Recurrent Neural Networks for Aortic Image Sequence Segmentation with Sparse Annotations

Wenjia Bai[1(✉)], Hideaki Suzuki[2], Chen Qin[1], Giacomo Tarroni[1], Ozan Oktay[1], Paul M. Matthews[2,3], and Daniel Rueckert[1]

[1] Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, UK
w.bai@imperial.ac.uk

[2] Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK

[3] UK Dementia Research Institute, Imperial College London, London, UK

Abstract. Segmentation of image sequences is an important task in medical image analysis, which enables clinicians to assess the anatomy and function of moving organs. However, direct application of a segmentation algorithm to each time frame of a sequence may ignore the temporal continuity inherent in the sequence. In this work, we propose an image sequence segmentation algorithm by combining a fully convolutional network with a recurrent neural network, which incorporates both spatial and temporal information into the segmentation task. A key challenge in training this network is that the available manual annotations are temporally sparse, which forbids end-to-end training. We address this challenge by performing non-rigid label propagation on the annotations and introducing an exponentially weighted loss function for training. Experiments on aortic MR image sequences demonstrate that the proposed method significantly improves both accuracy and temporal smoothness of segmentation, compared to a baseline method that utilises spatial information only. It achieves an average Dice metric of 0.960 for the ascending aorta and 0.953 for the descending aorta.

1 Introduction

Segmentation is an important task in medical image analysis. It assigns a class label to each pixel/voxel in a medical image so that anatomical structures of interest can be quantified. Recent progress in machine learning has greatly improved the state-of-the-art in medical image segmentation and substantially increased accuracy. However, most of the research so far focuses on static image segmentation, whereas segmentation of temporal image sequences has received less attention. Image sequence segmentation plays an important role in assessing the anatomy and function of moving organs, such as the heart and vessels. In

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 586–594, 2018.
https://doi.org/10.1007/978-3-030-00937-3_67

this work, we propose a novel method for medical image sequence segmentation and demonstrate its performance on aortic MR image sequences.

There are two major contributions of this work. First, the proposed method combines a fully convolutional network (FCN) with a recurrent neural network (RNN) for image sequence segmentation. It is able to incorporate both spatial and temporal information into the task. Second, we address the challenge of training the network from temporally sparse annotations. An aortic MR image sequence typically consists of tens or hundreds of time frames. However, manual annotations may only be available for a few time frames. In order to train the proposed network end-to-end from temporally sparse annotations, we perform non-rigid label propagation on the annotations and introduce an exponentially weighted loss function for training.

We evaluated the proposed method on an aortic MR image set from 500 subjects. Experimental results show that the method improves both accuracy and temporal smoothness of segmentation, compared to a state-of-the-art method.

1.1 Related Works

FCN and RNN. The FCN was proposed to tackle pixel-wise classification problems, such as image segmentation [1]. Ronnerberger et al. proposed the U-Net, which is a type of FCN that has a symmetric U-shape architecture for feature analysis and synthesis paths [2]. It has demonstrated remarkable performance in static medical image segmentation. The RNN was designed for handling sequences. The long short-term memory (LSTM) network is a type of RNN that introduces self-loops to enable the gradient flow for long durations [3].

In the domain of medical image analysis, the combination of FCN with RNN has been explored recently [4–9]. In some works, RNN was used to model the spatial dependency in static images [4–6], such as the inter-slice dependency in anisotropic images [4,5]. In other works, RNN was used to model the temporal dependency in image sequences [7–9]. For example, Kong et al. used RNN to model the temporal dependency in cardiac MR image sequences and to predict the cardiac phase for each time frame [7]. Xue et al. used RNN to estimate the left ventricular areas and wall thicknesses across a cardiac cycle [8]. Huang et al. used RNN to estimate the location and orientation of the heart in ultrasound videos [9]. These works on medical image sequence analysis [7–9] mainly used RNN for image-level regression. The contribution of our work is that instead of performing regression, we integrate FCN and RNN to perform pixel-wise segmentation for medical image sequences.

Sparse Annotations. Manual annotation of medical images is time-consuming and tedious. It is normally performed by image analysts with clinical knowledge and not easy to outsource. Consequently, we often face small or sparse annotation sets, which is a challenge for training a machine learning algorithm, especially neural networks. To learn from spatially sparse annotations, Cicek et al. proposed to assign a zero weight to unlabelled voxels in the loss function [10]. In this work, we focus on learning from temporally sparse annotations and

address the challenge by performing non-rigid label propagation and introducing an exponentially weighted loss function.

Aortic Image Segmentation. For aortic image sequence segmentation, a deformable model approach has been proposed [11], which requires a region of interest and the centre of aorta to be manually defined in initialisation. This work proposes a fully automated segmentation method.

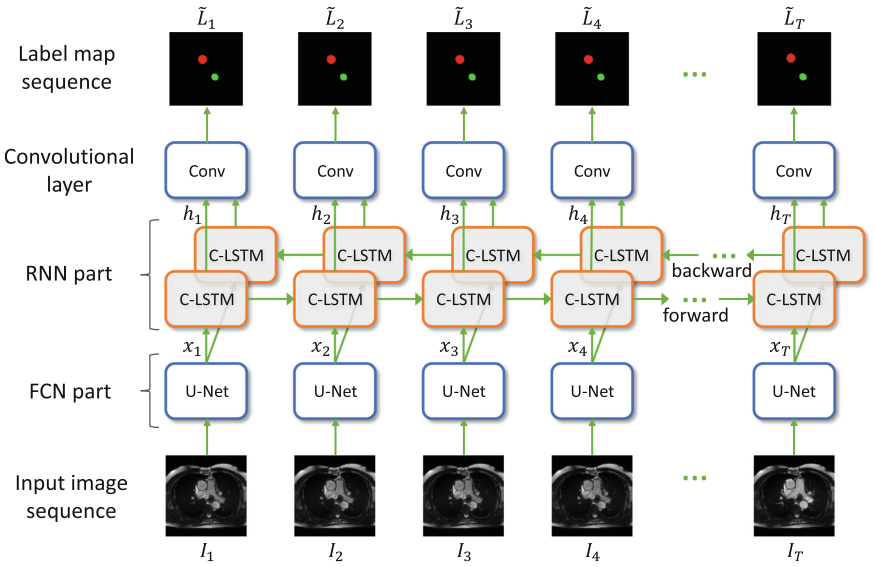


Fig. 1. The proposed method analyses spatial features in the input image sequence using U-Net, extracts the second last layer of U-Net as feature maps x_t, connects them using convolutional LSTM (C-LSTM) units across the temporal domain and finally predicts the label map sequence.

2 Methods

2.1 Network Architecture

Figure 1 shows the diagram of the method. The input is an image sequence $I = \{I_t | t = 1, 2, \ldots, T\}$ across time frames t and the output is the predicted label map sequence $\tilde{L} = \{\tilde{L}_t | t = 1, 2, \ldots, T\}$. The method consists of two main parts, FCN and RNN. The FCN part analyses spatial features in each input image I_t and extracts a feature map x_t. We use the U-Net architecture [2] for the FCN part, which has demonstrated good performance in extracting features for image segmentation.

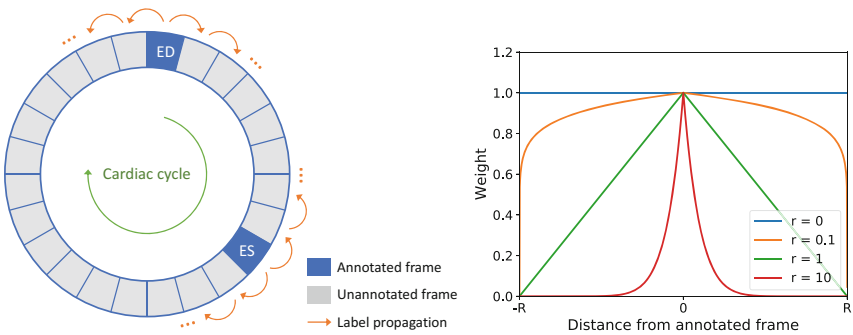
The second last layer of the U-Net [2] is extracted as the feature map x_t and fed into the RNN part. For analysing temporal features, we use the convolutional LSTM (C-LSTM) [12]. Compared to the standard LSTM which analyses

one-dimensional signals, C-LSTM is able to analyse multi-dimensional images across the temporal domain. Each C-LSTM unit is formulated as:

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

where $*$ denotes convolution[1], $\odot$ denotes element-wise multiplication, $\sigma(\cdot)$ denotes the sigmoid function, i_t, f_t, c_t and o_t are respectively the input gate (i), forget gate (f), memory cell (c) and output gate (o), W and b denote the convolution kernel and bias for each gate, x_t and h_t denote the input feature map and output feature map. The equation shows that output h_t at time point t is determined by both the current input x_t and the previous states c_{t-1} and h_{t-1}. In this way, C-LSTM utilises past information during prediction. In the proposed method, we use bi-directional C-LSTM, which consists of a forward stream and a backward stream, as shown in Fig. 1, so that the network can utilise both past and future information.

The output of C-LSTM is a pixel-wise feature map h_t at each time point t. To predict the probabilistic label map $\tilde{L}_t$, we concatenate the outputs from the forward and backward C-LSTMs and apply a convolution to it, followed by a softmax layer. The loss function at each time point is defined as the cross-entropy between the ground truth label map L_t and the prediction $\tilde{L}_t$.



(a) Label propagation (b) Weighting function

Fig. 2. Label propgation and the weighting function for propagated label maps.

2.2 Label Propagation and Weighted Loss

To train the network end-to-end, we require the ground truth label map sequence across the time frames. However, the typical manual annotation is temporally

[1] The standard LSTM performs multiplication instead of convolution here.

sparse. For example, in our dataset, we only have manual annotations at two time frames, end-diastole (ED) and end-systole (ES). In order to obtain the annotations at other time frames, we perform label propagation. Non-rigid image registration [13] is performed to estimate the motion between each pair of successive time frames. Based on the motion estimate, the label map at each time frame is propagated from either ED or ES annotations, whichever is closer, as shown in Fig. 2(a).

Registration error may accumulate during label propagation. The further a time frame is from the original annotation, the larger the registration error might be. To account for the potential error in propagated label maps, we introduce a weighted loss function for training,

$$E(\theta) = \sum_t w(t-s) \cdot f(L_t, \tilde{L}_t(\theta)) \tag{2}$$

where θ denotes the network parameters, $f(\cdot)$ denotes the cross-entropy between the propagated label map L_t and the predicted label map $\tilde{L}_t(\theta)$ by the network, s denotes the nearest annotated time frame to t and $w(\cdot)$ denotes an exponential weighting function depending on the distance between t and s,

$$w(t-s) = (1 - \frac{|t-s|}{R})^r \tag{3}$$

where R denotes the radius of the time window T for the unfolded RNN and the exponent r is a hyper-parameter which controls the shape of the weighting function. Some typical weighting functions are shown in Fig. 2(b). If $r = 0$, it treats all the time frames equally. If $r > 0$, it assigns a lower weight to time frames further from the original annotated frame.

2.3 Evaluation

We evaluate the method performance in two aspects, segmentation accuracy and temporal smoothness. For segmentation accuracy, we evaluate the Dice overlap metric and the mean contour distance between automated segmentation and manual annotation at ED and ES time frames. We also calculate the aortic area and report the difference between automated measurement and manual measurement. For evaluating temporal smoothness, we plot the curve of the aortic area $A(t)$ against time, as shown in Fig. 4, calculate the curvature of the time-area curve, $\kappa(t) = \frac{|A''(t)|}{(1+A'^2(t))^{1.5}}$, and report the mean curvature across time.

3 Experiments and Results

3.1 Data and Annotations

We performed experiments on an aortic MR image set of 500 subjects, acquired from the UK Biobank. The typical image size is 240×196 pixel with the spatial

resolution of $1.6 \times 1.6\,\mathrm{mm}^2$. Each image sequence consists of 100 time frames, covering the cardiac cycle. Two experienced image analysts manually annotated the ascending aorta (AAo) and descending aorta (DAo) at ED and ES time frames. The image set was randomly split into a training set of 400 subjects and a test set of 100 subjects. The performance is reported on the test set.

3.2 Implementation and Training

The method was implemented using Python and Tensorflow. The network was trained in two steps. In the first step, the U-Net part was trained for static image segmentation using the Adam optimiser for 20,000 iterations with a batch size of 5 subjects. The initial learning rate was 0.001 and it was divided by 10 after 5,000 iterations. In the second step, the pre-trained U-Net was connected with the RNN and trained together end-to-end using image and propagated label map sequences for 20,000 iterations with the same learning rate settings but a smaller batch size of 1 subject due to GPU memory limit. Data augmentation was performed online, which applied random translation, rotation and scaling to each input image sequence. Training took ∼22 h on a Nvidia Titan Xp GPU. At test time, it took ∼10 s to segment an aortic MR image sequence.

3.3 Network Parameters

There are a few parameters for the RNN, including the length of the time window T after unfolding the RNN and the exponent r for the weighting function. We investigated the impact of these parameters. Table 1 reports the average Dice metric when the parameters vary. It shows that a combination of time window $T = 9$ and exponent $r = 0.1$ achieves a good performance. When the time window increases to 21, the performance slightly decreases, possibly because the accumulative error of label propagation becomes larger. The exponent $r = 0.1$ outperforms $r = 0$, the latter treating the annotated frames and propagated frames equally, without considering the potential propagation error.

Table 1. Mean dice overlap metrics of the aortas when parameters vary.

(a) Varying T ($r = 0.1$)

T	AAo	DAo
5	0.959	0.952
9	**0.960**	**0.953**
13	0.959	0.950
17	0.959	0.952
21	0.958	0.951

(b) Varying r ($T = 9$)

r	AAo	DAo
0	0.955	0.949
0.1	**0.960**	**0.953**
1.0	0.959	0.951
10.0	0.959	0.948
100.0	**0.960**	0.949

Table 2. Quantitative comparison to U-Net. The columns list the mean dice metric, contour distance error, aortic area error and time-area curve curvature.

	Dice metric		Dist. error (mm)		Area error (mm^2)		Curvature	
	AAo	DAo	AAo	DAo	AAo	DAo	AAo	DAo
U-Net	0.953	0.944	0.80	0.69	51.68	35.96	0.47	0.38
Proposed	**0.960**	**0.953**	**0.67**	**0.59**	**39.61**	**27.98**	**0.41**	**0.28**

3.4 Comparison to Baseline

We compared the proposed method to the U-Net [2], which is a strong baseline method. U-Net was applied to segment each time frame independently. Figure 3

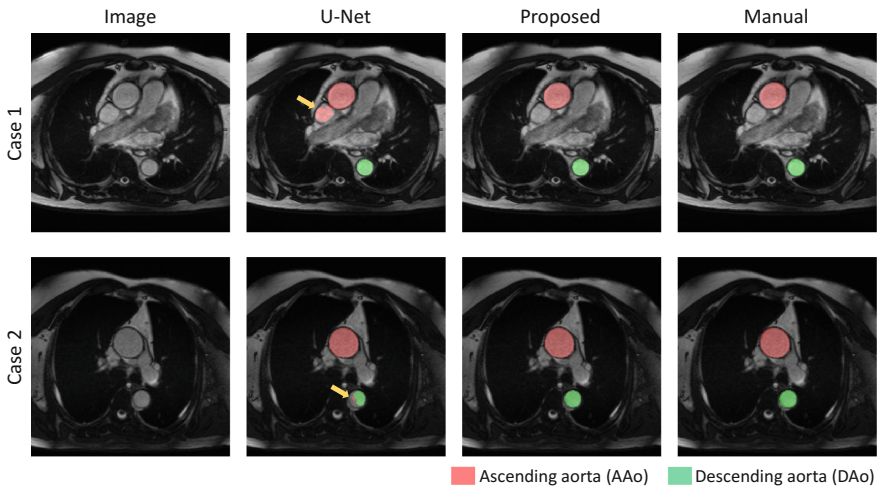
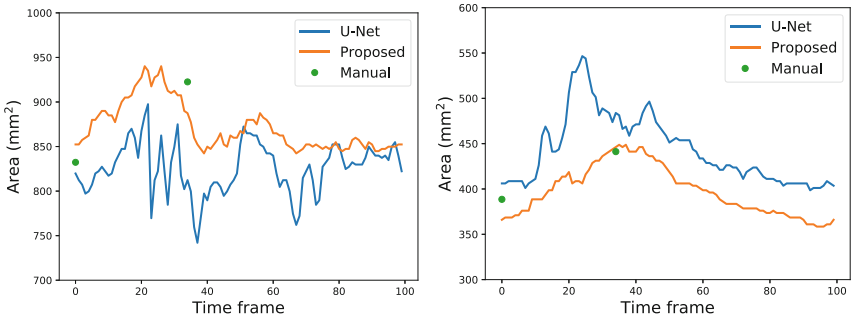


Fig. 3. Comparison of the segmentation results for U-Net and the proposed method. The yellow arrows indicate segmentation errors made by U-Net.



(a) Ascending aorta (b) Descending aorta

Fig. 4. Comparison of aortic time-area curves. The green dots indicate the manual measurements at ED and ES time frames.

compares the segmentation results on two exemplar cases. In Case 1, the U-Net misclassifies a neighbouring vessel as the ascending aorta. In Case 2, the U-Net under-segments the descending aorta. For both cases, the proposed method correctly segments the aortas. Figure 4 compares the time-area curves of the two methods on a exemplar subject. It shows that the curve produced by the proposed method is temporally smoother with less abrupt changes. Also, the curve agrees well with the manual measurements at ED and ES. Table 2 reports the quantitative evaluation results for segmentation accuracy and temporal smoothness. It shows that the proposed method outperforms the U-Net in segmentation accuracy, achieving a higher Dice metric, a lower contour distance error and a lower aortic area error (all with $p < 0.001$ in paired t-tests). In addition, the proposed method reduces the curvature of the time-area curve ($p < 0.001$), which indicates improved temporal smoothness.

4 Conclusions

In this paper, we propose a novel method which combines FCN and RNN for medical image sequence segmentation. To address the challenge of training the network with temporally sparse annotations, we perform non-rigid label propagation and introduce an exponentially weighted loss function for training, which accounts for potential errors in label propagation. We evaluated the method on aortic MR image sequences and demonstrated that by incorporating spatial and temporal information, the proposed method outperforms a state-of-the-art baseline method in both segmentation accuracy and temporal smoothness.

Acknowledgements. This research has been conducted using the UK Biobank Resource under Application Number 18545. This work is supported by the SmartHeart EPSRC Programme Grant (EP/P001009/1). We would like to acknowledge NVIDIA Corporation for donating a Titan Xp for this research. P.M.M. thanks the Edmond J. Safra Foundation, Lily Safra and the UK Dementia Research Institute for their generous support.

References

1. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
2. 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
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
4. Chen, J., et al.: Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. In: NIPS, pp. 3036–3044 (2016)

5. 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
6. Yang, X., et al.: Towards automatic semantic segmentation in volumetric ultrasound. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 711–719. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_81
7. Kong, B., Zhan, Y., Shin, M., Denny, T., Zhang, S.: Recognizing end-diastole and end-systole frames via deep temporal regression network. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 264–272. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46726-9_31
8. Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J., Li, S.: Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 276–284. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_32
9. Huang, W., Bridge, C.P., Noble, J.A., Zisserman, A.: Temporal HeartNet: towards human-level automatic analysis of fetal cardiac screening video. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 341–349. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_39
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
11. Herment, A., et al.: Automated segmentation of the aorta from phase contrast MR images: validation against expert tracing in healthy volunteers and in patients with a dilated aorta. J. Mag. Reson. Imag. **31**(4), 881–888 (2010)
12. Stollenga, M.F., et al.: Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: NIPS, pp. 2998–3006 (2015)
13. Rueckert, D., et al.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imag. **18**(8), 712–721 (1999)