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
3D 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, \( x_t \in \mathbb{R}^2 \)
- View Label, \( v_t \in \{\text{4C, LVOT, 3V}\} \) (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 diagnostic of CHD.

View Plane Definitions

View Plane Definitions

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

Architecture Overview

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.

- \( L_{cls} \): Classification (\( v_t \)): Softmax
- \( L_{loc} \): Localisation (\( x_t, \theta_t \)): Smooth-L1 loss
- **Total:**
  \[
  L = L_{cls} + \lambda L_{loc}
  \]

Intersection-over-Union (IoU) 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 (\( v_t \)): Softmax
- \( L_{loc} \): Localisation (\( x_t, \theta_t \)): IoU (intersection over union) [4]
- \( L_{ori} \): Orientation (\( \theta_t \)): Cosine loss

\[
L_{ori} = 1 - \cos(\theta_t - \theta)
\]

**Total:**
\[
L = L_{cls} + \lambda L_{loc} + \lambda L_{ori}
\]

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 constraints
- CNN trained on pre-image basis (start with pre-trained VGG-16)
- RNN trained with random-length sequences

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Class Error or Outside 0.25% (%)</th>
<th>Class Error or IoU &lt; 0.25 (%)</th>
<th>Orient. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circular Anchor</td>
<td>28.8</td>
<td>30.3</td>
<td>0.674</td>
</tr>
<tr>
<td>RNN + Circular Anchor</td>
<td>21.6</td>
<td>27.7</td>
<td>0.672</td>
</tr>
</tbody>
</table>

* Estimated inter-rater variation: 26%, intra-rater variation: 13%.

- Orientation Error = \( \frac{1}{2} \) \[1 - \cos(\theta - \theta_t)\]
- 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.

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

References