Project Proposal- Echocardiography Segmentation and Ejection Fraction Estimation

Cheng Che Tsai
Department of Computer Science, University of North Carolina at Chapel Hill
cctsai@cs.unc.edu

Alessandro Folloni
Division of Artificial Intelligence, Alma Mater Studiorum Università Di Bologna
afolloni@unc.edu

Abstract

Our project focuses on providing fast and comprehensive solutions for echocardiography analysis. Specifically, we aim to build a model for left ventricle segmentation and ejection fraction based on some video analysis techniques learned in class. Past studies have thoroughly explored these two topics, but most of them were based on image analysis techniques instead of video. The results were built on the CAMUS dataset, which only contains 500 video clips, compared to the 10,000 clips in the echonet dataset that we are going to use in this project.

1. Problem Definition

1.1. Echocardiography

Echocardiography, often referred to as "echo," is a valuable medical imaging technique for assessing heart structure and function. It aids in early heart problem detection, monitors cardiac conditions, and assesses the effectiveness of medical interventions. Compared to heart CT and MRI, echocardiography is fast and radiation-free.

However, echocardiography has two major limitations: operator dependency and the potential for low-resolution images or videos. Firstly, the quality of results varies with the operator's skill and experience, leading to inconsistent interpretations among professionals. Secondly, echocardiography may produce low-resolution images, particularly when compared to cardiac MRI or CT scans. In some cases, obtaining clear and detailed images can be challenging, limiting the ability to accurately assess certain cardiac structures or conditions.

These two limitations underscore the need for a standardized, non-operator-dependent method to provide heart echo analysis, which is the motivation behind our project. Specifically, we aim to focus on two key indices of heart condition: left ventricle (the major chamber of the heart pump) contour segmentation and ejection fraction estimation (a regression task).

1.2. Ejection Fraction Estimation (EF Estimation)

Ejection fraction estimates how well your heart is pumping blood. It is a percentage that tells us the proportion of blood pumped out of the heart with each beat. A healthy heart typically has an EF of around 55% to 70% and values below 50% will be considered as heart failure. Mathematically, it can be defined as:

$$EF = \frac{EDV - ESV}{EDV} \times 100\%$$

where

- EDV stands for End-Diastolic Volume, which is the amount of blood in the left ventricle at the end of diastole (when the heart is fully relaxed and filled with blood).
- ESV stands for End-Systolic Volume, which is the amount of blood in the left ventricle at the end of systole (when the heart has just contracted and pumped out blood).

In addition to the ratio, we will also try to estimate EDV and ESV, which provide values for deciding on medication for treating heart failure if it occurs.

1.3. Overall Goals

Our project goal is to build a model that can simultaneously provide segmentation masks for the left ventricle and estimate ejection fraction (and possibly EDV and ESV) in one step. The input to the model will be short clips of heart echo. These video clips can contain either a single cycle of

heart beating or multiple cycles. We'll build and test our models on both clip configurations.

2. Related Work

1) 2D CNNs: In 2019, Leclerc et al. [1] established a milestone in echocardiographic segmentation by publishing the CAMUS dataset. This is the first publicly available larger-scale dataset for performing segmentation on heart echo. The authors also demonstrated competitive results with different UNet-based models. A year later, Leclerc et al. [2] introduced their LU-Net, a two-step segmentation network inspired by Mask R-CNN. LU-Net showed improved results compared to the authors' previous UNet architecture, even beating intra-observer accuracy on the epicardium.

2) 2D+time CNNs: Alongside the work of Leclerc et al., Ouyang et al. introduced their EchoNet-Dynamic dataset [3]. Initially developed as a single-task model using the R(2+1)D architecture, they later enhanced its performance by transforming it into a multi-task model for LV segmentation and EF regression [4], while constraining it to a single beat cycle. The EF regression achieved an MAE of 4.10.

More recently, Wei et al. introduced CLAS [5], a unique 3D segmentation network aimed at achieving temporal consistency with only ED and ES annotations. To achieve this, CLAS predicts deformation fields and utilizes them for annotation propagation during training, improving the consistency between ED and ES predictions. However, their evaluation of temporal consistency across sequences was primarily qualitative, focusing on a limited number of patients. Dai et al. [6] proposed AdaCon, a novel contrastive learning framework for regression problems. These end-to-end regression methods yield more reliable estimates, avoiding error propagation from separate frames, resulting in a 3.3% reduction in MAE.

3) Limitations of past approaches: There are three major limitations in the past studies that we aim to address and improve:

- First, most EF estimation models discard the intermediate frames between end-diastole (ED) and end-systole (ES), even though these frames contain valuable information for characterizing other pathologies and can synergize with LV segmentation.

- Second, the models used to model temporal information have all been CNN-based models. However, these models are not considered state-of-the-art in the era of transformers.

- Third, these models have not fully leveraged the repetitive nature of the data. Video clips between cardiac cycles represent a natural augmentation that can be harnessed to enhance performance. An illustrative example can be found in [7].

Hence, we aim to develop a model that can provide LV segmentation masks across the videos, estimate the dynamic volumes of the beating heart, and provide EF estimation in the end.

3. Technical Approach

3.1. Architectures In Interest

We planed to start with the two most coceptually simple models introduced in the class: TimeSformer[8] and R-CNN [9].

Inspired by the success of the Transformer architecture in natural language processing and computer vision, TimesFormer leverages the self-attention mechanism to capture long-range temporal dependencies in videos and time series data. For our project, we like to build a segmentation model that is conditioned on the temporal information provided by the spatiotemporal module using TimesFormer's methods.

On the other hand, R-CNN is a powerful CNN-based segmentation model. Unlike popular R-CNN works in the natural image domain, we would like to explore whether the repetitive nature of the heart echo can help reduce the proposed regions and, therefore, accelerate this model.

We also want to note here that neither author of this project is familiar with video analysis works. These proposed models are likely to undergo changes as the project progresses.

4. Experiments

Provide details for your experimental setup:

4.1. Experiments

- LV segmentation: we will evaluate our segmentation results on three metrics.
  - The first one is the commonly used Dice score evaluated only at the end of the systolic phase and at the end of the diastolic phase, which is the conventional methods used in the previous study.
  - The second one is the temporal IoU [10]. This helps to establish the temporal evaluation of the generated masks.
  - The third metric is a second derivative-based image smoothness score. This score can help us to understand if the temporal segmentation masks is anatomically reasonable.

- EF regression: We will evaluate model performance using the most commonly used metrics: R^2, MAE and RMSE.

Regarding our expectations, we do not expect our model can perform better at the systolic/diastolic phase evaluation since past models have provided a decent 0.95 performance. However, we expect our model provide better smoothness between the frames and better temopral IoU because of incoporating the temporal information and the beat-cycle constraints imposed.

Regarding the

4.2. Dataset

For our project, the choice of dataset for use is not straightforward as the first glance. There are not many validated and meaningful sources in the literature, indeed. We opted for two datasets that were relatively popular in the past and suit perfectly for our goals: CAMUS and EchoNet-Dynamic.

CAMUS dataset (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) was introduced in 2019. It contains 2D echocardiographic sequences with two and four-chamber views (two different angles of viewing the heart) of 500 patients acquired at the University Hospital of Saint Etienne. Three cardiologists annually annotated the left ventricle endocardium, the myocardium, and the left atrium. In other words, we have keypoints of the inner and outer walls of the left ventricle.

Moreover, it has the peculiarity of having one subset annotated by the same physician 7 months apart; this is helpful to measure intra- and inter-operator variability. Regarding the dimensionality, there is not a fixed frame size; each video has a different resolution (all of them are larger than 1024×512). This is due to the fact that it enforced clinical realism. That is to say, there was no pre-processing on the data acquired when creating the dataset.

EchoNet-Dynamic was created in 2021 to provide images for comprehensive study of cardiac motion and chamber volumes using echocardiography on a large scale. These videos were acquired during real clinical practice for diagnostic and medical decision-making purposes. The dataset comprises 10,036 videos, each originating from a different individual. A typical full resting echocardiogram study consists of a series of 50-100 videos and still images, capturing various perspectives of the heart through different angles, locations, and image acquisition techniques. Ultimately, from each study, one apical-4-chamber 2D grayscale video is extracted. The uniqueness of the dataset is that each video is linked to clinical measurements and calculations and it is full of different annotations including paired electrocardiography (ECG) and ejection fraction.

These videos depict canonical apical-4-chamber views or zoomed-in apical-4-chamber echocardiographic views of sufficient quality for sonographers to determine left ventricular volumes as part of their standard clinical workflow. Each video underwent cropping and masking to remove medical text and information outside the scanning sector. Subsequently, all resulting images were downsized to a standard 112x112 size.

References

[1] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019. 2

[2] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, M. Belhamissi, S. Israilov, T. Grenier, *et al.*, "Lu-net: a multistage attention network to improve the robustness of segmentation of left ventricular structures in 2-d echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2519–2530, 2020. 2

[3] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning," in *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019. 2

[4] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020. 2

[5] H. Wei, H. Cao, Y. Cao, Y. Zhou, W. Xue, D. Ni, and S. Li, "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pp. 623–632, Springer, 2020. 2

[6] W. Dai, X. Li, W. H. K. Chiu, M. D. Kuo, and K.-T. Cheng, "Adaptive contrast for image regression in computer-aided disease assessment," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1255–1268, 2021. 2

[7] W. Dai, X. Li, X. Ding, and K.-T. Cheng, "Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos," *IEEE Transactions on Medical Imaging*, 2022. 2

[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, vol. 2, p. 4, 2021. 2

[9] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015. 2

[10] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5188–5197, 2019. 2