DeepDisc: Optic Disc Segmentation Based on Atrous Convolution and Spatial Pyramid Pooling

Zaiwang Gu[1,2(✉)], Peng Liu[1,3], Kang Zhou[4], Yuming Jiang[1,3], Haoyu Mao[1], Jun Cheng[1], and Jiang Liu[1]

[1] Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Beijing, China
guzaiwang@nimte.ac.cn
[2] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[3] University of Electronic Science and Technology of China, Chengdu, China
[4] School of Information Science and Technology, ShanghaiTech University, Shanghai, China

Abstract. The optic disc (OD) segmentation is an important step for fundus image base disease diagnosis. In this paper, we propose a novel and effective method called *DeepDisc* to segment the OD. It mainly contains two components: atrous convolution and spatial pyramid pooling. The atrous convolution adjusts filter's field-of-view and controls the resolution of features. In addition, the spatial pyramid pooling module probes convolutional features at multiple scales and encodes global context information. Both of them are used to further boost OD segmentation performance. Finally, we demonstrate that our DeepDisc system achieves state-of-the-art disc segmentation performance on the ORIGA and Messidor datasets without any post-processing strategies, such as dense conditional random field.

Keywords: Disc segmentation · Atrous convolution
Spatial pyramid pooling

1 Introduction

The optic disc (OD) contains lots of information and is an important anatomical structure in the retina. Locating and segmenting OD is an essential step in many retinal image analysis for disease detection and monitoring. For example, in age-related macular degeneration detection, localization of the macula is often conducted by finding the OD first. In retinal image analysis for coronary heart disease detection, OD is segmented for retinal vessel caliber measurement [10]. In glaucoma diagnosis, OD segmentation is often conducted before the cup to

© Springer Nature Switzerland AG 2018
D. Stoyanov et al. (Eds.): COMPAY 2018/OMIA 2018, LNCS 11039, pp. 253–260, 2018.
https://doi.org/10.1007/978-3-030-00949-6_30

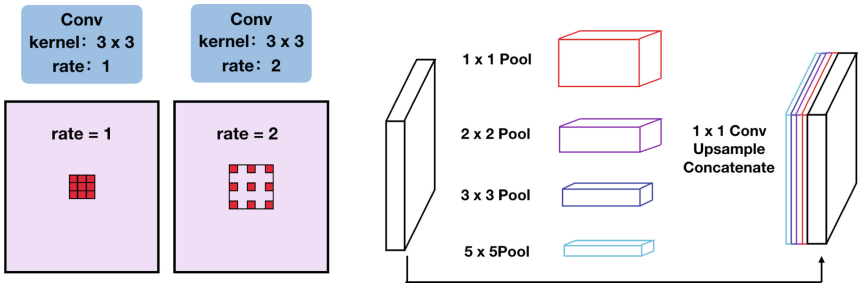
disc ratio (CDR) is computed [2,5]. Therefore, a reliable OD segmentation is necessary in automatic retinal image analysis for different diseases detection.

Many methods [2,3,7,9,15] have been proposed to segment the OD. Cheng *et al.* proposed a method based on peripapillary atrophy elimination [1]. Then, they improved the OD segmentation performance by adopting superpixel classification method [2]. Li *et al.* [9] built a shape-appearance model to segment the OD, which could learn a sequence for supervised decent directions between the coordinates of OD boundary and their surrounding visual appearances for OD segmentation.

With the development of convolution neural network (CNN) in image and video processing [8], automatic feature learning algorithms using deep learning have emerged as feasible approaches and are applied to handle the retinal image analysis. Recently, some OD segmentation algorithms [5,14] based on the fully convolutional network (FCN) [11] have been proposed. In FCN-like structure, the deep CNNs are originally designed for image classification tasks. The consecutive pooling and strided convolutional operations reduce the feature resolution and extract increasingly abstract feature representations. Meanwhile, the segmentation is a dense prediction task, which is a pixel-wise classification. The detailed spatial information is often neglected in the pooling and strided convolutional operations, however, such information is important for segmentation.

To overcome this limitation, Fu *et al.* [5] proposed M-Net, which is modified from U-Net [13]. The architecture contains multi-scale inputs and deep supervision. The methods attempted to solve the problem of losing some detailed information with pooling and strided convolution operations. The multi-scale inputs introduced the resized images into each middle training stage to ensure that the architecture could learn more from the full-size image. The deep supervision was adopted to optimize the deep network more easily. However, the multi-scale inputs and deep supervision cannot prevent losing some spatial information inevitably. More generally, the FCN-like structures (such as U-Net) can be considered as a Encoder-Decoder architecture. The Encoder aims to reduce the spatial dimension of feature maps gradually and capture more high-level semantic features. The Decoder aims to recover the object details and spatial dimension. FCN employs deconvolution operation to learn the upsampling of low resolution features, while U-Net mainly adds skip connections from the encoder features to the corresponding decoder features. Preserving more spatial information in the convolution is critical to improve the performance of the OD segmentation. Intuitively, maintaining high-resolution feature maps at the middle stage can boost segmentation performance. However, to accelerate training and ease the difficulty of optimization, the size of feature map should be small. Therefore, there is a trade-off between accelerating the training and maintaining the high resolution.

In this paper, different from the FCN-like structures, we propose a novel architecture for OD segmentation. It combines atrous convolution [16] and spatial pyramid pooling. In particular, the atrous convolution allows us to efficiently enlarge the field of view of filters to incorporate multi-scale context. It learns



(a) The standard convolution can be seen as atrous convolution with rate = 1

(b) Spatial pyramid pooling employs four-level pooling kernels to extract both local and global context information.

Fig. 1. The illustrations of atrous convolution and spatial pyramid pooling

high-level semantic features in high resolution and preserve more spatial details. The spatial pyramid pooling strategy is adopted to ensure the pooling operation at multiple kernel sizes and effective fields of view. Our proposed OD segmentation architecture is validated on ORIGA and Messidor datasets. It outperforms state-of-the-art methods and achieves an overlapping error of 0.069 in the ORIGA dataset and 0.064 in the Messidor dataset.

The rest of the paper is organized as follows. In Sect. 2, we introduce our proposed method in detail. Section 3 contains our adopted datasets including ORIGA and Messidor, as well as the experimental results. In the last section, the conclusion is presented.

2 Method

The OD segmentation can be considered as a pixel classification problem. It assigns each pixel a label, indicating whether this pixel belongs to the OD or not. To achieve high performance in this dense prediction task, it should not only maintain high-resolution features containing much spatial information, but also extract high-level semantic features. Therefore, we propose to use atrous convolution operation and spatial pyramid pooling strategy for OD segmentation.

2.1 Atrous Convolution

The atrous convolution is originally developed for the efficient computation of the wavelet transform. Mathematically, the atrous convolution under two-dimensional signals is expressed as the following formula:

$$\boldsymbol{y}[\boldsymbol{i}] = \sum_{\boldsymbol{k}} \boldsymbol{x}[\boldsymbol{i} + r\boldsymbol{k}]\boldsymbol{w}[\boldsymbol{k}], \tag{1}$$

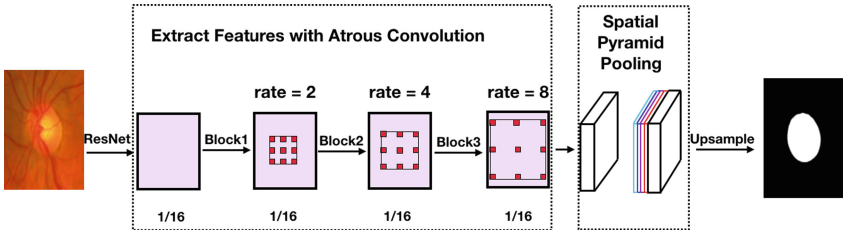


Fig. 2. The architecture of the proposed method. Given an input image, we use pretrained-ResNet to extract features. Then, atrous convolution and spatial pyramid pooling are applied to extract high-level semantic features under high resolution. Finally, upsampling operation is adopted to get the final per-pixel prediction map.

where the convolution of the input feature map $\boldsymbol{x}$ and a filter $\boldsymbol{w}$ yields the output $\boldsymbol{y}$, where the atrous rate r corresponds to the stride with which we sample the input signal. It is equivalent to convolute the input $\boldsymbol{x}$ with upsampled filters produced by inserting $r-1$ zeros between two consecutive filter values along each spatial dimension (hence the name atrous convolution where the French word atrous means holes in English). Standard convolution is a special case for rate $r=1$, and atrous convolution allows us to adaptively modify filter's field-of-view by changing the rate value. See Fig. 1(a) for illustration.

2.2 Spatial Pyramid Pooling

A challenge in segmentation is the large variation of OD sizes. In this paper, we adopt spatial pyramid pooling to address the problem, which mainly relies on multiple effective fields of view to detect objects at different sizes.

In the deep neural network, the size of receptive field can roughly indicates how much we can use the context information. The adopted pyramid pooling encodes global context information with four different-size receptive fields, $1\times1, 2\times2, 3\times3$ and 5×5 respectively. The four-level outputs contain the feature maps with varied sizes. To reduce the dimension of weights and computational cost, we use a 1×1 convolution after each level of pooling. It reduces the dimension of feature maps to the 1/N of original dimension, where N represents number of channels in original feature maps. Then we upsample the low-dimension feature map to get the same size features as the origin feature map via bilinear interpolation. Finally, we concatenate the original pyramid pooling features with upsampled feature maps.

2.3 Network Architecture

With the atrous convolution and spatial pyramid pooling, we propose our OD segmentation network as illustrated in Fig. 2. Given an input image, we adopted the ImageNet-pretrained ResNet [6] as our main body. Different from the ResNet

under classification task, we use atrous convolution to extract the dense high-level semantic features. Atrous convolution allows us to explicitly control the denseness of feature responses in neural convolution networks. Here, all subsequent convolutional layers are replaced with atrous convolution with rate $r = 2$. This allows us to extract denser feature responses without requiring learning extra parameters.

The pretrained-ResNet is used to extract the high-level feature maps and the image is 1/16 of the original input image. Then we continue with atrous convolution layers with rate r = 2, 4, 8. After that, we use spatial pyramid pooling module to gather the context information. The four-level pooling kernels are adopted to cover most multi-scale size objects. Finally, we bilinearly upsample and deconvolve the feature maps to the desired spatial dimension.

To summarize, we use the ImageNet-pretrained ResNet as the feature extractor, avoiding the training difficulty with lacking of sufficient training dataset. With the atrous convolution and spatial pyramid pooling, the proposed method ensures the spatial resolution. At the same time, it can also detect discs with different sizes.

Compared to the background in a fundus image, OD just occupied a smaller area. Therefore, common cross-entropy loss function could not achieve high performance on facing the problem of imbalanced data distribution. In this paper, we use dice-coefficient loss function [12].

3 Experiments and Results

3.1 Datasets and Evaluation Protocols

We validated the effectiveness of the proposed method using two datasets, the ORIGA [17] and Messidor datasets. The ORIGA dataset contains 325/325 images for training/testing respectively [4], and the OD boundaries have been manually demarcated. The Messidor dataset is a public dataset provided by the Messidor program partners. It consists of 1200 images with three different sizes: $1440 \times 960, 2240 \times 1488, 2340 \times 1536$. The Messidor dataset is originally collected for Diabetic Retinopathy (DR) grading. Later, disc boundary for each image has also been provided[1].

To evaluate the performance, we adopt the overlapping error [2], E given as:

$$E = 1 - \frac{Area(S \cap G)}{Area(S \cup G)}, \tag{2}$$

where S and G denote the segmented and the manual ground truth OD respectively.

[1] http://www.uhu.es/retinopathy/.

3.2 Implementation Details

The OD only occupies a small area in full-size image, which causes difficulty in the training. Therefore, we crop a 800×800 region containing the OD of ORIGA dataset and a 448×448 region of Messidor dataset. Data augmentation is adopted on the training set by flipping images, which simulates the relationship between left-right eyes and also simulates the image rotation caused in capturing the fundus images. Our architecture is implemented with PyTorch. During the training process of whole architecture, we trained the parameters of atrous convolution and spatial pyramid pooling modules with 0.001 learning rate. In addition, we fine-tuned the parameters of ResNet with one tenth of origin learning rate. Additionally, the ℓ_2 norm regularization with factor 0.00004 is applied in the final loss function. The whole code will be released at github after acceptance of the paper.

Table 1. Performance comparison of the different methods on ORIGA dataset

Method	$\boldsymbol{E}_{disc}$
R-Bend [7]	0.129
ASM [15]	0.148
Superpixel [2]	0.102
QDSVM [3]	0.110
U-Net [13]	0.089
M-Net [5]	0.083
Our	0.069

Table 2. Performance comparison of the different methods on Messidor dataset.

Method	$\boldsymbol{E}_{disc}$
Li's [9]	0.087
Superpixel [2]	0.125
U-Net [13]	0.069
DeepDisc	0.064

3.3 Experiment Results

Experiment Results on the ORIGA Dataset. Limited by the memory of our computational devices, we choose the ImageNet-pretrained ResNet-34 as main body of architecture. As can be seen in Table 1, we show the performances of most OD segmentation algorithms. Compared with some state-of-the-art OD segmentation methods, our DeepDisc achieves an overlapping error of 0.069 in the ORIGA dataset. It outperforms the other algorithms based on deep learning or traditional image processing method. From the comparison between M-Net and our DeepDisc, we observe that there is a clear improvement of overlapping error by 16.9% from 0.083 to 0.069.

Experiment Results on the Messidor Dataset. To further validate the performance of OD segmentation, the Messidor dataset is also tested. The image of Messidor only has rough boundary of the OD, however it is enough for our validation. As three different sizes in the Messidor, we first resize all images to a fix size 1440×960, then crop a 448×448 size image. For a fair comparison, we follow the same setting as in [9]: 1000 images for training and 200 images for testing. From the Table 2, our proposed method achieves an overlapping error of 0.064 and it outperforms the other algorithms. Different from the FCN-like structures, our DeepDisc system based on atrous convolution and spatial pyramid pooling can save more detailed information and improve performance on OD segmentation, as illustrated in Fig. 3.

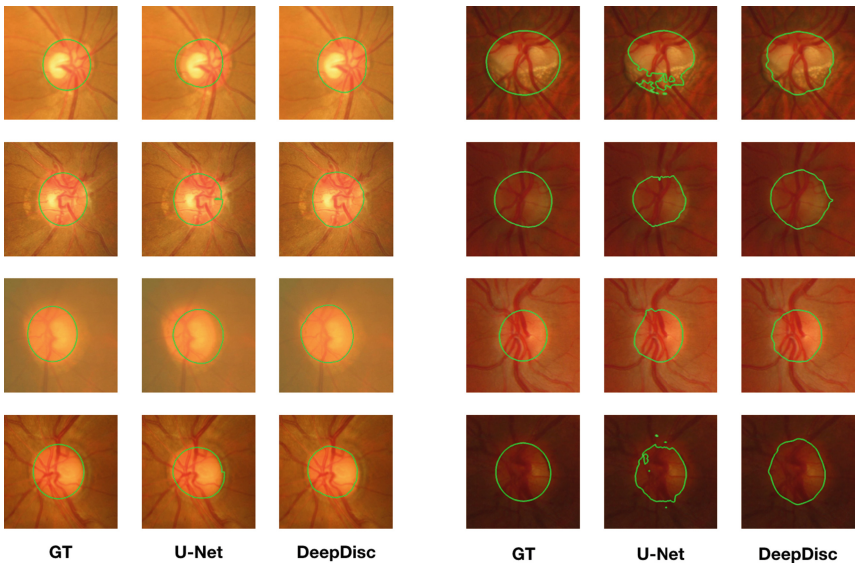


Fig. 3. The visual examples of optic disc segmentation. The left is examples from ORIGA dataset and the right is from Messidor dataset.

4 Conclusion

We propose an effective architecture to segment the OD, which employs atrous convolution to extract the dense image-level feature maps with high resolution and the spatial pyramid pooling module to collect contextual information with multiple effective field-of-views. The proposed method achieves an overlapping error of 0.069 in the ORIGA dataset and 0.064 in the Messidor dataset, better than other methods.

References

1. Cheng, J., et al.: Automatic optic disc segmentation with peripapillary atrophy elimination. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6224–6227. IEEE (2011)
2. Cheng, J., et al.: Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE Trans. Med. Imaging **32**(6), 1019–1032 (2013)
3. Cheng, J., Tao, D., Wong, D.W.K., Liu, J.: Quadratic divergence regularized SVM for optic disc segmentation. Biomed. Opt. Express **8**(5), 2687–2696 (2017)
4. Cheng, J., et al.: Similarity regularized sparse group lasso for cup to disc ratio computation. Biomed. Opt. Express **8**(8), 3763–3777 (2017)
5. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imaging (2018)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Joshi, G.D., Sivaswamy, J., Krishnadas, S.: Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. IEEE Trans. Med. Imaging **30**(6), 1192–1205 (2011)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Li, A., et al.: Learning supervised descent directions for optic disc segmentation. Neurocomputing **275**, 350–357 (2018)
10. Li, H., Hsu, W., Lee, M.L., Wong, T.Y.: Automatic grading of retinal vessel caliber. IEEE Trans. Biomed. Eng. **52**(7), 1352–1355 (2005)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
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. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. arXiv preprint arXiv:1704.00979 (2017)
15. Yin, F., et al.: Model-based optic nerve head segmentation on retinal fundus images. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2626–2629. IEEE (2011)
16. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
17. Zhang, Z., et al.: ORIGA-Light: an online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3065–3068. IEEE (2010)