CardioNet: Identification of fetal cardiac standard planes from 2D Ultrasound data

Krishna Chaitanya Kaluva
Department of Engineering Design
Indian Institute of Technology Madras
Chennai, India
kaluvakrishnachaitanya@gmail.com

Dr. Shanthi C
Senior consultant, Fetal and Pediatric Cardiology
Mediscan Systems
Chennai, India
drshanthi@mediscan.org.in

Arun Kumar Thittai
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai, India
akthittai@iitm.ac.in

Ganapathy Krishnamurthi
Department of Engineering Design
Indian Institute of Technology Madras
Chennai, India
gankrish@iitm.ac.in

Abstract

Congenital Heart Defects (CHDs) affect up to 1% of newborns globally and about 4-5 % in developing world. Identifying the fetal cardiac standard planes from 2D ultrasound videos is the preliminary step to diagnosis of CHDs. The task of manually identifying the standard cardiac planes such as four Chamber, three vessel, RVOT, LVOT from video frames requires extensive training and experience. In India, about 69 % of the entire population reside in rural areas & the acute shortage of trained and experienced sonographers in these parts of the country make detection of CHDs extremely difficult. In this paper, we propose a novel deep learning based approach for identification of fetal cardiac standard planes from 2D ultrasound videos. Such a tool would help not only the novice sonographers but also the experts in identifying all the cardiac standard planes from a free hand fetal ultrasound video. The dataset was acquired from a collaborating institute post appropriate ethical clearance. On the test data (n= 1636 slices), for the task of identifying the four standard planes, the proposed network achieved an accuracy of 92 % and an average F1-score of 0.919.

1 Introduction

Congenital heart defects are the most common type of birth defects and the severity of the defect can be life threatening. The defects include improper growth of the heart valves and walls separating the various cardiac structures or compartments. In India, the prevalence of CHDs is found to be 19 per 1000 infants. Ventricular septal defect (VSD) is the most common defect accounting for about 33% [1] of the prevalence which is followed by atrial septal defect (ASD). In a clinical setup, 2-D sonography is the primary modality used to identify CHDs due to its low cost and real time cardiac imaging capability. The scans are acquired during the 18-22 weeks gestation period. Detection of the four standard planes namely Four Chamber, Three Vessel, Left Ventricular Outflow Tract (LVOT), Right Ventricular Outflow Tract (RVOT) from 2-D ultrasound videos is the preliminary step to diagnose CHDs. Figure 1 illustrates the 4 standard planes while Fig. 2 shows some common CHDs. However, detecting the standard planes from a plethora of 2-D views requires a highly

1 Images courtesy: https://obgynkey.com/atrial-ventricular-and-atrioventricular-septal-defects/

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.
experienced sonographer. For an untrained/novice sonographers, identifying relevant structures to
detect the standard plane is a challenging task. In India, a large proportion of the population dwell
in the rural parts of the country. The acute shortage of trained sonographers and high prevalence of
CHDs in the Indian population makes development of a computer assisted device (CAD) tool crucial.

(a) 4 Chamber view  (b) 3 Vessel view  (c) LVOT  (d) RVOT

Figure 1: The figures show the 4 major fetal cardiac standard planes. (a) 4 Chamber view: All the 4
chambers (2 auricles, 2 ventricles) and the septa separating them are visible. (b) 3 Vessel view: In
this view we can see the pulmonary artery, ascending aorta and superior venacava. (c) LVOT : Left
ventricular outflow tract. (d) RVOT : Right ventricular outflow tract

(a) VSD  (b) ASD

Figure 2: The figures show the two most common congenital heart defects in infants. (a) Ventricular
Septal Defect (VSD) is a hole in the septum connecting the two ventricles (b) Atrial Septal Defect
(ASD) is a hole in the septum connecting the two auricles.

1.1 Related work

For various computer vision & pattern recognition tasks [10] [5] [6], convolutional neural networks
(CNNs) have either produced comparable results or outperformed traditional machine learning
techniques like support vector machines, random forest. Additionally, deep learning based methods
learn important features directly from the data and seldom requires feature engineering.

In the recent past, various machine learning based techniques have been introduced to detect
standard planes & presence of CHDs from 2-D fetal ultrasound video frames. In [11], 2-D images
were classified as presence of VSDs by first manually cropping a region of interest and then using
a Markov random field (MRF) based technique to segment the structures in the fetal heart. The
frames were classified as VSD if the H- shaped sonography marker was detected from the generated
segmentations. Recently, a CNN based technique, Sononet-32 [1], was utilized to detect 13 different
standard planes for various anatomies from the fetal free hand ultrasound videos. This real time
approach attains an overall F1 score of 0.79 & a retrospective retrieval accuracy of 90.9 %. On the
2 Materials & Methods

2.1 Data

Retrospective ultrasound data (n=16 patients) was acquired from the collaborating hospital, post appropriate ethical clearance. The data was obtained from subjects within the gestational age of 18-20 weeks in DICOM format. Anonymity of the subject was preserved by removing all related headers. The data contained a total of 25 volumes obtained from 16 patients with 16,328 slices in total. Depending on the machine parameters, the dimensions of data obtained were either 720 × 960 or 852 × 1136. The annotations of the acquired data were reviewed by an expert with an experience of about 10 years. In this paper all the views other than the four standard cardiac planes were considered to be background.

2.2 Data pre-processing

The images having dimensions of either 852 × 1136 or 720 × 960, Fig. 3(a) were manually cropped to have resolution 592 × 752, Fig. 3(b). A total variation filter, an edge preserving filter, was employed to minimize the speckle noise and to highlight the edges of the structures, Fig. 3(c). The data was further normalized to have intensity values between zero and one using Eq. (1), wherein \( \text{min} & \text{max} \) represents the minimum and maximum values of the pixel intensities in an ultrasound dicom image.

\[
\text{Normalized data} = \frac{\text{data} - \text{min(data)}}{\text{max(data)} - \text{min(data)}}
\]  

2.3 Model architecture

The proposed network architecture Fig. 4 was inspired from Densely connected convolutional networks [7], however we modified the architecture to optimize it for our task. The architecture comprises of 4 dense blocks. Each dense block Fig. 5 was composed of 3 convolution layers & the growth rate of the network was set to \( k = 16 \). Except for the input convolution layer, all other convolution layers in the network were preceded by a batch normalization [8] & a non-linearity activation function [ELU] [4]. In order to minimize over-fitting on the training data, the convolutional layers were succeeded by spatial dropouts [13], with a dropout rate of 0.2. In literature, a sequence of batch normalization, non-linearity, convolution and dropout is referred to as “Layer”. Fig. 6. Each dense block in the network was followed by a convolution and a spatial max-pooling so as to

Figure 3: In this figure, (a) shows the slice from a dicom volume, (b) shows the cropped image removing unnecessary details and (c) shows the output of the total variational filter applied on (b). All the grainy noise is smoothened while preserving the edges of the structure.
reduce the dimension of feature maps. The network accepts an input of size $592 \times 792$ & returns the posterior probability associated to each of the 4 standard cardiac planes and the background.

**Figure 4:** The figure shows the network architecture with 4 dense blocks. The layers between two adjacent blocks are referred to as transition layers and change the sizes of feature maps via convolution and max pooling. The input to the network is a slice/image of a dicom/video file and the output is the posterior probability for each class.

![Network Architecture](image)

**Figure 5:** The structure of a dense block comprising of 3 layers. Each layer’s outputs are channelwise concatenated with the inputs to every layer preceding it. The structure of each layer is shown in Fig. 6.

![Dense Block](image)

**Figure 6:** The figure shows the structure of a layer. Batch normalization followed by Elu activation were applied to the input of the layer before the convolutional operation. A dropout rate of 0.2 was applied to the convolution outputs to reduce over-fitting on the training data.

![Layer Structure](image)

### 2.4 Training

In the given dataset, we observe a large imbalance between background and other classes, Table 1. The total data (images = 16328) was split into 70% for training, 20% for validation, and 10% for testing. Stratified sampling was employed to ensure that 70% of each class goes to the training set, 20% goes to validation and 10% goes to test.

The network was trained using Adam optimizer with a learning rate of $1e^{-4}$. The magnitude of the class imbalance was reduced by utilizing weighted cross entropy loss function. The weights associated to each class was based on the class-frequency. The network was trained for 200 epochs and an epoch was defined to have 1000 and 500 training and validation iterations respectively. The network was trained with L2 regularization with a weight decay of $1e^{-6}$. The weights were initialized from a truncated normal distribution with mean of 0 and standard deviation of 1.
Table 1: Statistics of the data

<table>
<thead>
<tr>
<th>Class</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>6684</td>
<td>1910</td>
<td>956</td>
</tr>
<tr>
<td>4 Chamber</td>
<td>3005</td>
<td>859</td>
<td>430</td>
</tr>
<tr>
<td>3 vessel</td>
<td>541</td>
<td>155</td>
<td>78</td>
</tr>
<tr>
<td>LVOT</td>
<td>683</td>
<td>195</td>
<td>98</td>
</tr>
<tr>
<td>RVOT</td>
<td>513</td>
<td>147</td>
<td>74</td>
</tr>
<tr>
<td>Total</td>
<td>11426</td>
<td>3266</td>
<td>1636</td>
</tr>
</tbody>
</table>

2.5 Testing

During inference, each slice in the volume was input to the trained network & was classified as one of the five classes.

2.6 Retrieval of top 5 images for each class.

In a typical 2-D ultrasound video, there exists a few slices that satisfies the condition or quality to qualify as a standard plane. In a clinical environment, a sonographer diagnoses a patient or subject as CHD after analyzing multiple such slices from each of the standard planes. Henceforth, we make use the network to retrieve the top 5 slices in the volume for each fetal cardiac standard plane.

Each slice in the test volume was associated to the class & the posterior probability assigned by the network. For each cardiac standard plane, the slices were sorted based on posterior probability to retrieve the top \( K \) slices. The retrieved slices were then presented to the sonographer for further analysis.

3 Results and discussion

For classification related tasks, accuracy, sensitivity (Recall) & precision are the often used metrics to benchmark the performance of the algorithm. The definition of each of the aforementioned terminologies are given in Eq. 2, 3, 4, 5 respectively.

We consider all views or planes other than the 4 standard planes as “background”. The performance of the proposed network was tested on a held-out test data (n= 1636 slices).

On this limited test data, the performance of the classifier to accurately classify each standard plane is tabulated in Table 2.

Upon considering background class in the calculation of overall accuracy the model achieves an overall accuracy of 92 %. In order to understand whether the network has understood well to distinguish planes other than background class, an overall accuracy was calculated on the 4 standard planes and reported separately in Table 2 as 94.26 %.

\[
\text{Accuracy} = \frac{\sum_{i=0}^{\text{num classes}} \text{frequency}_i \times \text{accuracy}_i}{\sum_{i=0}^{\text{num classes}} \text{frequency}_i} \tag{2}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{3}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{4}
\]

\[
F1 - \text{score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}
\]
Table 2: Class-wise classification accuracy on test images

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.902</td>
</tr>
<tr>
<td>4 Chamber</td>
<td>0.977</td>
</tr>
<tr>
<td>3 vessel</td>
<td>0.936</td>
</tr>
<tr>
<td>LVOT</td>
<td>0.847</td>
</tr>
<tr>
<td>RVOT</td>
<td>0.878</td>
</tr>
<tr>
<td>Overall accuracy (including background)</td>
<td>0.92</td>
</tr>
<tr>
<td>Overall accuracy (excluding background)</td>
<td><strong>0.9426</strong></td>
</tr>
</tbody>
</table>

Table 3: Class-wise Precision, Recall and F1-score

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.9572</td>
<td>0.9022</td>
<td>0.9288</td>
</tr>
<tr>
<td>4 Chamber</td>
<td>0.8606</td>
<td>0.9767</td>
<td>0.9150</td>
</tr>
<tr>
<td>3 vessel</td>
<td>0.9359</td>
<td>0.9359</td>
<td>0.9358</td>
</tr>
<tr>
<td>LVOT</td>
<td>0.8556</td>
<td>0.8469</td>
<td>0.8512</td>
</tr>
<tr>
<td>RVOT</td>
<td>0.9286</td>
<td>0.8784</td>
<td>0.9028</td>
</tr>
</tbody>
</table>

Table 5 tabulates the class-wise performance metrics achieved by the network. Except for LVOT, the classifier achieves an F1 score \( \geq 0.90 \). We also compare the performance of the network with the state of the art technique SonoNet – 32 for the task of detecting fetal cardiac planes. As the dataset used by both techniques are different, it is not possible to directly compare the techniques. Table 4 compares the performance of the proposed network on the held out test data against results reported in literature, specifically sonoNet-32.

As mentioned above in Sec 2.6, we retrieved the top 5 images for each class using the posterior probabilities from the test data and plotted them Fig. 8. All the best images for each class were verified with the clinical expert for a qualitative analysis of the retrieval.

Despite being trained on a limited dataset (n=11426), the classifier achieved good performance on the held-out test data. The absence of multiple fully connected layers in DenseNets aids in reducing the number of parameters in the network when compared to networks with multiple fully connected layers e.g AlexNet[9], VGG-16, VGG-19[12]. This reduction in number of parameters in-turn reduces the requirement of large number of training examples.

Figure 7: The figure show the complete flow of the proposed model. The inputs are first fed to our proposed deep learning model which outputs the posterior probabilities of all classes for each slice, which is then passed through a sorting algorithm which outputs the best 5 slices for each fetal cardiac standard plane.
Table 4: Comparison of the proposed network against Sononet-32.

<table>
<thead>
<tr>
<th>Class</th>
<th>SonoNet-32</th>
<th>CardioNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.99</td>
<td>0.9288</td>
</tr>
<tr>
<td>4 Chamber</td>
<td>0.77</td>
<td>0.9150</td>
</tr>
<tr>
<td>3 vessel</td>
<td>0.63</td>
<td>0.9358</td>
</tr>
<tr>
<td>LVOT</td>
<td>0.78</td>
<td>0.8512</td>
</tr>
<tr>
<td>RVOT</td>
<td>0.59</td>
<td>0.9028</td>
</tr>
</tbody>
</table>

Figure 8: Top 5 images retrieved for each class using posterior probabilities. 4 Chamber view: (a) - (e), 3 Vessel view: (f) - (j), LVOT: (k) - (o), RVOT: (p) - (t)

4 Conclusion

We present a novel deep learning based approach to help novice as well as expert sonographers in identifying the fetal cardiac standard planes. The network was trained & tested on a limited amount of dataset acquired from an Indian population. From our preliminary results, we observe that, based on performance metrics such as F1 score and precision, our network was able to produce results comparable with the state-of-the-art techniques. For the task of distinguishing each of 4 planes in the volume appropriately, the network achieves an overall accuracy of 94.26%. Based on the posterior probability associated to each slice by the network, the top 5 slices in the volume corresponding to each standard plane was retrieved & displayed to the sonographer for further analysis. Figure 7 shows the detailed flow of the proposed model.

As part of the future work,
• We plan to test the performance of the network on a larger dataset.
• Based on the retrieved slices, classify the volume as normal or abnormal (various types of CHDs).
• Extend the work to extract standard planes associated with other anatomical structures.
References


