A Unified Framework Integrating Recurrent Fully-Convolutional Networks and Optical Flow for Segmentation of the Left Ventricle in Echocardiography Data

Mohammad H. Jafari[1](✉), Hany Girgis[1,2], Zhibin Liao[1], Delaram Behnami[1], Amir Abdi[1], Hooman Vaseli[1], Christina Luong[1,2], Robert Rohling[1], Ken Gin[1,2], Terasa Tsang[1,2], and Purang Abolmaesumi[1]

[1] University of British Columbia, Vancouver, Canada
mohammadj@ece.ubc.ca
[2] Vancouver General Hospital, Vancouver, Canada

Abstract. Accurate segmentation of left ventricle (LV) from echocardiograms is a key step toward diagnosis of cardiovascular diseases. Manual segmentation of the LV done by sonographers or cardiologists can be time-consuming, and its accuracy is subjective to the operator's experience and skill level. Automation of LV segmentation is a challenging task due to a number of factors such as the presence of speckle and a high operator-dependent variability in acquiring echocardiography data. In this paper, we present a method that integrates deep recurrent fully-convolutional networks and optical flow estimation to accurately segment the LV in the apical four-chamber (A4C) view. Our method analyzes the temporal information in echocardiogram cines with the use of convolutional bi-directional long short-term memory units. Furthermore, it uses optical flow motion estimation between consecutive frames to improve the segmentation accuracy. The proposed method is evaluated over an echo cine dataset of 566 patients. Experiments show that the proposed system can reach a noticeably high mean accuracy of 97.9%, and mean Dice score of 92.7% for LV segmentation in A4C view.

Keywords: Fully convolutional network · Recurrent neural network
Convolutional bi-directional LSTM · Deep learning
Video segmentation · Left ventricle segmentation · Echocardiography

1 Introduction

Cardiovascular disease is the foremost cause of mortality worldwide, resulting in an estimated 17.7 million deaths annually [1]. Assessment of left ventricle (LV) function is considered as a key metric to determine the risk of heart disease.

H. Girgis—Joint first authors.

T.Tsang and P. Abolmaesumi—Joint senior authors.

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 29–37, 2018.
https://doi.org/10.1007/978-3-030-00889-5_4

Echocardiography (echo) is an imaging technique that is often used to inspect cardiovascular function. Segmentation of the LV in echo images is used to derive clinically important measurements such as LV ejection fraction (EF) estimation and wall motion abnormality detection [10]. In particular, the current clinical practice of LV EF estimation requires an expert to manually trace the endocardial border of LV on both end-diastole (ED) and end-systole (ES) frames of an echo cine clip. However, manual LV segmentation is a laborious procedure and its accuracy is often dependent on the operator's experience, resulting in a low test-retest reliability [3].

A number of research groups have attempted to automate the segmentation of LV in echo and also other modalities [3,4,9,11,14,18]. Methods to-date can be categorized into active contour models, deformable templates, level sets, and supervised learning approaches [3,10]. Specifically, in recent years, deep learning [7] has been proposed for segmentation and quantification of LV in computed tomography (CT) and cardiac magnetic resonance imaging (CMR) [9,17,18]. For CT images, Zreik *et al.* [18] propose a two stage LV segmentation method, where the first stage detects a bounding box containing LV by using Convolutional Neural Networks (CNN), and the second stage performs LV segmentation by using voxel classification within the bounding box. An extensive literature review of methods for LV segmentation in CMR is presented in [9,17]. Specifically, Ngo *et al.* propose a level-set model, initialized by LV map obtained from a first deep belief network (DBN), and constrained by the location of endocardial and epicardial borders computed by a second DBN. Xue *et al.* [17] propose a deep network model to quantify LV measurements in CMR as a multi-task relationship learning. In [17], features extracted from cardiac cine using CNNs are fed into two branches of recurrent neural networks, one combined with a Bayesian-based multi-task relationship module for LV quantification, and another branch is ended with a softmax layer to detect the cardiac phase. Most recently, several works investigated deep learning for LV segmentation in echo [4,11,14]. In [11], anatomical priors based on the heart structure are used to regularize training of a deep network for segmentation of LV in 3D ultrasound. Also the works of [4,14] propose to use U-Net and its variations [13] for per frame segmentation of LV in echo cine.

Temporal information encoding is a key research problem in video analysis. Various methods in computer vision have shown that by combining temporal information with shape features, using tools such as recurrent neural networks and optical flow maps, the accuracy of video classification [8], segmentation [16], and interpretation [6] can be improved. Recently, in the area of medical imaging, adaptation of recurrent fully convolutional neural networks have shown promising results for detection of measurement points in echo [15], segmentation of the heart in CMR [12], and 3D biomedical image segmentation [5].

In this paper, we present a deep learning architecture for automatic segmentation of the LV from an entire echo cine. The individual frames of a cardiac echo cine are first processed by a U-Net encoder. The encoded temporal dependency information of the past frames are maintained via stacked bidirectional convolutional LSTM. Furthermore, temporal displacement information of moving objects between the consecutive frames is provided to the network by externally computed optical flow motion vectors. During the training phase, our method only requires LV annotation in ES and ED frames. Therefore, our architecture can be easily trained on most clinically obtained patient data without providing annotation beyond those that are normally recorded as part of standard-of-care in echo. In the test phase, our method can be used to infer accurate LV segmentation for the entire cine loop. Our method is quantitatively evaluated on an echo cine dataset consisting of 648 A4C echo cines that were gathered from 566 patients. We demonstrate that the proposed method can achieve a noticeably high segmentation accuracy of 97.9% with standard deviation of less than 1%.

2 Materials and Method

2.1 Dataset Information and Clinical Background

Our echo imaging data is collected from the Picture Archiving and Communication System at Vancouver General Hospital, with ethics approval of the Clinical Medical Research Ethics Board, in consultation with the Information Privacy Office. Our data consist of a collection of 648 A4C view echo studies from 566 patients, with about 34,000 total number of frames, captured by using Philips iE33 and GE Vivid-i/-7 ultrasound machines. In clinical practice, A4C is one of the primary standard views for LV EF estimation and other cardiac functions analyses. Each study was performed by an expert sonographer, where the LV boundary is traced in two frames (*i.e.*, ED and ES phase frames). The ED phase refers to the cardiac structure at the end of relaxing, *i.e.*, the end of ventricle loading, and the ES phase refers to the cardiac structure at the end of contraction, *i.e.*, the beginning of ventricle filling, respectively. We consider existing annotations at ED and ES phase as ground truth to train our model. In order to evaluate the performance of the model on the entire cine, we sought assistance from an experienced cardiologist, who helped us with annotation of a randomly selected frame between ED and ES frame in our test set. The cardiologist also validated our existing annotation of ED and ES frames by sonographers. An example of sample frames in our dataset and the corresponding cardiologist's annotation of LV segmentation is shown in Fig. 1.

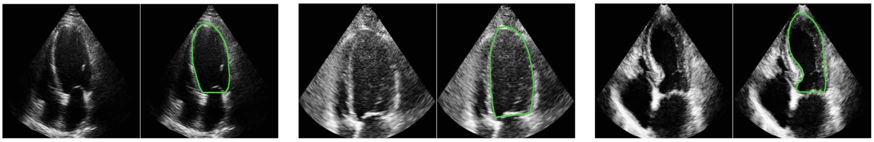

Fig. 1. Examples of sample frames and corresponding annotations.

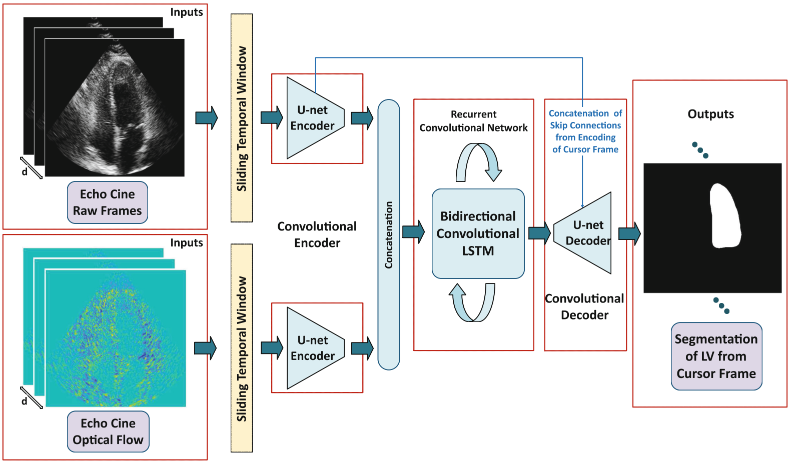

Fig. 2. Block diagram of the proposed architecture to integrate shape, temporal and motion information for LV segmentation in echo cine.

2.2 Network Architecture

The proposed LV segmentation architecture is depicted in Fig. 2, where the individual components of the pipeline are explained below.

Temporal Window: In the first stage, we define a collection of d consecutive frames as a temporal window. This set is fed to the network and the final output would be the segmentation mask of the last frame in the window. The last frame in the temporal window is called the “cursor” frame. The segmentation prediction of the entire cine can be obtained by sliding the model over the temporal dimension, with $stride = 1$.

U-Net Encoder: In the next stage of the network, we use U-Net’s [13] encoder schema to process the input echo frames. More specifically, each frame is passed through a number of stacked convolutional layers and pooling layers. A dense representation of the per-frame encoded features is obtained by the end of the encoding stage.

Optical Flow Integration: A second U-Net encoder model is used to process the optical flow motion vector maps between each pair of consecutive frames. We use the optical flow algorithm to track the motion of walls of heart chambers, providing the network with additional information for deriving segmentation. In our method, optical flow is calculated between each two consecutive frames in a temporal window with the use of Horn-Schunck algorithm [2]. Each optical flow input to the network is a two-channel image, showing the direction and distance of movement in both x and y axes. The processed optical flow information goes though a separate U-Net encoder, which is then concatenated with the intensity image encoded representation. Since the speckle motion of background tissue has a much lower velocity than the heart muscle motion, the motion of background tissue can be filtered out by convolutional layers in the U-Net encoder model.

Convolutional LSTM: In the third stage, the concatenated features from echo frames intensity information and optical flow maps are processed by a stack of two convolutional bi-directional recurrent long short term memory (Bi-directional LSTM) layers. Our intention of using convolutional Bi-directional LSTM comes in two-fold: (1) Bi-directional LSTM does not only encode temporal feature from the past context but also from the future context, which has been observed to handle noisy data well in speech recognition, thus making it a good candidate to handle noisy echo data; (2) the convolutional implementation of recurrent neural networks can capture spatio-temporal correlation better than conventional fully-connected recurrent neural networks, which based on our experiments, can be beneficial to localize the segmentation prediction.

U-Net Decoder: During the decoding stage, the representation generated from the Bi-directional LSTM is fed through a pipeline of up-sampling layers in order to obtain the final prediction of segmentation mask, where the architecture of the up-sampling layers is in the reverse order of the U-Net encoder architecture. The skip connections by-pass layers to connect an encoder feature map with corresponding decoder feature map of the same size. In each slide of the temporal window over echo clip, the output segmentation map corresponds to the LV location in the last frame of the temporal window.

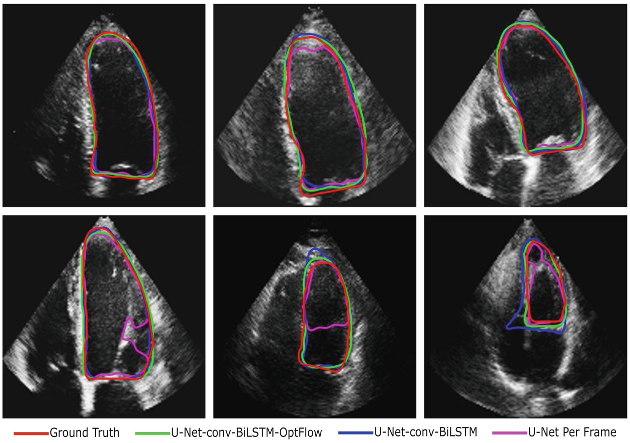

Fig. 3. Example LV segmentation results on six different subjects.

3 Experiments

The echo studies of the 566 patients are randomly assigned into training and test sets, with a split ratio of 80% and 20% of total amount of patients, respectively. This results in 453 patients (with 520 echo studies) in the training set and remaining 113 patients (with 128 echo studies) in the test group. Also, 20% of the training data is held as a validation set for cross-validation of the training hyper-parameters. The echo cine frames are resized to 128×128. The network is implemented in Keras with the use of Tensorflow (Google Corp., Mountain View, CA) backend. The network weights are initialized by using a normal distribution. ReLU activation is used in all constitutional layers of the network, and the activation in the prediction layer is a sigmoid function. Dice loss is used as the network's objective function. We use Adam optimizer with the learning rate of $1e-4$, and batch size of 10. Finally, d in the temporal window is set to 4 frames.

Testing Criteria: Note that in the standard clinical procedure, the LV tracing is routinely done in only the ED and ES frames of the A4C view, therefore we report the Dice score and accuracy on the ED and ES frames. In order to report segmentation accuracy for in-between frames, since developing per-frame ground truth for all echo cine frames is very time consuming, we approximated the full cine segmentation performance by evaluating the performance on a randomly selected frame between the ED and ES frames against an expert manual annotation. This frame is named RF (Random Frame) in Table 1.

Example visual results of the LV tracking by the proposed method compared to the ground truth are shown in Fig. 3. As can be seen, the proposed model accurately detects the LV wall and shape.

Table 1. Empirical evaluation of the proposed method. Best results are in bold.

Method	Dice Score(%)						Accuracy(%)					
	ED		RF		ES		ED		RF		ES	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
U-Net Per Frame	91.2	3.9	90.2	5.3	88.9	4.9	97.2	1.1	97.3	1.0	97.3	0.9
U-Net-conv-BiLSTM	93.3	3.4	92.1	3.8	90.1	8.8	97.8	0.9	97.7	1.0	97.8	1.0
U-Net-conv-BiLSTM-OptFlow	**93.6**	3.0	**92.5**	3.5	**92.1**	4.1	**97.9**	0.9	**97.8**	1.0	**97.9**	1.0

Model Comparison: We compare the performance of the proposed deep learning architecture (*i.e.*, U-Net-conv-BiLSTM-OptFlow) with the off-the-shelf 2D U-Net implementation [14] (*i.e.*, U-Net (Per Frame) in Table 1) that was trained with only the ED and ES frame segmentation ground truth, and also with an architecture of combining 2D U-Net with convolution Bi-directional LSTM (*i.e.*, U-Net-conv-BiLSTM), in Table 1. It is clear that the proposed architecture improves all segmentation metrics. In particular, the combination of U-Net and convolutional Bi-directional LSTM architecture consistently increases the Dice score on all ED, RF and ES frames. Furthermore, the integration of Bi-directional LSTM and optical flow information shows further improvement of segmentation performance. Most importantly, using optical flow information increases the robustness of LV tracking in echo data given the standard deviation of the reported results. Paired t-tests indicate there is a statistically significant difference between every pairs of the compared network architectures for both Dice

Score and Accuracy ($p < 0.05$). Also, in terms of area under the Receiver Operating Characteristic Curve (AUC), our analysis show U-Net per frame has substantially lower AUC (AUC = 0.94) than U-Net-conv-BiLSTM and U-Net-conv-BiLSTM-OptFlow (AUC = 0.97 for both methods). In addition, it can be seen in Fig. 3 that per frame U-Net can be misled by image artifacts and reduction in image quality. Both U-Net-conv-BiLSTM and U-Net-conv-BiLSTM-OptFlow, which utilize temporal information, show more consistent segmentation results.

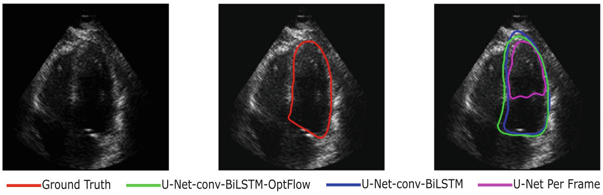

Fig. 4. Sample failed case of our method. Left to right: input echo frame, ground truth by cardiologist, and segmentation by the compared methods.

4 Conclusion and Discussion

Accurate LV segmentation in echocardiograms is an important component to diagnose critical cardiovascular disease. In this work, we present a method based on deep recurrent fully convolutional networks and optical flow for LV tracking in A4C echo cine data. We use convolutional Bi-directional LSTM to encode temporal information from a short number of frames. We also use optical flow information as an additional input to improve the segmentation accuracy and robustness. The proposed model is evaluated on an echo dataset consist of 648 echo studies from 566 patients, and shows advantageous over two compared models. Sample visual comparison of our proposed method can be seen in Fig. 3. The first row in Fig. 3 shows sample cases where all of the three compared methods provide an acceptable tracking of LV. The second row of Fig. 3 shows samples where U-Net per frame has been mislead by artifacts in echo data. Also, poor quality of captured echo in the cursor frame has resulted in high errors by per frame U-Net. This is while incorporating temporal and motion information in U-Net-conv-BiLSTM-Optflow results in a more smooth and accurate tracking of LV. The sample in the right column of the second row in Fig. 3 shows a case where adding information of optical flow has been advantageous comparing the blue contour (U-Net-conv-BiLSTM) with the green segmentation (U-Net-conv-BiLSTM-OptFlow). A sample failed case of our proposed method (U-Net-BiLSTM-OptFlow) is shown in Fig. 4. Captured echo with a low quality throughout the whole cine could be more challenging in terms of accurate segmentation of LV, as is the case with the shown sample. Low quality echo misses the location of the heart wall chambers and makes it hard to annotate LV even for expert human. Future work will include using the proposed architecture to automatically estimate various cardiac measurements, including the Left Ventricle Ejection Fraction.

References

1. World health organization. http://www.who.int/mediacentre/factsheets/fs317/en/
2. Achmad, B., Mustafa, M.M., Hussain, A.: Inter-frame enhancement of ultrasound images using optical flow. In: Badioze Zaman, H., Robinson, P., Petrou, M., Olivier, P., Schröder, H., Shih, T.K. (eds.) IVIC 2009. LNCS, vol. 5857, pp. 191–201. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-05036-7_19
3. Carneiro, G.: The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. IEEE Trans. Image Process. **21**(3), 968–982 (2012)
4. Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A., Zhou, S.K.: Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 487–495. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_56
5. Chen, J., et al.: Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: NIPS, pp. 3036–3044 (2016)
6. Li, Z.: Videolstm convolves, attends and flows for action recognition. Comput. Vis. Image Underst. **166**, 41–50 (2018)
7. Litjens, G.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
8. Ng, J.Y.H., et al.: Beyond short snippets: Deep networks for video classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4694–4702 (2015)
9. Ngo, T.A.: Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Med. Image Anal. **35**, 159–171 (2017)
10. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Trans. Med. Imaging **25**(8), 987–1010 (2006)
11. Oktay, O.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**(2), 384–395 (2018)
12. Poudel, R.P.K., Lamata, P., Montana, G.: Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In: Zuluaga, M.A., Bhatia, K., Kainz, B., Moghari, M.H., Pace, D.F. (eds.) RAMBO/HVSMR -2016. LNCS, vol. 10129, pp. 83–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52280-7_8
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Smistad, E., et al.: 2D left ventricle segmentation using deep learning. In: 2017 IEEE International Ultrasonics Symposium (IUS), pp. 1–4 (2017)

15. Sofka, M., Milletari, F., Jia, J., Rothberg, A.: Fully convolutional regression network for accurate detection of measurement points. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 258–266. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_30
16. Valipour, S., et al.: Recurrent fully convolutional networks for video segmentation. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 29–36 (2017)
17. Xue, W.: Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. **43**, 54–65 (2018)
18. Zreik, M., et al.: Automatic segmentation of the left ventricle in cardiac CT angiography using convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 40–43 (2016)