Optic Disc Segmentation via Deep Object Detection Networks

Xu Sun, Yanwu Xu, Wei Zhao, Tianyuan You, Jiang Liu

***Abstract*— Accurate optic disc (OD) segmentation is a fundamental step in computer aided diagnosis for various types of eye diseases. In this paper, we propose a new pipeline to segment OD from retinal fundus images based on deep object detection networks. The fundus image segmentation problem is defined as a more straightforward object detection task. This then allows us to find the OD boundary simply by transforming the predicted bounding box into a vertical ellipse. Using Faster R-CNN as the object detector, our method achieves state-of-the-art OD segmentation results on ORIGA dataset, outperforming existing methods in this field.**

I. INTRODUCTION

The optic disc (OD) is the exit point where ganglion cell axons leaves the eye. Reliable OD segmentation is a necessary step in the diagnosis of various retinal diseases such as diabetic retinopathy and glaucoma. With this in mind, a natural question arises as if we could lend impressive object detection results of deep learning from computer vision community to investigate the OD segmentation problem.

A plenty of research have been conducted to segment the OD and/or OC, and they are mainly based on segmentation techniques like thresholding, edge-based, and region-based methods. In order to achieve satisfactory result, ellipse fitting algorithm is often performed in these methods for generating smooth boundaries. Given the truth that localizing a bounding box requires exactly the same parameters as an ellipse (regardless of its inclination), *i.e.*, its width, height, and central point (horizontal and vertical coordinates), in this paper, we propose a simple yet effective method to jointly segment optic disc and cup from a retinal fundus image using deep object detection architectures.

The main steps of our method are illustrated in Fig. Firstly, a color fundus image is fed into a deep object detection networks to produce bounding boxes for both OD and OC. The parameters of these boxes are then used to generate the required ellipse-like boundaries and finally output the segmentation masks. Using faster R-CNN [1] in the first step , our detection based method achieves state-of-the-art segmentation results on ORIGA dataset [2], with the average non-overlapping ratio of 9.77% and 22.9% for OD and OC segmentation, respectively. Furthermore, the CDR is calculated with bounding boxes for glaucoma screening. Our proposed method obtains satisfactory glaucoma screening performance with the area under the curve (AUC) of 0.825 on the same dataset.

Xu Sun and Yanwu Xu are with Guangzhou Shiyuan Electronics Co., Ltd. (CVTE), Guangzhou, China {sunxu, xuyanwu}@cvte.com

II. RELATED WORK

A. Deep Object Detection Networks

With the resurgence of deep learning, computer vision community has significantly improved object detection results over a short period of time. Modern object detection systems can mainly be divided into two groups: one-stage detectors and two-stage detectors. OverFeat [3] was one of the pioneered modern one-stage object detector based on deep networks. More recent works like SSD [4],YOLO [5] and RetinaNet [6], have demonstrated their promising results. Generally, these approaches are applied over regularly sampled candidate object locations across an image. In contrast, two-stage detectors are based on a proposal-driven mechanism, where a classifier is applied to a sparse set of candidate object locations. Following the R-CNN work [7], recent progresses on two-stage detectors have focused on processing all regions with only one shared feature map, and on eliminating explicit region proposal methods by directly predicting the bounding boxes. Various extensions to this framework have been presented, *e.g.*, Faster R-CNN [1], R-FCN [8], Feature Pyramid Networks [9] and Mask-R-CNN [10].

B. Optic Disc Segmentation

A plenty of research have been conducted to segment the OD, and they are mainly based on segmentation techniques like thresholding, edge-based, and region-based methods.

Earlier, the template based methods are proposed firstly to obtain the OD boundary. For example, Lowell et al. employed the active contour model [19] to detect the contour based on image gradient. In [20], [21], the Circular-based Transformation techniques are employed to obtain the OD boundary. In [9], the local texture features around each point of interest in multidimensional feature space are utilized to provide robustness against variations in OD region. Recently, the pixel classification based method is proposed to transfer the boundary detection problem into a pixel classification task, which obtains a satisfactory performance. Cheng et al. [10] utilizes superpixel classifier to segment the OD and OC, which exploits using various hand-crafted visual features at superpixel level to enhance the detection accuracy. In [22], the disparity values extracted from stereo image pairs are introduced to distinguish the OD and background. However, reliance on hand-crafted features make these methods susceptible to low quality images and pathological regions.

III. METHOD

In this paper, we propose a new pipeline for OD segmentation from retinal fundus images. The pipeline of our basic

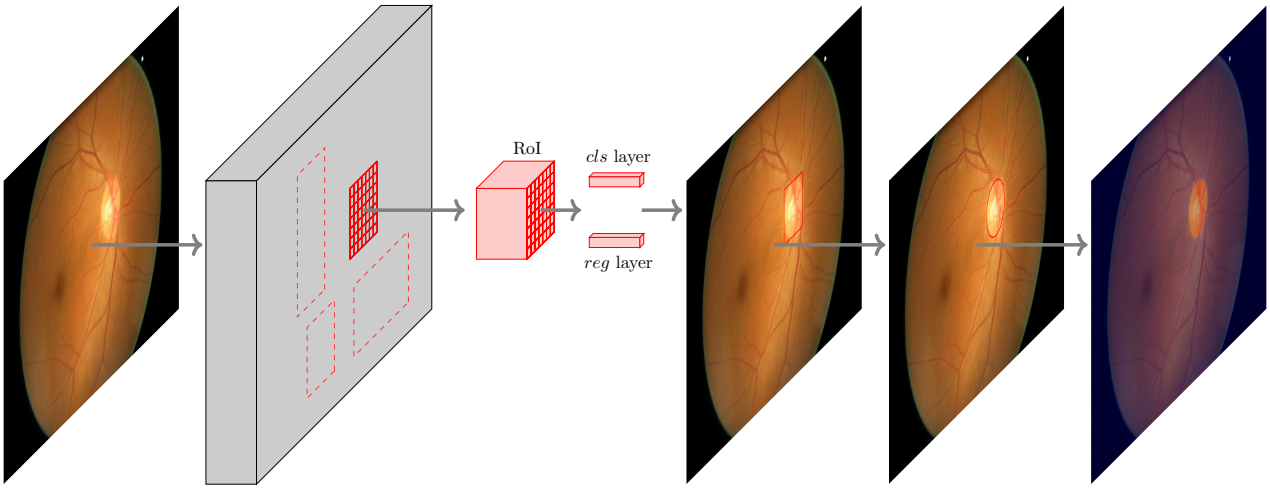

Fig. 1. Illustration of the proposed pipeline for OD segmentation from color fundus images, where red region denotes the detected OD.

algorithm is shown in Fig. 1.

- We first feed the color fundus image into a deep object detection networks to get the most confident bounding box for OD.
- The parameters of this bounding box are then used to generate an ellipse-like boundary which well approximates the OD appearance.
- If necessary, segmentation masks can be obtained based on the generated boundary.

A. Network Architecture

In this paper, we adopt *Faster R-CNN* as the object detector due to its flexibility and robustness to many follow-up improvements (*e.g.*, [**?**]). Faster R-CNN consists of two stages. The first stage, called region proposal network (RPN), processes images with a deep convolutional network *(e.g.*, VGG-16), and predicts a set of rectangular candidate object locations using features at some selected intermediate layer (*e.g.*, "conv5"). During training, the loss for this first stage is defined as

$$L = L_{cls} + L_{reg} \tag{1}$$

where L_{cls} and L_{reg} denote the classification loss and bounding box regression loss, respectively. We refer readers to [**?**] for more details of these two quantities..

In the second stage, these (*e.g.*, 300) candidate bounding boxes are mapped to the same intermediate feature space, and then fed to the subsequent layers of the convolutional network (*e.g.*, "fc6" followed by "fc7") to output a class label and a bounding box offset for each proposal. The loss function for this second stage box classifier also takes the form of (1) using proposals produced from the RPN as anchors.

B. Segmentation Generation

The object detector predicts a probability score for each candidate bounding box in the input image. Non-maximum suppression (NMS) is often required to reduce redundancy. However, for OD detection, a retinal fundus image contains one and only one object. Therefore, NMS is no longer necessary in our pipeline since we only need to retain the bounding box with the highest confidence score.

Given the bounding box, the question now is how to generate a satisfactory OD boundary. Since the OD appears as a bright yellowish elliptical region in color fundus images, it is promising to use an ellipse to approximate the shape of OD. Furthermore, we also observe that localizing a bounding box requires exactly the same parameters as a vertical ellipse, *i.e.*, its width, height, and central point (horizontal and vertical coordinates). With this in mind, we propose to generate the OD boundary by simply redrawn the predicted bounding box as a vertical ellipse.

IV. EXPERIMENTS

A. Dataset and Evaluation Criteria

We use the ORIGA dataset [2] for our experiments, and evaluate the proposed disc segmentation method using the manual mask as ?ground truth?. The ORIGA dataset contains 650 images with a uniform size of $3,072 \times 2,048$. These images are divided into 325 images for training and 325 images for testing. The OD are labeled by vertical ellipse.

Various evaluation criteria have been applied to OD detection/segmentation, among which the overlapping ratio may be the most widely used one. For space limitation, the proposed technique is only evaluated based on the overlapping ratio here. Let R denote the overlapping ratio between the manually labeled OD region and the estimated one as follows:

$$R = \frac{A_{\text{GT}} \cap A_{\text{ET}}}{A_{\text{GT}} \cup A_{\text{ET}}} \tag{2}$$

where A_{GT} and A_{ET} denote the area of the image region enclosed by the reference OD boundary and the area of the image region enclosed by the OD boundary determined by the tested method, respectively.

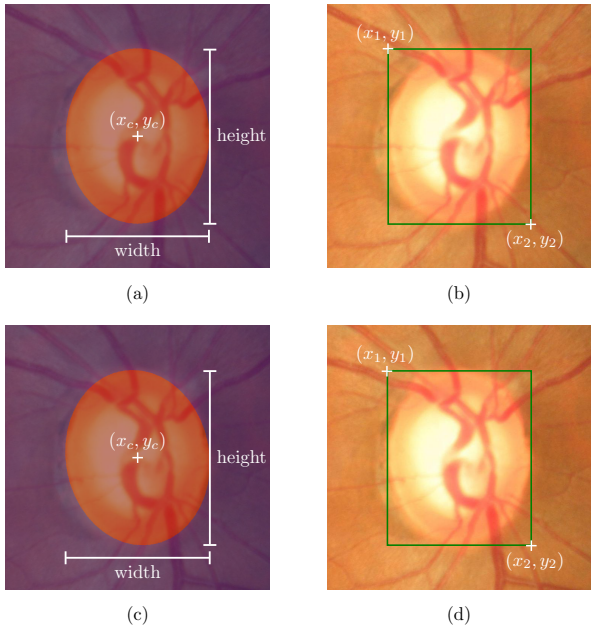

Fig. 2. One example from the training set to illustrate how we generate "ground truth" bounding box with the manual segmentation mask.

B. Implementation Details

Preprocessing: The manual "ground truth" provided by the ORIGA dataset is the segmentation masks, while what we need to train the object detector is the bounding box for OD. Therefore, we have to preprocess the dataset to enable training of Faster R-CNN. As mentioned in the previous subsection, the OD are labeled by vertical ellipse. It means that we can determine the OD boundary by only four parameters, *i.e.*, its width w, height h and center location (x_c, y_c), as illustrated in Fig. 2 (a). These parameters, on their own, can then be easily converted into coordinates of the upper left corner point (x_1, y_1) and the lower right corner point (x_2, y_2) of the bounding box, which is exactly what we want. But we found that there is another simpler way to accomplish that which allows us to bypass the tedious processing of fitting an ellipse first and then converting it parameters. To enable the bounding box to tightly localize the labeled OD segmentation mask, we simply let x_1 and x_2 denote the minimum and maximum horizontal coordinates of OD mask, and similarly, y_1 and y_2 denote minimum and maximum vertical coordinates.

Data argumentation: The training set of ORIGA dataset contains only 325 images, which is insufficient to learn so many parameters of the deep neural networks without overfitting. The easiest and most common method to reduce overfitting on image data is to artificially enlarge the dataset using label-preserving transformations. We employ two ways to argument our data. The first way is to rotate images by a set of angles over -10(2)10 degrees, where the notation -10(2)10 stands for a list starting from -10 to 10 with an increment of 2. Since image rotation also changes the location of OD in it, we have to modify the "ground truth" bounding box accordingly. This can be easily accomplished using the same method as what we do for the original dataset, provided that the manual segmentation masks are also rotated (as illustrated in bottom row of Fig. 2). The second way to argument our data is to use horizontal reflection, on both the original training set and it rotated counterparts. This increases the size of our training set by a factor of 20.

TABLE I
COMPARISON OF MEAN OVERLAPPING RATIO ON ORIGA DATASET.

Method	R (%)
MCV [13]	87.1
ASM [14]	88.7
EHT [15]	89.7
MDM [16]	89.2
SP+ASM [17]	90.5
SDM [18]	91.09
U-Net [19]	88.5
M-Net [20]	92.9
Proposed	**93.12**

Training details: The join-training scheme is adopted to train the Faster RCNN detection framework. The network is implemented using `Tensorflow` based on the publicly available code provided by Chen *et.al.* [11]. Several minor revisions are made in this implementation, which give potential improvements. Interested readers may refer to the technical report [11] for more details about the modifications. We trained and test the proposed method on a single-scale image using a single model. We rescale the images such that their shorter side is $s = 600$ pixels before feeding them to detector. VGG-16 [12] is used as the backbone of the Faster R-CNN and is initialized with an ImageNet pre-trained model. For anchors, we use 3 scales with box areas of 128^2, 256^2, and 512^2 pixels, and 3 aspect ratios of $1:1$, $1:2$, and $2:1$. The entire network is fine-tuned end-to-end with a training set of 7,150 images for 200,000 iterations (about 28 epochs) on a single NVIDIA TITAN XP GPU. The learning rate is set to 0.001 at the beginning of the training process and then changed to 0.0001 after 100,000 iterations.

Testing: The pixels inside the predicted bounding box are labeled as OD. This region is then used to calculated the overlapping ratio with the manual "ground truth" from ORIGA dataset.

C. Segmentation Results

In Table I, we compare the proposed method with the modified ChanVese method (MCV) [13], active shape model (ASM) [14], elliptical Hough transform (EHT) [15], modified deformable models (MDM) [16], superpixel-based method with ASM post-processing (SP+ASM) [17], supervised descent method (SDM) [18], and two deep learning based methods, *i.e.*, U-Net [19] and M-Net [20]. As shown in Table I, our detection based method achieves state-of-the-art segmentation results on ORIGA dataset, with the average overlapping ratio of 93.12% for OD segmentation. Fig. 3 shows some visual examples of the segmentation results, where the first two columns are images from which the

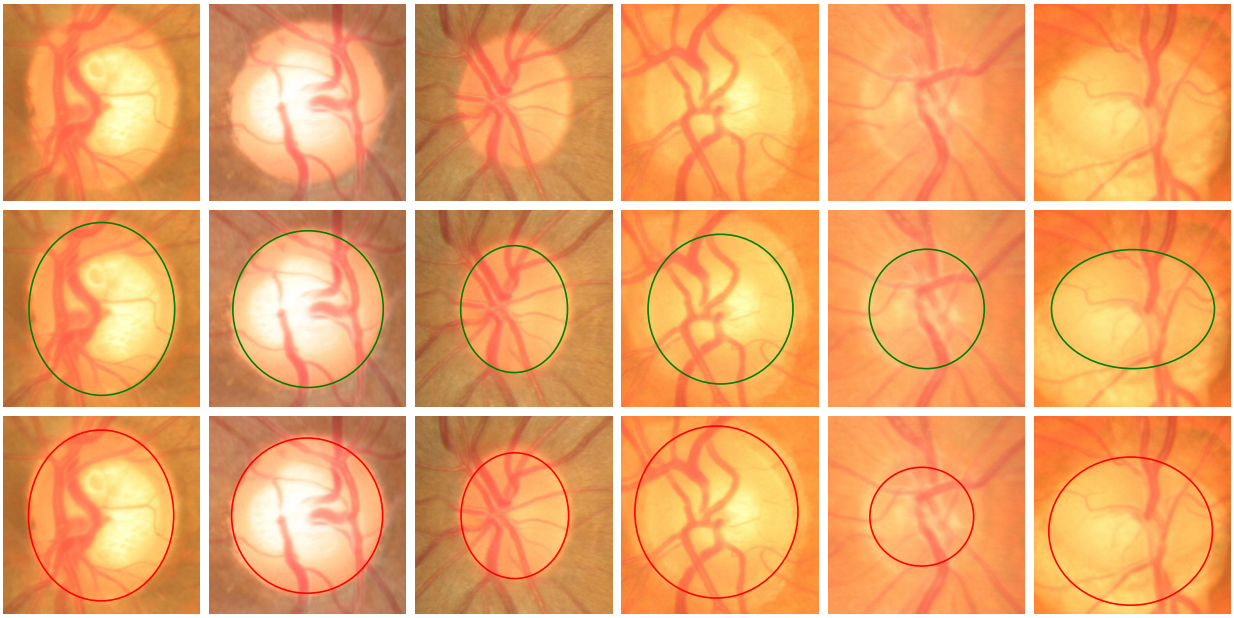

Fig. 3. Sample results. From top to bottom: the cropped original images, the manual "ground truth" and outlines by the proposed method. From left to right, the overlapping ratio by the proposed method are 98.23%, 98.20%, 98.06%, 77.01%, 75.00% and 73.68%, respectively

proposed method obtains highest overlapping ration and the rest columns are lowest ones.

V. CONCLUSION

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. It may be used for A4 paper size if the paper size setting is suitably modified. References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[2] Z. Zhang, F Yin, J. Liu, W. Wong, N. Tan, B. Lee, J. Cheng, and T. Wong, "Origa-light: An online retinal fundus image database for glaucoma analysis and research," in *EMBC*. IEEE, 2010, pp. 3065–3068.

[3] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*, 2014.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "SSD: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *CVPR*, 2017, pp. 2980–2988.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *NIPS*, 2016, pp. 379–387.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *ICCV*. IEEE, 2017, pp. 2980–2988.

[11] Xinlei Chen and Abhinav Gupta, "An implementation of faster rcnn with study for region sampling," *arXiv preprint arXiv:1702.02138*, 2017.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[13] Gopal Datt Joshi, Jayanthi Sivaswamy, and SR Krishnadas, "Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment," *IEEE Trans. medical imaging*, vol. 30, no. 6, pp. 1192–1205, 2011.

[14] Fengshou Yin, Jiang Liu, Sim Heng Ong, Ying Sun, Damon WK Wong, Ngan Meng Tan, Carol Cheung, Mani Baskaran, Tin Aung, and Tien Yin Wong, "Model-based optic nerve head segmentation on retinal fundus images," in *EMBC*. IEEE, 2011, pp. 2626–2629.

[15] Jun Cheng, Jiang Liu, Damon Wing Kee Wong, Fengshou Yin, Carol Cheung, Mani Baskaran, Tin Aung, and Tien Yin Wong, "Automatic optic disc segmentation with peripapillary atrophy elimination," in *EMBC*. IEEE, 2011, pp. 6224–6227.

[16] Juan Xu, Opas Chutatape, Eric Sung, Ce Zheng, and Paul Chew Tec Kuan, "Optic disk feature extraction via modified deformable model technique for glaucoma analysis," *Pattern recognition*, vol. 40, no. 7, pp. 2063–2076, 2007.

[17] Jun Cheng, Jiang Liu, Yanwu Xu, Fengshou Yin, Damon Wing Kee Wong, Ngan-Meng Tan, Dacheng Tao, Ching-Yu Cheng, Tin Aung, and Tien Yin Wong, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Transactions on Medical Imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

[18] Annan Li, Zhiheng Niu, Jun Cheng, Fengshou Yin, Damon Wing Kee Wong, Shuicheng Yan, and Jiang Liu, "Learning supervised descent directions for optic disc segmentation," *Neurocomputing*, vol. 275, pp. 350–357, 2018.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[20] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Transactions on Medical Imaging*, 2018.