# Detection of Cardiac Events in Echocardiography using 3D Convolutional Recurrent Neural Networks

Adrian Meidell Fiorito
*Dept. of Engineering Cybernetics*
*NTNU*
Trondheim, Norway
adrianf@stud.ntnu.no

Andreas Østvik
*Centre for Innovative Ultrasound Solutions*
*NTNU*
Trondheim, Norway
andreas.ostvik@ntnu.no

Erik Smistad
*Centre for Innovative Ultrasound Solutions*
*NTNU*
Trondheim, Norway
erik.smistad@ntnu.no

Sarah Leclerc
*CREATIS*
*Universite de Lyon*
Lyon, France
sarah.leclerc@insa-lyon.fr

Olivier Bernard
*CREATIS*
*Universite de Lyon*
Lyon, France
olivier.bernard@insa-lyon.fr

Lasse Lovstakken
*Centre for Innovative Ultrasound Solutions*
*NTNU*
Trondheim, Norway
lasse.lovstakken@ntnu.no

***Abstract*—A proper definition of cardiac events such as end-diastole (ED) and end-systole (ES) is important for quantitative measurements in echocardiography. While ED can be found using electrocardiography (ECG), ES is difficult to extract from ECG alone. Further, on hand-held devices ECG is not available or cumbersome. Several methods for automatic detection of cardiac events have been proposed in the recent years, such as using a 2D convolutional neural network (CNN) followed by 1D recurrent layers. This structure may be suboptimal, as tissue movement has a spatio-temporal nature which is ignored in the CNN.**

**We propose using a 3D CNN to extract spatio-temporal features directly from the input video, which are fed to long short term memory (LSTM) layers. The joint network is trained to classify whether frames belong to either diastole or systole. ES and ED are then automatically detected as the switch between the two states. The 3D CNN + LSTM model performs favourably at detecting cardiac events on a dataset consisting of standard B-mode images of apical four- and two-chamber views from 500 patients. The mean absolute error between events in the apical four-chamber view is $1.63$ and $1.71$ frames from ED/ES reference respectively. Model inference is fast, using $(30 \pm 2)$ ms per 30 frame input sequence on a modern graphics processing unit.**

## I. Introduction

Detection of end-systole (ES) and end-diastole (ED) in echocardiography is an important step when assessing cardiac function. ED and ES are defined as the time points when the mitral valve and aortic valve closes respectively [1]. Several clinical metrics, such as ejection fraction and global longitudinal strain [2] are determined using the ES and ED images. The current approach for detecting ED usually involves finding the QRS-complex in additional measurements from electrocardiograms (ECG), or by visual inspection of the videos. Finding ES is more difficult in ECG alone, making visual inspection of ultrasound (US) images necessary. In clinical practice, this constitutes a significant amount of work that potentially could be automated. An additional benefit is that accurate detection of ES and ED solely using echocardiographic frames removes the need for applying ECG-patches, further reducing time and resources. This is especially useful for smaller devices such as the pocket-sized US scanners.

A multitude of machine learning methods have been proposed for learning video representations. Recently, deep learning have been able to perform on par or better than traditional approaches. These methods differ in the way spatial and temporal features are combined. In the two-stream network [3], one CNN is trained to extract features from still images, and another CNN is trained to capture motion patterns using a stack of optical flow frames. Several methods have been proposed to increase the temporal capacity of these models, such as extending the CNN to 3D [4]. Another popular approach is the Long-Term Recurrent Convolutional Network [5], which uses a CNN to extract features for individual frames. These features are input into a Long Short-term Memory (LSTM) [6] recurrent network for temporal fusion. Similarly, [7] use a shallow 3D CNN to extract features from short clips, which are passed to an LSTM network. Other methods use deeper 3D CNNs to learn spatio-temporal features [8], [9].

Several methods have been proposed for detecting cardiac events automatically in echocardiography. Cardiac cycle start and length are estimated without the use of ECG in [10]. To detect cycle start, the motion of a point near the mitral annulus is found using speckle tracking. This is compared to a database of left ventricle (LV) displacement curves to estimate the cycle start corresponding to the QRS complex in ECG. Other methods explore manifold learning and dimensionality reduction [11], [12]. Frames in an echocardiogram are mapped to a learned manifold, and the fact that ED and ES occur in periods with small volumetric changes is used to detect these events as dense regions on the manifold. CNNs have been used to extract ED and ES with high precision in cine magnetic resonance imaging [13]. Here, a pretrained CNN is used as a feature extractor, and features are passed on to an LSTM layer. The model is trained to regress a typical volume curve of the LV over a single heartbeat. ED and ES is then identified

as the largest and smallest regressed volume in the sequence, respectively. A similar approach applied to echocardiography replaced the pretrained CNN with a residual network [14].

In this work, we replaced the standard CNN with a 3D CNN for spatiotemporal feature learning. Further, we propose training the model on a target which is more suited for detecting ED and ES. The model is trained on variable length sequences, whereas previous deep learning approaches use fixed length input videos.

## II. Methodology

### A. Problem formulation

To train models for detecting ED and ES frames in a supervised manner, the target output must be generated. An intuitive approach involves posing this as classification with three classes: ED, ES, or neither. However, this introduces a class imbalance problem, as ED and ES frames are underrepresented. An easy way to achieve low loss is then to output neither for all frames.

In [13], [14] the problem is formulated as a regression task. Here, the target is set to approximate a typical LV volume curve, by using a cubic function and normalizing the target 0 to 1. The representation is thus not the actual volume curve for a given sample, and therefore the model must attempt to learn a mapping which is not exactly present in the data. For some cases of pathology, such as in the event of post-systolic contraction, the volume might not be smallest at the time of ES. In addition, detecting ED/ES as extrema in the estimated volume curve might be difficult due to flat regions during isovolumetric periods, resulting in several candidates.

In this work, the problem is formulated as a binary classification task. The target is set to $0$ for frames belonging to systole, and $1$ for frames in diastole. This alleviates the issue of class imbalance, as there is a comparable number of diastole and systole frames. ES and ED is detected as the frames where the output at the next timestep crosses $0.5$ from below or above, respectively.

### B. Network architecture

A 3D CNN architecture is presented which is capable of handling arbitrarily long sequences (until GPU memory is full). Due to high GPU memory utilization of 3D convolutions, the model contains few filters and use pooling frequently compared to state-of-the-art image recognition models. The CNN consists of five 3D convolutional layers, each followed by batch normalization, ReLU activation and max pooling layers. As one prediction should be made for each input frame, pooling is only performed along the spatial axes, and not along the temporal axis. For the same reason, each convolutional layer pads the input with zeroes to preserve the length of the data. As in [9], kernels have spatial and temporal size of 3, except from the first layer which uses a spatial size of 7. The number of feature maps double every convolutional layer, starting at 16 and ending at 256. At the output of the 3D CNN, dropout with a probability of 0.3 is performed to prevent overfitting. The output of shape $[t, 4, 2, 256]$ is then reshaped into $t$ vectors of shape [2048]. LSTM layers are added to filter the CNN predictions and increase the capability to remember longer movements. Both LSTM layers have a cell state of size 32, resulting in 32 output features per timestep. An $L^2$ regularization of $1 \times 10^{-4}$ is used for recurrent and convolutional kernels. A 1D convolutional layer with a sigmoid activation is placed at the end of the model, operating along the temporal axis. The layer has a single kernel of temporal size 3, with the aim of smoothing the output of the model and reduce the likelihood of the output erroneously crossing 0.5 as a result of noisy data. The model is implemented in Keras with the Tensorflow backend. Fig. 1 shows the overall layout.

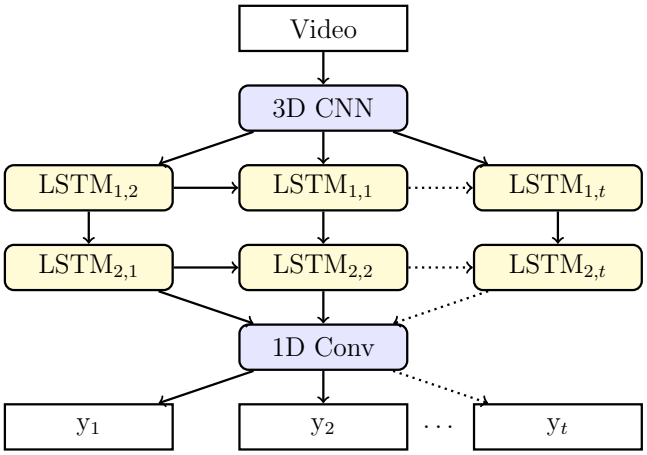


Fig. 1. Schematic of the network architecture with a 3D CNN followed by LSTM layers and a 1D convolutional layer at the end.

### C. The dataset

The dataset consists of apical four-chamber (A4C) and two-chamber (A2C) echocardiograms from 500 patients, acquired at the University Hospital of St-Etienne (France) using a GE Vivid E95 ultrasound system (GE Vingmed Ultrasound, Horten, Norway) [15]. For most patients, a corresponding electrocardiogram (ECG) is aligned with each sequence, giving one ECG measurement per video frame. The data is representative for a typical outpatient clinic. The videos have varying sector geometries, sampling rates and durations. The sample time per frame is between 11.99 ms and 21.05 ms. Each video contains a varying number of cardiac cycles. For each video, one frame corresponding to ED and one frame corresponding to ES is labeled by an expert. The labeled ES and ED belongs to the same heart cycle, with ED labeled first for 498 of the A2C videos, and for 481 of the A4C videos. The dataset is split randomly into three folds, with 300 patients used for training, 100 for validation during training, and 100 for testing. Both the A4C and A2C videos for a single patient are placed in the same fold to avoid data leakage.

As only one ES and ED is labeled for each video in the dataset, no labeled input data contains a full heart cycle. In addition, a majority of the frames between the labeled ED/ES

belongs to systole, as the labeled ED most commonly occurs before ES. To have training data for any part of the heart cycle, an additional ED is labeled by considering the accompanying ECG signal. The QRS-complex is used to label the ED that yields a fully labeled heart cycle. From 500 patients, 333 and 334 of the ECG signals corresponding to A4C and A2C videos respectively are considered of high enough quality to accurately identify a second ED.

A number of frames before and after the labeled ED/ES are included to further expand the dataset size, and to make sure ED and ES does not occur at the first and last frames. The resulting dataset contains 26818 frames of A4C and 26170 frames of A2C echocardiograms, belonging to both diastole and systole. The frames are resized to size $128 \times 80$ using bicubic interpolation, and normalized by subtracting the mean and dividing by the standard deviation over all pixels in the training data.

### D. Learning details

Training is done for 100 epochs with cross-entropy loss applied over each time step. The Adam optimizer with a learning rate of $1 \times 10^{-4}$ was used. At the end of every epoch, the training data is shuffled. Both A4C and A2C views are used as training data. Model weights are saved at the epoch with the lowest mean absolute error (MAE) on the validation set. Training on only A4C views was also tested, but resulted in worse performance. The model is trained using mini-batches of four videos, and shorter videos and targets are padded at the end with zeroes. The loss is set to zero for padded frames before backpropagation.

Data augmentations were important for preventing overfitting. Sequences are downsampled temporally by a factor of 2 by discarding every other input frame with a probability of $0.2$. Sequences are temporally cropped by randomly discarding between $0 - 80\%$ of the original duration, starting and ending at a random frame. After this, videos are rotated randomly between -10 to 10 degrees. Next, videos are randomly cropped spatially, removing between 0 and 20 pixels along each border. After cropping, the videos are resized to the input size expected by the model. Training and model evaluations were performed on a NVIDIA Titan V GPU with 12 GB RAM.

### E. Evaluation

The error is defined as the difference between the time of a labeled event $E$ and a detected event $\hat{E}$, either ED or ES. Using the notation of [13], the MAE in frames is denoted the average frame difference (aFD),

$$\text{aFD} = \frac{1}{N} \sum_{1}^{N} |E - \hat{E}|, \tag{1}$$

where $N$ is the number of events in the dataset. The mean ($\mu_e$) and standard deviation ($\sigma_e$) of the error is also presented in milliseconds (ms).

In order to evaluate if the model is invariant to the cardiac cycle starting point, a variable number of additional frames are included at the beginning and end of the sequence. For each video in the test set, results are measured with $0\%$, $33\%$ and $66\%$ of the duration between the labeled ED and ES included at the beginning and end of the input data. The model output for the included frames are then discarded as there is no ground truth for these frames.

## III. Results

Table I shows the resulting performance on the 100 patients in the test set. Three of the labeled EDs and ESs are not detected by the model for the A2C view, due to the event occurring near the first or last frames of the input. More than one detection of the same event occur three times for ED, and six times for ES. In all these cases, inspection reveals that the data is from a non-standard view or noisy. These cases are thus excluded in the result metrics. Fig. 2 shows a patient from the dataset along with labeled and detected events, while Fig. 3 shows the model output for the patient.

TABLE I
ERRORS OF DETECTED ED AND ES RELATIVE TO LABELED ED AND ES

| View | Event | aFD | $\mu_e(ms)$ | $\sigma_e(ms)$ |
|---|---|---|---|---|
| A2C | ED | 1.40 | -5.68 | 35.8 |
| | ES | 1.25 | -1.94 | 29.9 |
| A4C | ED | 1.63 | 0.50 | 29.8 |
| | ES | 1.71 | 0.60 | 37.8 |

Table II shows the model compared to results reported in [14] for other deep learning approaches.

TABLE II
COMPARISON TO METRICS REPORTED IN [14] ON THE A4C VIEW

| Model | aFD (ED) | aFD (ES) |
|---|---|---|
| CNN + LSTM [13] | 6.3 | 7.3 |
| ResNet + LSTM [14] | 3.7 | 4.1 |
| 3D CNN + LSTM | **1.6** | **1.7** |

The time used to predict a single video consisting of 30 frames is measured 100 times and averaged. This resulted in $(30 \pm 2)$ ms used on average for predicting 30 frames.

## IV. Discussion

The 3D CNN is able to detect both ED and ES accurately both for A4C and A2C views, as seen in Table I. This suggests that the network has learned general features for both cardiac phases, such as movement of the atrioventricular valves and the contraction / relaxation of the myocardium. The model is suited for learning these features, as the 3D convolutional layers are able to learn motions between adjacent pixels. A 3D CNN alone might result in a noisy output, due to the noisy input data. This is where the LSTM layers can do a good job of filtering the CNN output. There are few visible difference between the labeled and detected ES frame in Fig. 3, and the most noticeable difference between the labeled and detected ED frames is the slightly more closed mitral valve for the labeled ED. As seen from Fig. 3, the model output closely resembles a square wave corresponding to systole and diastole

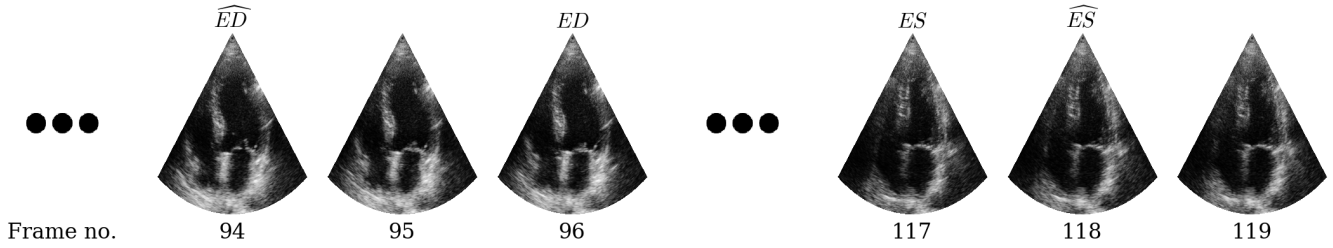


Fig. 2. Example input sequence (apical four-chamber) along with the labeled frames (ED, ES) and frames detected by the model ($\widehat{ED}$, $\widehat{ES}$).

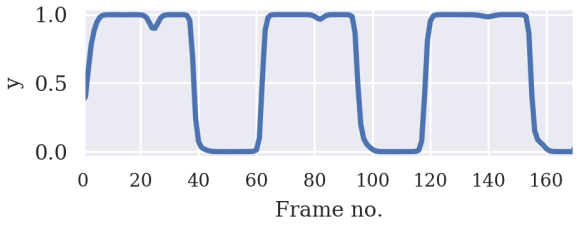


Fig. 3. Output of the model on the sequence shown in Fig. 2. The model output, $y$, is close to 1 for frames corresponding to the diastole phase and 0 for frames in systole. ED and ES is detected as frames where the $y$ crosses 0.5.

frames, with only a few noticeable dips, showing how well the model separates between systole and diastole. As seen in Table II, the aFD is less than half of [14]. Comparing the performance must however be performed with caution, due to the models being evaluated on different datasets.

An issue is that the labels are not guaranteed to be correct. Determining the exact moment of ED and ES can be difficult for a human annotator due to small differences between consecutive frames. These errors increase as the sampling rate increases. Therefore, it would be interesting to compare the variability of human annotators.

Frequent pooling and few convolutional kernels ensures that the model runs efficiently. It also has a regularization effect, as a small network is less likely to overfit to the training data. The approach has shown to work using a variable number of input frames, instead of limiting the input to a fixed number of frames. This means that the model can operate on an arbitrary long input sequence, and is not restricted to using a single heart cycle as input. Thus, the method may be used to automatically extract heart cycles when considering the distinct differences between output for diastole and systole frames.

## V. Conclusion

In this paper, a novel method for detecting cardiac events in echocardiography using deep learning was proposed. A 3D CNN was employed followed by recurrent layers to facilitate the learning of spatio-temporal features. State-of-the-art results are achieved on a large dataset, which indicate that the chosen components enhances the solution of the task.

## References

[1] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Rietzschel, L. Rudski, K. T. Spencer, W. Tsang, and J.-U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal - Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[2] A. Østvik, E. Smistad, T. Espeland, E. A. R. Berg, and L. Lovstakken, "Automatic myocardial strain imaging in echocardiography using deep learning," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing, 2018, pp. 309–316.

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 4724–4733.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 29–39.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[10] S. A. Aase, S. R. Snare, H. Dalen, A. Støylen, F. Orderud, and H. Torp, "Echocardiography without electrocardiogram," *European Journal of Echocardiography*, vol. 12, no. 1, pp. 3–10, 2010.

[11] P. Gifani, H. Behnam, A. Shalbaf, and Z. A. Sani, "Automatic detection of end-diastole and end-systole from echocardiography images using manifold learning," *Physiological Measurement*, vol. 31, no. 9, p. 1091, 2010.

[12] A. Shalbaf, Z. Alizadehsani, and H. Behnam, "Echocardiography without electrocardiogram using nonlinear dimensionality reduction methods," *Journal of Medical Ultrasonics*, vol. 42, no. 2, Apr 2015.

[13] B. Kong, Y. Zhan, M. Shin, T. Denny, and S. Zhang, "Recognizing end-diastole and end-systole frames via deep temporal regression network," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Cham: Springer International Publishing, 2016, pp. 264–272.

[14] F. T. Dezaki, N. Dhungel, A. H. Abdi, C. Luong, T. Tsang, J. Jue, K. Gin, D. Hawley, R. Rohling, and P. Abolmaesumi, "Deep residual recurrent neural networks forcharacterisation of cardiac cycle phase from echocardiograms," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing, 2017, pp. 100–108.

[15] S. Leclerc, E. Smistad, T. Grenier, A. Østvik, F. Espinosa, L. Lovstakken, and O. Bernard, "Deep learning applied to multi-structures segmentation

in 2D echocardiography: a preliminary investigation of the required database size," in *IEEE International Ultrasonics Symposium, IUS*, 2018.