Segmentation of 4D images via space-time neural networks

Sun, Changjian, Udupa, Jayaram, Tong, Yubing, Sin, Sanghun, Wagshul, Mark, et al.

Changjian Sun, Jayaram K. Udupa, Yubing Tong, Sanghun Sin, Mark Wagshul, Drew A. Torigian, Raanan Arens, "Segmentation of 4D images via space-time neural networks," Proc. SPIE 11317, Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, 113170J (28 February 2020); doi: 10.1117/12.2549605

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Segmentation of 4D images via space-time neural networks

Changjian Sun[1, 2], Jayaram K. Udupa[2*], Yubing Tong[2], Sanghun Sin[3], Mark Wagshul[4], Drew A. Torigian[2], Raanan Arens[3]

[1] College of Electronic Science and Engineering, Jilin University, Changchun, China.

[2] Medical Image Processing Group, 602 Goddard building, 3710 Hamilton Walk, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, United States.

[3] Division of Respiratory and Sleep Medicine, The Children's Hospital at Montefiore, Albert Einstein College of Medicine, Bronx, New York 10467, United States.

[4] Department of Radiology, Gruss MRRC, Albert Einstein College of Medicine, Bronx, New York 10467, United States.

ABSTRACT

Medical imaging techniques currently produce 4D images that portray the dynamic behaviors and phenomena associated with internal structures. The segmentation of 4D images poses challenges different from those arising in segmenting 3D static images due to different patterns of variation of object shape and appearance in the space and time dimensions. In this paper, different network models are designed to learn the pattern of slice-to-slice change in the space and time dimensions independently. The two models then allow a gamut of strategies to actually segment the 4D image, such as segmentation following just the space or time dimension only, or following first the space dimension for one time instance and then following all time instances, or vice versa, etc. This paper investigates these strategies in the context of the obstructive sleep apnea (OSA) application and presents a unified deep learning framework to segment 4D images. Because of the sparse tubular nature of the upper airway and the surrounding low-contrast structures, inadequate contrast resolution obtainable in the magnetic resonance (MR) images leaves many challenges for effective segmentation of the dynamic airway in 4D MR images. Given that these upper airway structures are sparse, a Dice coefficient (DC) of ~0.88 for their segmentation based on our preferred strategy is similar to a DC of >0.95 for large non-sparse objects like liver, lungs, etc., constituting excellent accuracy.

1. INTRODUCTION

Medical imaging techniques currently produce 4D images that portray the dynamic behaviors and phenomena associated with internal structures. The segmentation of 4D images poses challenges different from those arising in segmenting 3D static images due to different patterns of variation of object shape and appearance in the space and time dimensions. For the segmentation of 4D medical image objects, the common method is to segment the 2D data set of the 4D medical image, given that a 4D medical image is a set of 2D images that tracks time and space. This kind of method cannot take advantage of the texture change information between successive slices in the temporal and spatial directions. Another commonly used method is to consider all the 3D volumes at all time points simultaneously, using a continuous 3D volume as the segmentation target, but this method often requires complicated calculations or heavy manual intervention [1].

Our application of focus is obstructive sleep apnea (OSA), a common and serious health problem in both adults and children associated with partial or complete upper airway collapse during sleep [2]. When studying such a condition, it is important to consider the dynamic characteristics of the upper respiratory tract. The general way of analyzing OSA medical images is to establish a patient-specific upper respiratory tract biomechanical model by making use of the anatomic information derived from the patient images to simulate airway dynamics [3]. The model parameters and behaviors are

Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, edited by Andrzej Krol, Barjor S. Gimi, Proc. of SPIE Vol. 11317, 113170J · © 2020 SPIE · CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2549605

used to characterize OSA. Dynamic magnetic resonance imaging (MRI) is the preferred method for studying the upper respiratory tract in these diseases [4]. Unfortunately, because of the sparse tubular nature of the upper airway and the surrounding low-contrast structures, inadequate contrast resolution obtainable in the MR images leaves many challenges for effective segmentation of the dynamic airway in 4D MR images [5].

In this paper, different network models are designed to learn the pattern of slice-to-slice change in the space and time dimensions independently. The two models then allow a gamut of strategies to actually segment the 4D image, such as segmentation following just the space or time dimension only, or following first the space dimension for one time instance and then following all time instances, or vice versa, etc. This paper investigates these strategies in the context of the OSA application and presents a unified deep learning framework to segment 4D images. The strategy of combing different networks will affect the segmentation accuracy. By comparing the segmentation accuracy of different network combinations, we can provide a basis for future segmentation network selection.

2. MATERIALS AND METHODS

Image data

The 4D acquisition protocol consists of a T1-weighted inversion prepared gradient recalled echo sequence, acquired in the sagittal plane. MR images utilized were acquired by a retrospective gating method [4]. Image data acquisition was triggered only if the input respiratory signal was within predefined temporal tolerances. Abnormal volumes arising due to events such as swallowing and deep inhalation were discarded. Images were collected on a 3T Achieva MRI scanner (Philips Center, Amsterdam, The Netherlands). Thirty-six 1.1-mm thick sagittal slices were acquired per subject. The slices were 240×240 pixels with a pixel size of 1×1 mm. 4D image data from 20 female subjects with OSA and each 4D image with 10 equally spaced time points over the respiratory cycle (a total of 200 3D volumes) were used in our experiments. Subjects were between 14 and 18 years of age.

ST-network strategy

4D image segmentation is different from 2D or 3D spatial image segmentation in that it involves both temporal and spatial characteristics which are usually independent and different. In other words, the manner in which shape changes from slice to slice spatially is different from the manner of shape change in slices along the time dimension. The idea behind the proposed space-time (ST) network is to have a separate space-network, denoted as S-network, to segment spatial slices for a given fixed time instance t, and analogously, a time-network, denoted as T-network, to segment time slices for a given fixed space location s. The premise (hypothesis) is that the patterns of change in the two directions are different and that it will be better to have separate networks to learn those patterns. For example, the slice-to-slice variations in the s-dimension portray how the object shape and appearance change spatially while the same in the t-dimension depict changes due to object dynamics.

We design and train separate deep S- and T-networks to carry out segmentation in a slice-by-slice manner under two basic strategies for each network (see Figure 1): (i) An initial slice segmentation is given (say via manual segmentation) and the task for the network is to predict the segmentation of all other slices in s or t dimension. We will denote these networks by S^+ and T^+, respectively. For the S^+-network for illustration, the idea is that we initially specify a segmentation for one s-location for each time instance as shown in Figure 1. The S^+-network is trained with three entities: The previous s-slice, its binary mask, and the current s-slice. Its task is to predict the segmentation of the current slice. For the next slice, the binary mask for the previous slice is the predicted mask. The T^+-network behaves

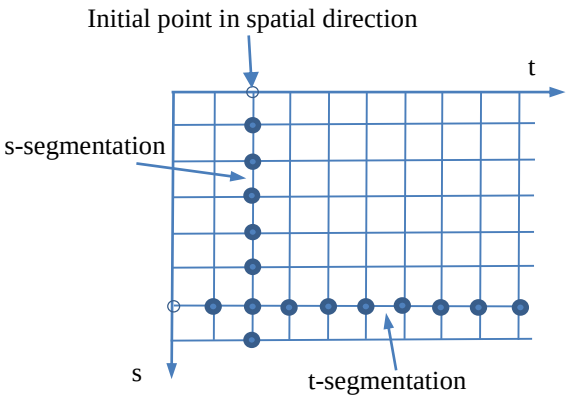

Figure 1. The idea of S- and T- networks.

analogously with s replaced by t. (ii) No initial segmentation is provided and the task for the network is to predict the segmentation of all slices in s or t dimension. We will denote these networks by S^- and T^-, respectively. For the S^--network for illustration, the idea is that it is trained with three entities: The previous, current, and next s-slice. Its task is to predict the current s-slice segmentation.

By combining these two basic strategies in different ways, we can generate several new strategies: S^+T^+, S^+T^-, S^-T^+, S^-T^-, T^+S^+, T^+S^-, T^-S^+, and T^-S^-. Note that for all these strategies, the previously trained S^+-, S^--, T^+-, and T^-- networks will be used and no retraining or additional network is needed. For illustration, consider S^+T^+. We initially use the S^+-network to perform segmentation of all s-slices for some t-location. This will require specification of the segmentation mask for one initial s-location. Subsequently, for all other s-locations, the predicted mask for the previous s-location will serve as the previous slice mask. After segmentation of all s-slices using the S^+-network in this manner for one t-location, the predicted s-slices are used as the initial mask for carrying out segmentation in the t-dimension using the T^+-network. Note that for S^+T^+, S^+T^-, T^+S^+, and T^+S^- strategies, only one initial 2D slice mask is required to be specified. For the remaining 4 strategies, no such mask is needed and they are fully automatic.

The focus of our segmentation is pharynx and larynx. Choosing a reasonable region of interest (ROI) is important for reducing the computational complexity and improving the training efficiency of the model. We select an ROI of size 192x96 automatically (Figure 2). The ROI selected guarantees that the pharynx and larynx of all subjects are included.

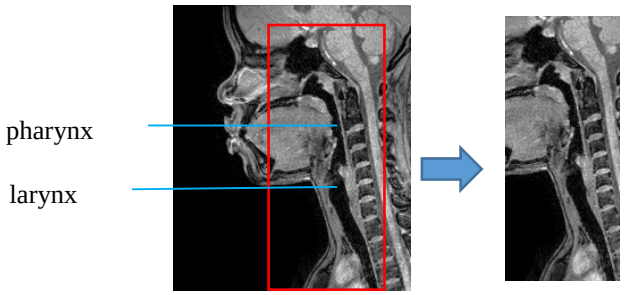

Figure 2. The red box illustrates the selection of ROI.

Network architecture

We utilize a 2D U-net [6] as the basic network architecture to implement our ST-networks. A 2D U-net cannot directly learn information between consecutive slices while a 3D convolutional network can take advantage of serial slices but increases the training complexity of the segmentation model substantially. To reduce the computational complexity and utilize slice continuity information, we used a multi-channel 2D U-net with the number of channels set to 3. The U-net is designed to contain 9 convolution layers (C1-C9), 4 pooling layers (P1-P4), 4 up-sampling layers (Up6-Up9), and 4 merge layers (M6-M9). Slices of adjacent time points or slices of adjacent positions are input into different channels of the U-net for feature learning.

Figure 3 shows the architecture of the T⁻network and the input and output of the network during testing. The three channels of the network are three slices ($F_{(t-1)}$, $F_{(t)}$, $F_{(t+1)}$) of three consecutive time points in order, and the mask $G_{(t)}$ of the middle slice is used as the ground truth during the training phase. In the test phase, the input consists of three consecutive slices

($F_{(n-1)}$, $F_{(n)}$, $F_{(n+1)}$) in order; note that this network is fully automatic. For S$^-$-networks, we only need to replace slices of continuous time by slices of consecutive spatial positions.

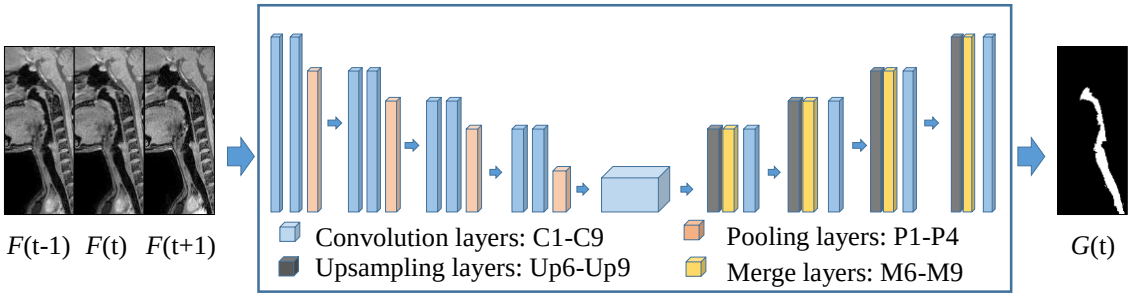

Figure 3. The input, network architecture, and ground truth of T$^-$-network.

We redesigned the input and output of 2D-Unet based on S$^-$ and T$^-$ network configurations described above to design the four basic S$^+$-, S$^-$-, T$^+$-, and T$^-$- networks. For the T$^+$-network (see Figure 4), we used two consecutive slices ($F_{(t-1)}$, $F_{(t)}$) and the mask of the previous slice $G_{(t-1)}$ as the training input during training, the order of the three channels being $F_{(t-1)}$, $G_{(t-1)}$, $F_{(t)}$. For the S$^+$-network, the input of the three channels is previous location slice $F_{(s-1)}$, previous location mask $G_{(s-1)}$, current location slice $F_{(s)}$. The mask of the current slice $G_{(s)}$ is used as the ground truth.

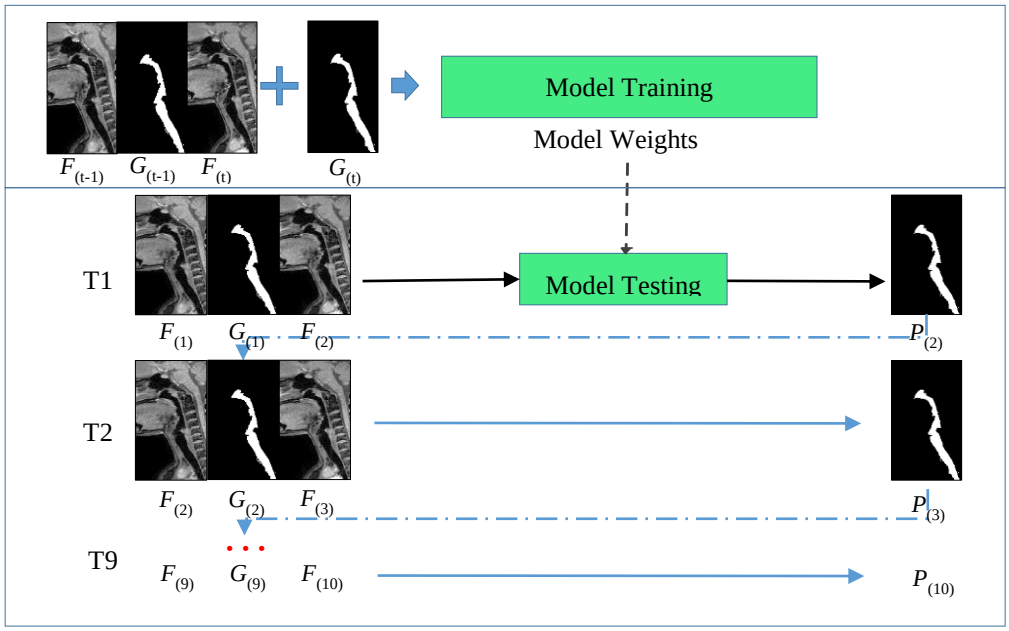

Figure 4. Training and testing configuration of the T$^+$-network as an example.

In the test phase, we manually labeled the mask $G_{(n-1)}$ of the initial time or spatial point $F_{(n-1)}$. Use the starting slice $F_{(n-1)}$, the pre-labeled mask $G_{(n-1)}$ of the starting slice, and the next slice $F_{(n)}$ as the input to get the predicted label of slice $F_{(n)}$. The predicted label $P_{(n)}$ then participates in the prediction of the next slice as a new pre-labeled mask $G_{(n)}$. Figure 4 shows the process for the T$^+$-network (training and testing).

3. EXPERIMENTS AND RESULTS

For each model, we use 4-fold cross-validation to test the segmentation accuracy from 4D image data sets from 20 patients, each patient containing 3D volumes for 10 time points, 36 slices in each 3D volume. Thus, our data set contains a total of 7,200 slices. For the S$^+$-network, we manually label the 18th s-location slice for all subjects as the initial position. For the

T^+ network, we manually label all the positions at the first time point for each subject as the initial position time point among 10 time points. Figure 5 shows exemplar segmentation results of the T^+-network. The mean and standard deviation of the Dice coefficient (DC) for all 12 tested strategies are summarized in Table 1.

Table 1. Mean and standard deviation (SD) of Dice coefficient (DC) from different models.

	S^+	T^+	S^-	T^-	S^- T^+	T^- S^+	S^+ T^+	T^+ S^+	S^- T^-	T^- S^-	S^+ T^-	T^+ S^-
Mean	0.83	0.90	0.88	0.88	0.86	0.81	0.81	0.82	0.88	0.88	0.88	0.89
SD	0.09	0.06	0.05	0.05	0.05	0.09	0.08	0.09	0.05	0.05	0.04	0.06

Among the four models, T^+ achieves the highest segmentation accuracy $0.90 \pm 0.06\%$, because the change between adjacent time points in the same position is slight, and the initial manually labeled mask guides the segmentation of the entire time cycle very well. For the S^+ model, the accuracy of the segmentation is lower than that of T^+. This is because slice-to-slice variation in the s-dimension is larger in this application and less smooth than that in the t-dimension, and the prediction error of each segmentation can mislead segmentation in subsequent positions. In the S^- and T^- models without manually labeled mask, the segmentation accuracies are similar. Also, S^- and T^- results are similar to T^+ results, and the three are statistically indistinguishable ($P > 0.05$). Among the dual-network methods, two groups of behavior can be observed: S^+T^-, S^-T^+, S^-T^-, T^+S^-, and T^-S^- with higher accuracy, statistically indistinguishable ($P > 0.05$) from the higher-accuracy single-network models, and the rest of the dual-network methods with lower performance comparable to the lower-performing single-network methods.

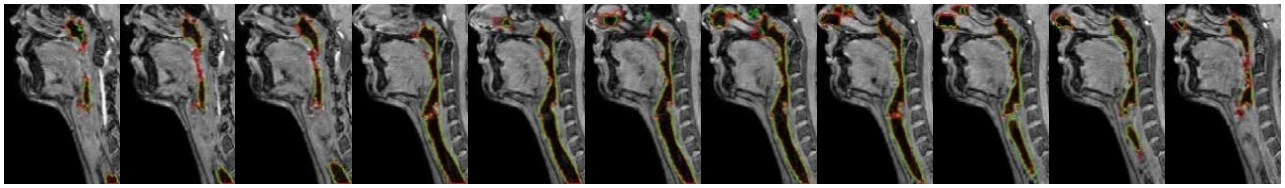

Figure 5. Segmentation results at different locations for the T^+-network. Red: Ground truth. Green: Segmentation result.

4. CONCLUSIONS

Recognizing that spatial and temporal variations in object characteristics are different in quality and magnitude, we proposed independent S- and T-networks to be designed to handle 4D dynamic images. We then showed how the two models can be combined in different ways to yield a whole set of new strategies to segment 4D imagery. We demonstrated the feasibility of the new approach on 4D MRI data sets of the upper airway of OSA patients. Our preferred method would be T^-S^- although these results are preliminary and need to be checked in other applications (such as heart, lungs, etc.) and on larger data sets. Given that these upper airway structures are sparse, a DC of ~0.88 for their segmentation is similar to a DC of >0.95 for large non-sparse objects like liver, lungs, etc., constituting excellent accuracy. Thorough validation of these networks on much larger data sets in this application is currently ongoing from data sets obtained from patients during sleep and wake conditions. We are also exploring usefulness of these ST-networks for segmentation of the lungs via free-breathing dynamic acquisitions of pediatric subjects with various types of malformations of the thorax and spine. These data sets will be utilized in the future in thoroughly testing the top performing fully automatic dual-network methods.

5. ACKNOWLEDGEMENTS

This work is partly supported by an NIH grant HL130468.

REFERENCES

[1] Maria Lorenzo-Valdés, Sanchez-Ortiz, G. I., Elkington, A.G., et al. (2004). Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. Medical Image Analysis, 8(3), 255-265.

[2] Ward, S. L. D., Amin, R., Arens, R., et al. (2014). Pediatric sleep-related breathing disorders: advances in imaging and computational modeling. IEEE pulse, 5(5), 33-39.

[3] Xu, C., Sin, S. H., Mcdonough, J. M., et al. (2006). Computational fluid dynamics modeling of the upper airway of children with obstructive sleep apnea syndrome in steady flow. Journal of Biomechanics, 39(11), 2043-2054.

[4] Wagshul, M. E., Sin, S., Lipton, M. L., et al. (2013). Novel retrospective, respiratory‐gating method enables 3D, high resolution, dynamic imaging of the upper airway during tidal breathing. Magnetic resonance in medicine, 70(6), 1580-1590.

[5] Tong, Y., Udupa, J. K., Odhner, D., et al. (2016). Minimally interactive segmentation of 4D dynamic upper airway MR images via fuzzy connectedness. Medical physics, 43(5), 2323-2333.

[6] Ronneberger O, Fischer P, Brox T. (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 2015: 234-241.