Recognizing End-Diastole and End-Systole Frames via Deep Temporal Regression Network

Bin Kong[1], Yiqiang Zhan[2], Min Shin[1], Thomas Denny[3], and Shaoting Zhang[1(✉)]

[1] Department of Computer Science, UNC Charlotte, Charlotte, NC, USA
szhang16@uncc.edu
[2] Siemens Healthcare, Malvern, PA, USA
[3] MRI Research Center, Auburn University, Auburn, AL, USA

Abstract. Accurate measurement of left ventricular volumes and Ejection Fraction from cine MRI is of paramount importance to the evaluation of cardiovascular functions, yet it usually requires laborious and tedious work of trained experts to interpret them. To facilitate this procedure, numerous computer aided diagnosis (CAD) methods and tools have been proposed, most of which focus on the left or right ventricle segmentation. However, the identification of ES and ED frames from cardiac sequences is largely ignored, which is a key procedure in the automated workflow. This seemingly easy task is quite challenging, due to the requirement of high accuracy (*i.e.*, precisely identifying specific frames from a sequence) and subtle differences among consecutive frames. Recently, with the rapid growth of annotated data and the increasing computational power, deep learning methods have been widely exploited in medical image analysis. In this paper, we propose a novel deep learning architecture, named as temporal regression network (TempReg-Net), to accurately identify specific frames from MRI sequences, by integrating the Convolutional Neural Network (CNN) with the Recurrent Neural Network (RNN). Specifically, a CNN encodes the spatial information of a cardiac sequence, and a RNN decodes the temporal information. In addition, we design a new loss function in our network to constrain the structure of predicted labels, which further improves the performance. Our approach is extensively validated on thousands of cardiac sequences and the average difference is merely 0.4 frames, comparing favorably with previous systems.

1 Introduction

Stroke volume (SV) and left ventricle ejection fraction, defined by the unnormalized and normalized difference between End-Diastole (ED) and End-Systole (ES) volumes respectively, are the most commonly used clinical diagnostic parameters for cardiac systolic funtion. This is because it reflects the contractile function of myocardium. However, current practice to calculate these parameters is mostly done manually by experts. Many previous methods have been devoted to automating this process, and the majority of them focus on left

S. Ourselin et al. (Eds.): MICCAI 2016, Part III, LNCS 9902, pp. 264–272, 2016.
DOI: 10.1007/978-3-319-46726-9_31

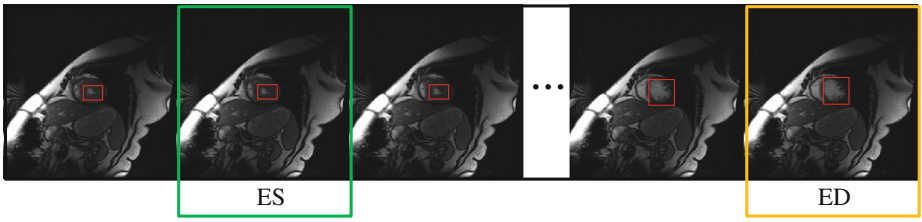

Fig. 1. A typical example of cardiac sequences (bright areas in red rectangles are left ventricles, the green and yellow rectangles indicate ES and ED frames respectively). (Color figure online)

ventricle segmentation [9,17]. However, the very first step of this automation, recognizing the ES and ED frames is largely ignored, while it is also an important process in the automatic system. In addition, even when SV and EF are computed manually or semi-automatically, reliably automating this step could reduce both inter and intra observer errors. Although the identification of ES and ED frames seems to be relatively easy, at least for human experts, the main challenges are the following: (1) the semantic gap between the high-level ES and ED concepts and low-level cardiac image sequence images, (2) the complex temporal relationships in cardiac cycles and subtle differences among consecutive cardiac frames (demonstrated in Fig. 1), and (3) the requirement of high accuracy since mislabeling even one frame may affect the diagnosis results. Therefore, determining ES and ED frames still remains a manual or semi-automatic task in many scenarios. Currently, this process could be time-consuming and error-prone, especially when dealing with large-scale datasets. It becomes a road block of a fully automatic solution.

Several attempts have been made to automate this process. A pioneer work [4] took advantage of rapid mitral opening in early diastole. However, it requires the identification by the user of three important landmarks: the apex and each angle of the mitral annulus, indicating a semi-automatic approach. Saeed Darvishi *et al.* [3] used a segmentation paradigm. In particular, they segmented every left ventricle region of the cardiac sequence by using level set. The frames corresponding to the largest ventricular area are the ED frames and the smallest ventricular area the ES frames. Since the initial contour has to be placed by the user, this method still remains semi-automatic. In addition, the final result largely relies on the quality of the initial contour. Another widely used method [6] tackled this problem with unsupervised learning. For this method, every frame of the cardiac sequence is embedded on a low-dimensional manifold, and a Euclidean distance between every two consecutive points in the manifold is computed to determine the ED and ES frames. However, a cardiac cycle is extremely complex and one individual's cardiac cycle may differ greatly from another's. Thus, this simple distance rule may not be applicable to other special patients, *e.g.*, those with cardiac diseases.

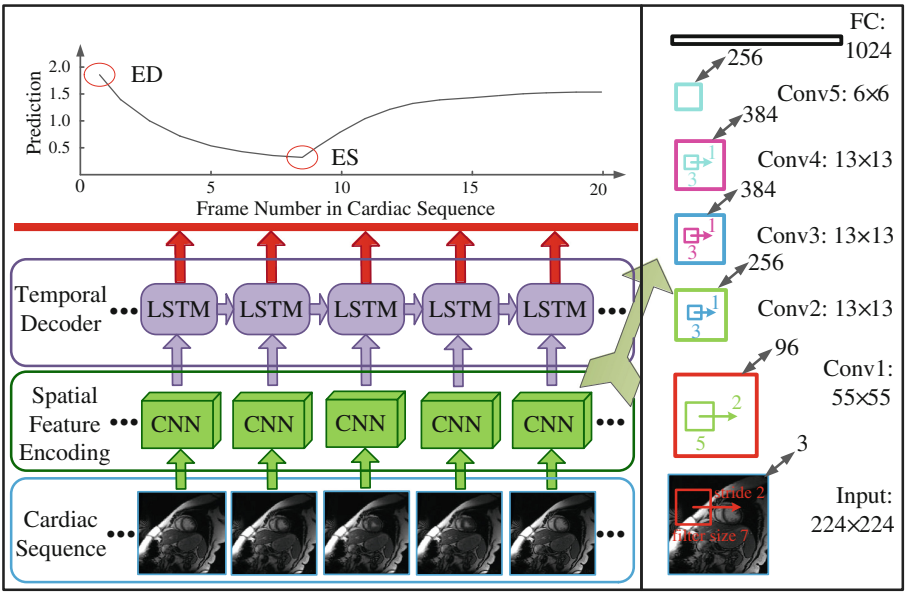

Fig. 2. An overview of the proposed framework, temporal regression network (TempReg-Net). Note that only convolutional layers are shown and Conv1 and Conv2 and Conv5 layers are followed by Pooling layers of size 3×3 and stride 2.

To overcome the above drawbacks, a joint network which combines Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) has been designed to automate the whole detection process. Specifically, our framework has two phases, *i.e.*, encoding and decoding. During the first phase, the CNN acts as an encoder to encode the spatial pattern of the cardiac image sequence, transforming every frame of the sequence into a fixed-length feature vector to facilitate the decoding phase. During the second phase, the RNN is used to decode the temporal information of the above mentioned feature vectors. The joint network can be trained to learn the complex spatial and temporal patterns of the cardiac sequences, and give predictions for the ED and ES frames during testing. The contribution of our work is twofold: (1) A deep temporal regression network is designed to recognize the ED and ES frames; and (2) A temporal structured loss is proposed to improve the accuracy of the network. Although deep learning has been widely used for medical image analysis [2,7,12,16], our network architecture is novel and carefully designed for this use case. This approach has several advantages compared to the previous methods: (1) No prior information or interaction is needed in the detection framework, since our system automatically learns everything from the patterns of the data. (2) Since RNN is able to learn long-term patterns, our framework can detect the complex and long temporal dynamics in the cardiac sequence.

2 Methodology

In this section, we provide an overview of our TempReg-Net framework. Then, we show that our framework can be trained end-to-end by jointly optimizing the regression and temporal structured constraints.

2.1 TempReg-Net Architectures

Figure 2 shows an overview of the proposed TempReg-Net framework, combining CNN and RNN (more specifically, the Long Short Term Memory (LSTM)). First, a feature encoder based on CNN is trained to encode the input into vectors. Then, the LSTM model takes over by exploring the temporal information sequentially. Finally, the ES and ED frames are detected according to the predictions from the LSTM model. At the training time, instead of using classification to identify the ES and ED frames, the network is trained to regress the location of the ES and ED frame numbers. During the testing phase, we examine the output sequence from TempReg-Net, where the ED frame is the local maximum and the ES frame is the local minimum.

Cardiac Frame Encoding with CNN: To fully capture the spatial information relevant to the left ventricle in every frame, we employ a CNN as the feature extractor in order to efficiently encode the spatial information. Recent years have witnessed numerous different kinds of CNN architectures. The Zeiler-Fergus (ZF) model is employed in our framework. The architecture of ZF model is illustrated in Fig. 2 (right). The reason of our choice is twofold: (1) Leveraging transferred knowledge across similar tasks is very useful when the labeled data is not adequate [14] and the architecture proposed in [5] achieved intriguing results in several image sequence analysis tasks; (2) the ZF model is reasonably deep and produces prominent results so we have a balance between computational complexity and the results. Essentially, a CNN acts as a feature transformer $\psi(S;V)$ parametrized by V to map cardiac sequence S to fixed-length vector sequence representations $< x_1, x_2, ..., x_T >$, in which V is the learnt weights of the CNN model and $x_t \in \mathbb{R}^q$ ($t = 1, 2, ..T$ and $q = 1024$ in our experiments).

Recognizing ED and ES Frames via RNN: Temporal information in cardiac sequence provides contextual clues regarding left ventricle volume changes. We tap into the temporal dimension by passing the above mentioned feature vector sequence into a RNN model. Instead of using traditional vanilla RNN, the LSTM model is adopted to avoid the vanishing or exploding gradients problem during back-propagation. The difference between a LSTM model

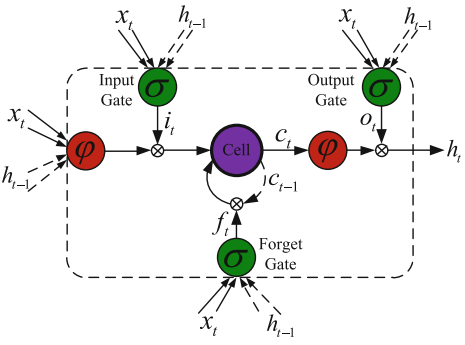

Fig. 3. A diagram of a LSTM memory block.

and a vanilla RNN is that a LSTM contains memory blocks instead of regular network units. A slightly simplified LSTM memory block [15] is used in this article, shown in Fig. 3. Benefited from the memory blocks, a LSTM learns when to forget previous hidden states h_{t-1} and when to update them given new information, shown as:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \varphi(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ h_t &= o_t \odot \varphi(c_t) \end{aligned} \tag{1}$$

where $\varphi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and $\sigma(x) = (1 + e^{-x})^{-1}$ are nonlinear functions which squash its inputs to $[-1, 1]$. i_t, f_t, o_t are input gate, forget gate and output gate respectively. $\odot$ denotes element-wise product.

The memory cell c_t is a function of the previous memory cell c_{t-1} and the current input x_t and the previous hidden state h_{t-1}. i_t and f_t enable the memory cell to selectively forget its previous memory or consider new input. These additional units enable the LSTM to learn very complex temporal dynamics.

The final step in estimating a regression at time t is to take a fully-connected layer over the output of the RNN. The sequential weights W are reused at every frame, forcing the model to learn generic time-to-time cardiac motion dynamics and preventing the parameter size from growing linearly with respect to the sequence length T.

2.2 Jointly Optimize the Regression and Temporal Structured Constraints

Essentially, TempReg-Net gives a prediction for every single frame in a cardiac sequence and there is no constraint among the prediction sequences. However, TempReg-Net is designed to model the left ventricle volumes in a cardiac cycle, *i.e.*, the predictions for a cardiac sequence should decrease during the systole phase and increase during the diastole phase. Solely doing a regression cannot ensure such a structured output. In order to address this problem, we explicitly model this sequential constraint by penalizing predictions with wrong structures. Suppose that we are given the ground truth label y, which will be discussed later, and the TempReg-Net regressor η. Ideally, given two consecutive ground truth labels y_{k-1} and y_k and $y_{k-1} < y_k$, *i.e.*, the kth frame is in a systole phase, we expect that the predictions for these two frames should subject to $\eta_{k-1} < \eta_k$ as well, and vice versa. To enforce this constraint in TempReg-Net, a new loss function which we name as temporal structured loss is defined:

$$
\begin{aligned}
L_{temp} &= \tfrac{1}{2}(L_{inc} + L_{dec}) \\
L_{inc} &= \tfrac{1}{T}\sum_{k=2}^{T} \mathbb{1}(y_k > y_{k-1}) \max(0, \eta_{k-1} - \eta_k) \\
L_{dec} &= \tfrac{1}{T}\sum_{k=2}^{T} \mathbb{1}(y_k < y_{k-1}) \max(0, \eta_k - \eta_{k-1})
\end{aligned}
\tag{2}
$$

where $\mathbb{1}(\cdot)$ is the indicator function. L_{inc} penalizes the decreasing predictions during a diastole phase, *i.e.*, the left ventricle volume is increasing. L_{dec} penalizes the increasing predictions during a systole phase, *i.e.*, the left ventricle volume is decreasing.

Having defined the temporal structured loss, the training criteria for TempReg-Net can be further explored. We denote the training example as (S, N_{es}, N_{ed}), where N_{es} and N_{ed} stand for the ES and ED frame numbers respectively in the sequence S. Given the training data $\mathbb{S} = \{S, N_{es}, N_{ed}\}$, the training objective becomes the task of estimating the network weights $\lambda = (U, V)$ (U and V are the parameters for the CNN and RNN, respectively):

$$
\begin{aligned}
\lambda &= \arg\min_{\lambda}\{\sum_{S\in\mathbb{S}}\sum_{k} ||y_k - \eta_k(S,\lambda)||^2 + \alpha L_{reg}(\lambda) + \beta L_{temp}\} \\
L_{reg}(\lambda) &= \tfrac{1}{2}(||U||_2^2 + ||V||_2^2)
\end{aligned}
\tag{3}
$$

where k is the kth frame of training sequence S. L_{reg} is the regularization penalty term which ensures the learnt weights are sparse. α and β are hyper-parameters which are cross-validated in our experiments. At training stage, the ground truth label y_k is synthesised to imitate the left ventricle volume changes during a typical cardiac cycle [3]:

$$
y_k = \begin{cases} \left|\frac{k-N_{es}}{N_{es}-N_{ed}}\right|^{\delta}, & \text{if} \quad N_{ed} < k \leq N_{es} \\ \left|\frac{k-N_{es}}{N_{es}-N_{ed}}\right|^{\upsilon}, & \text{otherwise} \end{cases}
\tag{4}
$$

where δ and υ are hyper-parameters, set as 3 and $\frac{1}{3}$ respectively to imitate the typical left ventricle volume changes in cardiac cycle.

3 Experiments

Experimental Setting and Implementations: Our experiments are conducted on MRI sequences, acquired from our collaborative hospital and labeled by experts. Specifically, this dataset comprises of cardiac sequences (consists of around 113,000 cardiac frames) gathered from 420 patients, from different imaging views and different positions, including long-axis, short-axis, four-chamber and two-chamber views. There are about 18 cardiac sequences for each patient (around 15 for short-axis view, and one for long-axis, four-chamber and two-chamber views, respectively). Every sequence has 20 frames with 256 × 256 pixels. ED and ES frames are carefully identified by cardiologists. Four-fold cross-validation is performed to obtain quantitative results.

Regarding implementation, TempReg-Net's implementation is based on Caffe [8]. In order to fully utilize the CNN architecture and make use of transferred knowledge, we squash all gray-scale cardiac frames to the range of $[0, 255]$, and these single-channel cardiac frames are replicated for three times, resulting in three-channel images. In order to get a reasonable initialization and avoid over-fitting, we fine-tune our TempReg-Net on a pre-trained model based on the 1.2M ImageNet [10]. In our experiment, the learning rate for the last fully-connected layer is set to be 1×10^{-3}, which is 10 times larger than the rest layers. Regarding the RNN, the LSTM is used to avoid vanishing and exploding gradient problems. All the parameters of the LSTM are randomly initialized within $[-0.01, 0.01]$. Since each cardiac sequence in our dataset contains 20 frames, the LSTM is unrolled to 20 time steps. All the hyper-parameters of the proposed TempReg-Net are cross-validated for the best results. During the training stage, we augment our datasets by randomly cropping the resized cardiac images. The whole network is trained end-to-end with back propagation.

Results and Analysis: To quantitatively evaluate our method, we use average Frame Difference (*aFD*) to quantify the error of the prediction, following the convention [6,11]. Assuming that the frame label for the mth patient is N_m and the predicted frame label is $\hat{N}_m$, *aFD* can be defined as:

$$aFD = \frac{1}{M} \sum_{m=1}^{M} |\hat{N}_m - N_m|, \tag{5}$$

where M is the total number of examples in the testing set.

Table 1 shows the evaluation of our framework. Even without using the temporal structured constraints (TSC), our TempReg-Net is already a competitive solution to detect ED and ES from MRI. It has achieved good performance, *i.e.*, *aFD* 0.47 and 0.52 for identifying ED and ES, respectively, meaning that the error is within one frame. This is a promising result considering that our framework is end-to-end and automatic, with no requirement for interactions. After adding the temporal structured constraints, the *aFD* is improved to 0.38 and 0.44 for ED and ES, reducing the errors by around 15 %. This shows that the structures enforced upon the predictions contribute positively to the network. Regarding the computational efficiency, this framework takes merely 1.4 seconds to process one sequence. Therefore, it has the potential to be integrated with cardiac analysis systems owing to the small overhead.

We also compare our method with other types of systems or algorithms. For example, the system in [1] first segments the left ventricle, and then identifies the ED and ES by measuring areas of segmented regions. We have developed a similar system, using variations of level set (used in [3]) or graph cut (used in [13]) to segment the left ventricle. This type of segmentation-based system is very intuitive and widely used. However, compared to our solution, it has several limitations, including the computational efficiency, human interactions, and the segmentation errors. In our experiments, the system takes 2.9 and 3.5 seconds to segment the left ventricle from one sequence using level set and graph

Table 1. Average frame differences, standard deviation and running time of different methods.

Methods		Seg-based: Level Set [3]	Seg-based: Graph Cut [13]	Reg-based: CNN + Reg	TmpReg-Net (without TSC)	TempReg-Net
aFD	ED	1.54	2.27	1.30	0.47	**0.38**
	ES	1.24	1.65	1.97	0.52	**0.44**
STD	ED	1.93	2.89	1.77	0.49	**0.39**
	ES	1.64	1.96	2.42	0.53	**0.46**
Time (s)		2.9	3.5	1.5	1.4	**1.4**

cut, respectively, which are slower than our method. Note that we do not count the time of human interactions to initialize the segmentation procedure (e.g., graph cut method needs to specify foreground and/or background), which could take extra time. The *aFD* is 1.54 for ED and 1.24 for ES when using level set, and 2.27 as well as 1.65 when using graph cut, both of which are much worse than ours. The reason is that these segmentation algorithms cannot perfectly segment left ventricles in all frames, while even small errors adversely affect the prediction results based on the subtle difference of areas. Moreover, a similar regression framework is implemented. In this framework, The only difference is that logistic regression is used to predict the ventricle volumes. The *aFD* is 1.30 for ED and 1.97 for ES when using this method, still worse than the proposed method. Note that [13] is a recently proposed method and it has achieved sound accuracy for the segmentation of myocardium. The standard deviation is 0.39 and 0.46 for ED and ES respectively when using the proposed methods, which compares favorably against other methods. This further proves its effectiveness. Therefore, our end-to-end deep learning solution is more competitive in this task.

4 Conclusion

In this paper, we proposed a novel deep neural network, TempReg-Net, by integrating the CNN and RNN to identify specific frames in a cardiac sequence. In our method, a CNN and RNN tap into the spatial and temporal information respectively. Since the predictions should be temporally ordered, we explicitly model this constraint by adding a novel loss function in our framework. Extensive experiments on cardiac sequences demonstrate the efficacy of the proposed method. As deep learning methods have advanced segmentation tasks as well, future work will be devoted to develop a segmentation framework to fully automate the calculation of cardiac functional parameters.

References

1. Abboud, A.A., et al.: Automatic detection of the end-diastolic and end-systolic from 4d echocardiographic images. JCS **11**(1), 230 (2015)

2. Chen, H., Dou, Q., Ni, D., Cheng, J.-Z., Qin, J., Li, S., Heng, P.-A.: Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 507–514. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_62
3. Darvishi, S., et al.: Measuring left ventricular volumes in two-dimensional echocardiography image sequence using level-set method for automatic detection of end-diastole and end-systole frames. Res. Cardiocvasc. Med. **2**(1), 39 (2013)
4. Dominguez, C.R., Kachenoura, N., Mulé, S., Tenenhaus, A., Delouche, A., Nardi, O., Gérard, O., Diebold, B., Herment, A., Frouin, F.: Classification of segmental wall motion in echocardiography using quantified parametric images. In: Frangi, A.F., Radeva, P., Santos, A., Hernandez, M. (eds.) FIMH 2005. LNCS, vol. 3504, pp. 477–486. Springer, Heidelberg (2005)
5. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR, pp. 2625–2634 (2015)
6. Gifani, P., et al.: Automatic detection of end-diastole and end-systole from echocardiography images using manifold learning. PMEA **31**(9), 1091 (2010)
7. Greenspan, H., et al.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE TMI **35**(5), 1153–1159 (2016)
8. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: ACMMM, pp. 675–678. ACM (2014)
9. Marino, M., et al.: Fully automated assessment of left ventricular volumes, function and mass from cardiac MRI. In: CinC, pp. 109–112. IEEE (2014)
10. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
11. Shalbaf, A., et al.: Automatic detection of end systole and end diastole within a sequence of 2-d echocardiographic images using modified isomap algorithm. In: MECBME, pp. 217–220. IEEE (2011)
12. Shin, H.C., et al.: Interleaved text/image deep mining on a very large-scale radiology database. In: CVPR, pp. 1090–1099 (2015)
13. Uzunbaş, M.G., et al.: Segmentation of myocardium using deformable regions and graph cuts. In: ISBI, pp. 254–257. IEEE (2012)
14. Yosinski, J., et al.: How transferable are features in deep neural networks? In: NIPS, pp. 3320–3328 (2014)
15. Zaremba, W., et al.: Learning to execute. CoRR abs/1410.4615 (2014)
16. Zhang, W., et al.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage **108**, 214–224 (2015)
17. Zhen, X., et al.: Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. MedIA **30**, 120–129 (2015)