ISSN: 2414-1895 DOI: 10.6919/ICJE.202211_8(11).0012

Ultrasound Video Object Segmentation of Renal Parenchyma

Ru Wang

College of Computer Science and Technology, Qingdao University, Qingdao 266071, China

*wru0217@163.com

Abstract

We propose a scheme to apply semi-supervised video target segmentation to kidney parenchyma images. Depending on the nature of the problem, the final obtained video frames with object masks become more refined with intermediate prediction. In our framework, the past frames will be stored as a querier for the current frame, and the mask information newly entered in the memory will be segmented. The extensive use of guidance information allows us to better deal with the variable nature of renal parenchyma sequences compared to previous approaches. We validated our method on the renal parenchyma dataset and achieved state-of-the-art performance with fast runtimes.

Keywords

Deep Learning; Renal Parenchyma; Video Object Segmentation.

1. Introduction

The kidneys are one of the most important organs in the human body, which can ensure the stability of the internal environment of the human body and enable the metabolism to proceed normally. Studies have shown that the area of renal parenchyma in renal ultrasound images may be closely related to acute and chronic renal diseases, and the thickness of renal parenchyma can also be used as a marker of chronic renal failure. Therefore, computer automatic tracking and segmentation of renal parenchyma can provide a premise support for further measurement of renal related indicators, which has important medical research value. M. Freimen et al. [1]defined kidney segmentation as a minimum-cut solution problem for maximum posterior probability estimation in a Markov follower field. By solving the model, unknown model variables and segmentation results were obtained; D. Turco et al. [2] Using a fully automatic segmentation technique, the total segmentation volume was calculated based on the non-differential CT (Computer Tomography) information of patients with autosomal dominant polycystic kidney disease. After the rapid application of the neural network, a large number of experiments have also proved its advantages. The Unet model proposed by Ronneberger et al. [3] is also improved based on FCN. It can be integrated into lower-level details, and a large number of convolution blocks are used in the upsampling part to extract features, making full use of context information, so the segmentation accuracy is also relatively high, and it was originally proposed to segment medical images. Chen et al. [4] proposed the Deeplab family of models with four versions. In Deeplabv1, atrous convolution was introduced to solve the problem of reduced resolution after multiple pooling, and then a fully connected conditional random field (CRF) was used to improve the model to better learn detailed information. However, the renal parenchyma in the renal ultrasound video is complex and diverse in shape, there are many artifacts and noise, and there is continuity between the video sequences. If the images are analyzed separately, it will cause greater errors. In addition, in many ultrasound images, the edge of the segment to be segmented is fused with the background, which leads to the fact that the existing methods cannot segment the renal parenchyma well. How to accurately segment the renal parenchyma is still a challenging task. In the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202211_8(11).0012

experiment, we created a renal parenchyma dataset and proposed a deep learning method for renal parenchyma segmentation in renal ultrasound videos in view of the difficult segmentation of this dataset.

2. Related Work

2.1 Video Object Segmentation

Video object segmentation methods include propagation-based methods, detection-based methods, hybrid methods, and online/offline learning. Detection-based methods [5,6] learn an object mask propagator, a deep network that refines unaligned masks to target objects. To make the network object specific, online training data is generated from the first frame by deforming object masks or synthesizing images for fine-tuning. Detection-based methods [7,8] work by learning an object detector using the object appearance on the first frame. Hybrid methods aim to take advantage of both detection and propagation methods. In [9], a sequence-to-sequence network that learns long-term information in videos is proposed. Online learning-based methods: To distinguish the target object from the background and distractors, online learning-based methods fine-tune the segmentation network on the first frame. OSVOS [10] fine-tunes the pretrained segmentation network on the first frame of the test video. Offline methods utilize the initial frame and pass target information to subsequent frames through propagation or matching. MaskTrack [11] concatenates the prediction mask of the previous frame with the image of the current frame to provide spatial guidance. The framework of STM [12] maintains intermediate outputs in external memory, rather than fixing which frame to use as a guide, and adaptively selects necessary information at runtime without learning a deep network model that fine-tunes the initial object mask in the first frame. Although online learning improves accuracy, it is computationally expensive and cannot be used in practical applications, and offline learning methods try to bypass online learning while maintaining accuracy.

2.2 FCNs for Segmentation

Fully Convolutional Neural Network (FCN) is a framework for semantic segmentation, which is proposed on the basis of Convolutional Neural Network (CNN). The classic convolutional neural network has the AlexNet model, which is mainly used for image classification and regression, and the final result will output a value. The difference between FCN and CNN is that the fully connected layer in CNN is replaced by a convolutional layer, and the final output result is not only a numerical value, it classifies each pixel, and the output result is a marked image. This change successfully solves the image segmentation problem at the semantic level. The original fully convolutional network was proposed by LongJ [13], and related models have been widely used in semantic segmentation of various types of images. Currently commonly used medical image segmentation networks usually use FCN-like structures [14,15]. Vorontsov et al. used two types of FCNs for liver and liver lesion segmentation [15]. Although these methods have achieved reasonable segmentation results in medical image segmentation, there is still a problem of blurred boundaries in preserving the boundaries.

3. Proposed Method

In this paper, a semi-supervised video object segmentation method for renal ultrasound video is proposed. First, the convolutional neural network is used to extract features, and then the position similarity and attention mechanism modules are used for further processing to estimate the semantics of the object to be segmented. Appearance models are combined with semantic priors, memory pools store past data, and finally the architecture is trained to determine foreground pixels for a particular image, initializing the convolutional neural network with weights at test time and fine-tuning and iterating. This experiment can effectively extract useful information in the video, and analyze the continuous video frames, which greatly reduces the time, manpower and material resources spent on labeling each video frame. The segmentation is more refined, and the accuracy is also improved.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202211_8(11).0012

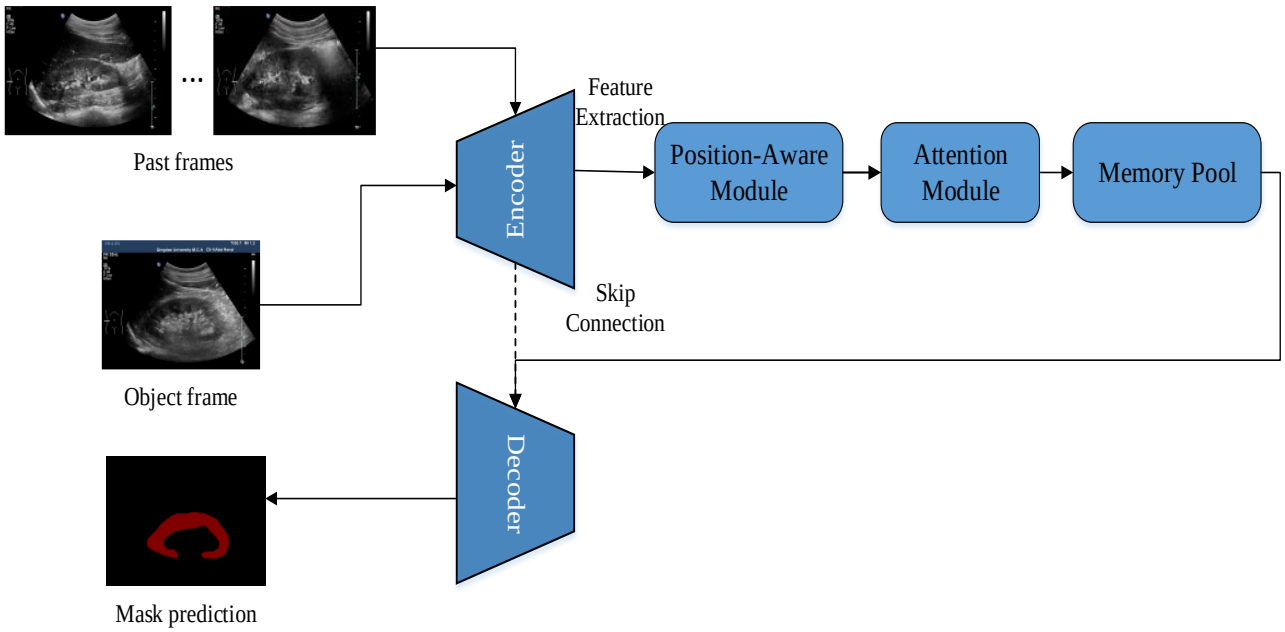


Fig. 1 Model framework

Automatically segmenting the renal parenchyma provides great convenience for subsequent measurement of the area, gray value, and long axis of the kidney, and can provide doctors with quantitative data to assist doctors in diagnosing acute and chronic kidney diseases. The method in this paper is shown in Figure 1. The input sequence enters the encoder, usually the mask of the first image is given, and the video frames are processed sequentially from the second frame. The encoder is robust against appearance changes by performing feature extraction on the input image, and in the position correlation module, computing the similarity between the key features of the query and the memory frame. By adding an attention module, the model learns whether the mask information of each feature belongs to the foreground or the background. In the read operation of a memory cell, the corresponding weights are first calculated by measuring the similarity between the input and all pixels of the memory map. Similarity matching is performed by comparing each corresponding location in the memory map with each spatial location in the input.

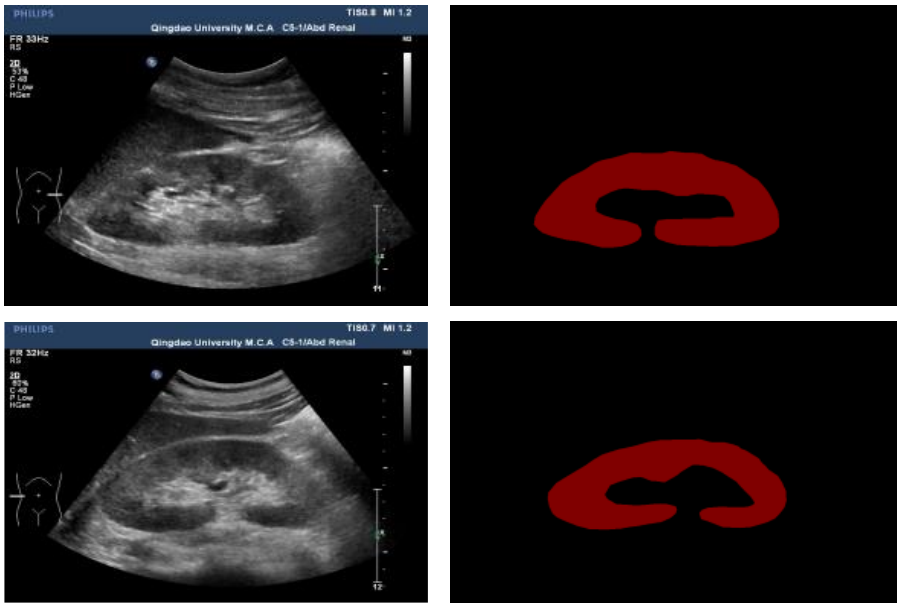
In the decoder stage, we train a fully convolutional neural network (FCN) for the binary classification task of separating foreground objects from background. The decoder is connected with the encoder through the FPN structure, which can fuse the features of each level, and has strong semantic information and spatial information at the same time, which is easy to deploy. The skip connection is used between the decoder and the encoder, and the feature maps of the same level are connected horizontally, which can be fused into the detailed information of the lower level, making full use of the context information and improving the accuracy of segmentation. The entire joint segmentation network fuses an encoder, a position-based similarity encoder, and a global feature attention mechanism and decoder to produce mask predictions. In this work, the backbone network used by the encoder is ResNet101. The experimental training steps are divided into two steps: first, we train a large number of video sequence frame objects offline to build a model that can distinguish the general concept of foreground objects; second, at test time, we fine-tune the network to perform a small number of iterations on the specific instance we want to segment and make mask predictions for the next frame in a much faster way.

4. Experiment

In this section, we conduct experiments on the renal parenchyma ultrasound video dataset using the method proposed in this paper, and then present the segmentation results on the test set to demonstrate the effectiveness of our proposed method. The experimental environment of the model in this paper

ISSN: 2414-1895 DOI: 10.6919/ICJE.202211_8(11).0012

is as follows: Windows 10 (64-bit) operating system, memory 32.00GB, and GPU as NVIDIA GeForce GTX2080Ti. The software environment used is Spyder under Anaconda3, the deep learning framework Pytorch is used for experiments, and the program running Python environment is Python3.6. The visualization of experimental results is implemented by Matplotlib. Set the experimental parameters used by each network model: the number of epoch training rounds is 50, the backbone network is Resnet101, the batchsize of each iteration input sample is 1, Adam is used as the optimizer, the initial learning rate is 0.01, and the data created in this paper is used. set for segmentation experiments.



(a) (b)

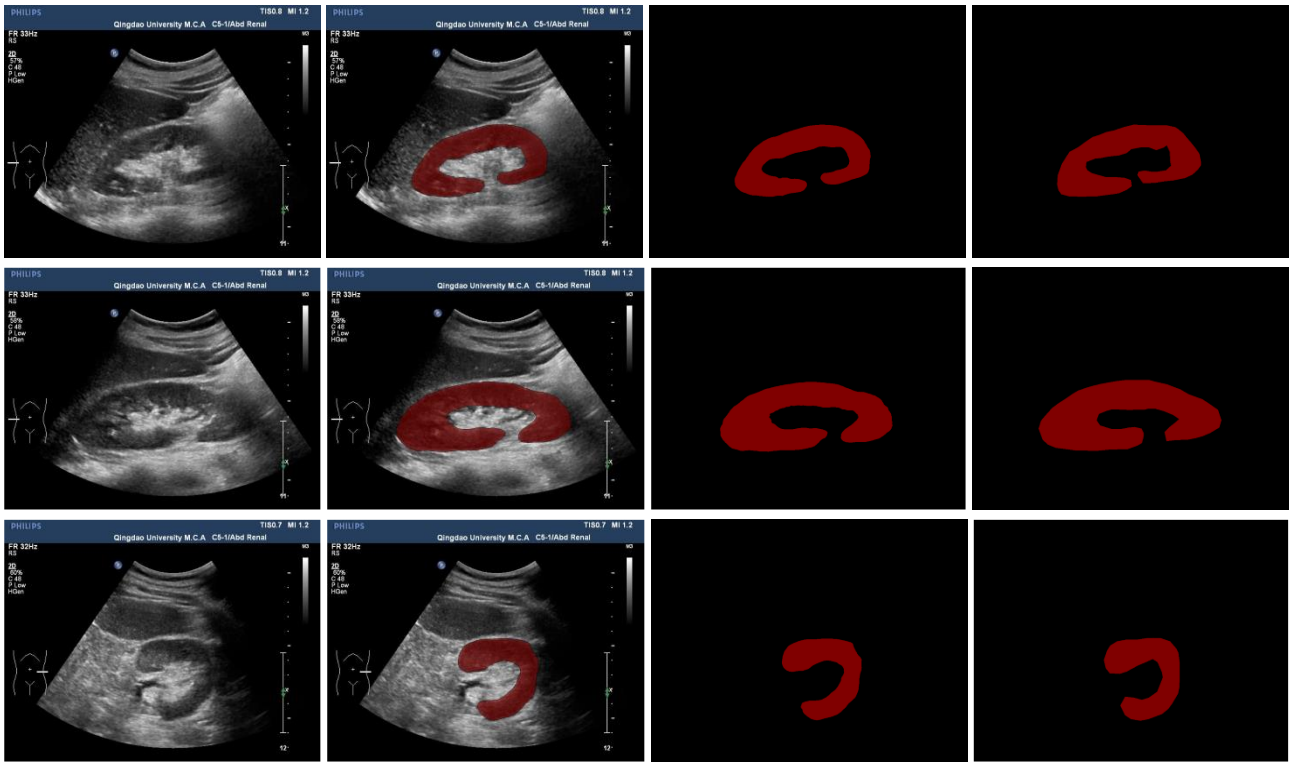
Fig. 2 (a) represents the original image, (b) represents the label, and the position and size of the renal parenchyma

The data set used in this experiment comes from the Department of Imaging, Affiliated Hospital of Qingdao University. 750 kidney ultrasound images were extracted in the experiment, and a segmentation data set of renal parenchyma was created. The data set includes 600 training samples and 150 test samples. All kidney scan images of the dataset are sagittal planes, collected by a Philips ultrasound scanner and abdominal transducer, as shown in Fig. 2(a). All images have a resolution of 1024x768, the number of dots per inch of the image (DPI, DotsPerInch) is 96, the pixel value range is [0,255], and all personal information has been removed from the images. The labeling of all images in the dataset was done manually by ultrasound radiologists. The dataset creation tool in the experiment was Labelme image labeling software written in python language. The labeling of each dataset was made under the guidance of professional doctors. This article uses single-frame sequences of the dataset as well as video images. The original image of the dataset and the label map of each segment are shown in Figure 2. Figure 2(a) is an ultrasound image of the kidney, and Figure 2(b) is the outline of the renal parenchyma marked by a doctor with a special tool. For the problem of a small number of training sets, data enhancement methods such as horizontally flipping the data set, increasing contrast, and changing brightness are adopted. For artifacts and noise in the images of the data set, methods such as increasing contrast and increasing brightness are adopted.

The effect diagram of renal parenchyma segmentation is shown in Figure 3, and the picture is the prediction result of the intermediate frame of the renal ultrasound video. It can be seen from the figure that the method in this paper has a better effect on the segmentation of renal parenchyma video, and can more accurately capture the position of the renal parenchyma. The algorithm can detect the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202211_8(11).0012

blurred boundary that may be lost, and can refine the image edge segmentation to a certain extent, so as to retain more detailed information more effectively.



(a) (b) (c) (d)

Fig. 3 (a) represents the original image, (b) represents prediction mask for intermediate frames, (c) represents the segmentation results of the algorithm in this paper, (d) represents the label, and the position and size of the renal parenchyma

In the test part, the test dataset, which contains consecutive frame sequences of multiple cases, is compared using the algorithm of this paper and other methods, as shown in Table 1. Table 1 compares the five models of U-Net, OSVOS, STM, SiamMask [16] and our algorithm. The test data were compared with PPV, TPR, TNR, ACC, F1-scores and JS metrics and are shown in Table 1. It can be seen from the table that the algorithm in this paper has reached the highest value on most of the indicators.

Table 1. Computation time comparison of the five models.

Models	PPV	TPR	TNR	ACC	F1	JS
U-Net	85.37	77.21	98.14	95.61	81.26	95.45
OSVOS	86.14	79.52	97.58	95.50	82.43	96.36
STM	86.74	79.35	97.72	95.62	82.74	96.82
SiamMask	87.45	80.05	97.94	96.38	83.62	96.93
Ours	87.63	80.28	98.06	96.59	83.87	97.02

5. Conclusion

In this paper, we propose a scheme for applying semi-supervised video object segmentation to images of renal parenchyma. Depending on the nature of the problem, the resulting video frames with object

ISSN: 2414-1895 DOI: 10.6919/ICJE.202211_8(11).0012

masks become more refined with intermediate predictions. In our framework, a location-based attention module and an attention mechanism for feature extraction are added, so that the model can use the relevant features of the previous frame to be added to the target segmentation of the current frame, and improve the accuracy of the model. Good results were obtained in renal parenchyma segmentation. We validate our method on the renal parenchyma dataset and achieve state-of-the-art performance with fast runtime. Furthermore, the method converges quickly and has a small parameter size, and future work includes applying our framework to other video object segmentation datasets.

References

[1] Freiman M, Kronman A, Esses S J, et al. Non-parametric iterative model constraint graph min-cut for automatic kidney segmentation [J]. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2010, 13(3):73-80.

[2] Turco D, Valinoti M, Martin E M, et al. Fully automated segmentation of polycystic kidneys from noncontrast computed tomography [J]. Academic Radiology, 2018, 25(7):850-855.

[3] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation [J]. Springer Cham, 2015, 9351(11):234-241.

[4] Chen L-C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]. Proceedings of the European conference on computer vision (EC-CV), 2018:801-818.

[5] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. Maskrnn: Instance level video object segmentation. In Advances in Neural Information Processing Systems, 2017. 1, 2, 6.

[6] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. arXiv preprint arXiv:1703.09554, 2017. 2.

[7] Linchao Bao, Baoyuan Wu, and Wei Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higherorder spatio-temporal mrf. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2, 6.

[8] Y uhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc V an Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2.

[9] Ning Xu, Linjie Y ang, Dingcheng Y ue, Jianchao Y ang, Brian Price, Jimei Y ang, Scott Cohen, Y uchen Fan, Y uchen Liang, and Thomas Huang. Y outube-vos: Sequence-to-sequence video object segmentation. In European Conference on Computer Vision (ECCV), 2018. 2, 4, 5, 6, 7.

[10] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc V an Gool. Oneshot video object segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017. 1, 2, 6.

[11] Sangho Lee, Jinyoung Sung, Y oungjae Y u, and Gunhee Kim. A memory network approach for story-based temporal summarization of 360 videos. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1410–1419, 2018. 2.

[12] Oh, Seoung Wug, et al. "Video object segmentation using space-time memory networks." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[13] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017(4):640-651.

[14] Christ P F, Elshaer M E A, Ettlinger F, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2016: 415-423.

[15] Vorontsov E, Tang A, Pal C, et al. Liver lesion segmentation informed by joint liver segmentation[C] //2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018: 1332-1335.

[16] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1328– 1338, 2019.