressed separately. The loss functions are:

- L_{cls} **Classification** (v): Softmax
- L_{loc} **Localisation** (**b**): IoU (intersection over union) [4]
- L_{ori} **Orientation** (θ): Cosine loss

$$L_{\text{ori}} = 1 - \cos\left(\hat{\theta} - \theta\right)$$

- **Total:** $L = L_{cls} + \lambda_1 L_{loc} + \lambda_2 L_{ori}$

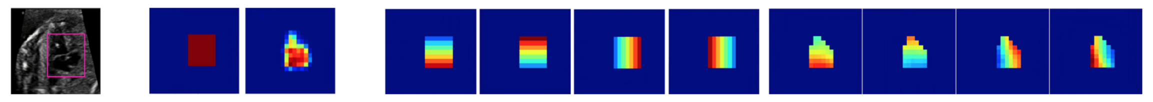

Figure 6: Example of IoU loss. Left to right: input image, ground truth classification map, predicted classification map, ground truth localisation maps ($\times 4$), predicted localisation maps ($\times 4$)

Experiments

- Database of 91 videos from 12 subjects
- Leave-one-subject-out cross-validation
- Multiple views and range of gestational ages (20–35 weeks), orientations, magnifications
- CNN and RNN trained separately due to GPU memory contraints:
  - CNN trained on per-image basis (start with pre-trained VGG-16)
  - RNN trained with random-length sequences

Results

Method	Class Error *or* Outside $0.25\hat{r}$ (%)*	Class Error *or* IoU < 0.25 (%)	Orient. Error†
Circular Anchor	28.8	30.3	0.074
IoU Loss	26.8	28.7	0.084
RNN + Circular Anchor	21.6	27.7	0.072

* Estimated inter-rater variation: 26%, intra-rater variation: 15%

† Orientation Error $= \frac{1}{2}\left(1 - \cos\left(\theta - \hat{\theta}\right)\right)$

- The IoU layer reduces localisation error over the circular anchor mechanism
- Inclusion of the RNN significantly improves all results by considering temporal context

Conclusions

We demonstrated a multi-task deep architecture for estimating multiple quantities of interest from fetal cardiac screening videos. Experiments demonstrate that the region-level temporal information from the RNN improves accuracy on all tasks.

Acknowledgements

The authors would like to thank Christos Ioannou clinical advice, and the EPSRC Seebibyte programme for funding.

References

[1] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014).

[2] S. Ren, K. He, R. B. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In: *NIPS*. Ed. by C. Cortes et al. 2015, pp. 91–99.

[3] A. Graves and J. Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural Networks* 18 (2005), pp. 602–610.

[4] J. Yu et al. "UnitBox: An Advanced Object Detection Network." In: *CoRR* abs/1608.01471 (2016).