AUTOMATIC OPTIC DISK AND CUP SEGMENTATION OF FUNDUS IMAGES USING DEEP LEARNING

Venkata Gopal Edupuganti, Akshay Chawla, Amit Kale

Bosch Research and Technology Centre India, Bangalore - 560095
Email: {VenkataGopal.Edupuganti, fixed-term.Akshay.Chawla, AmitArvind.Kale}@ in.bosch.com

ABSTRACT

Automatic segmentation of optic disk (OD) and cup regions in fundus images is essential in deriving clinical parameters, such as, cup-to-disk ratio (CDR), to assist glaucoma diagnosis. This paper presents a deep learning system using fully convolutional neural networks (FCN) to perform such segmentation, discusses various strategies on how to leverage multiple doctor annotations and prioritize pixels belonging to different regions while training the neural network. Experimental evaluations on Drishti-GS dataset demonstrate that the presented method achieves comparable and superior F-score to prior work on optic disk and cup segmentation, respectively.

Index Terms— Semantic Segmentation, Neural Networks, Glaucoma, Deep Learning, Fundus Imaging

1. INTRODUCTION

Glaucoma[1, 2] is a chronic and irreversible neuropathy caused by the progressive degeneration of retinal nerve fibers. Large scale screening programs can detect it at an early stage and inhibit its progression through proper treatment. Glaucoma causes structural changes in the optic nerve head, consisting of the optic disk (OD) and cup, that can be measured by the cup-to-disk diameter ratio (CDR). A CDR of at least 0.65 is considered as glaucomatous in clinical practice [1]. The difference between a glaucomatous and healthy eye is shown in Figure 1a and 1b, respectively.

Advanced imaging methods [1, 2], such as, optical coherence tomography (OCT), capture 3D information that is used by experts to identify cup and disk boundaries. However as these methods are expensive, color fundus imaging (CFI) [3] is used as non-invasive and inexpensive alternative for large scale screening. In CFI, a monocular camera captures a projection of the fundus onto a 2D plane. Due to a lack of depth information, an expert takes eight minutes to annotate each image [4]. Hence, an automated system that marks the boundaries will be extremely helpful. Although several segmentation methods exist [5, 6, 7], generation of a reliable cup boundary from the CFI is still a challenging task due to the lack of a clear visual demarcation between the cup and
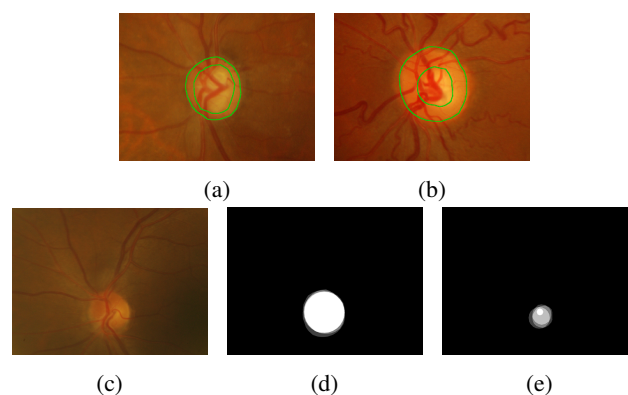

Fig. 1: Cup and Disk boundaries (green) for (a) Glaucomatous eye and (b) Healthy eye. Disagreement amongst annotators for a (c) sample image on (b) optic disk boundary, and (c) cup boundary; brighter pixels mean higher consensus among experts.

the disk. This is also shown in Figure 1e as a large amount of disagreement among experts while marking the cup boundary for a sample image from Drishti dataset [3].

Prior work using classical methods can be broadly classified into monocular and stereo based methods. In [6], the authors estimate depth discontinuity at the cup boundary by relating sequentially acquired images via a motion model. Chakravarthy et al. [5] model relative depth and discontinuity on cup boundary by correlating color, shading, and texture gradients to depth maps from OCT. Aquino et al. [8] use a combination of morphological features, edge detection, and hough transforms to fit OD boundary while [9] uses an active-contour based approach to parameterize OD. In [10], edge and wavelet transforms are used to identify small vessel bends (kinks) to establish the cup boundary.

Zilly et al. [7] train a set of convolution filter banks as linear-regressors over the groundtruth of extracted patches. Pre-processing consists entropy filtering, color space conversion and contrast normalization. Due to the complex training regime, test images are cropped on the optic disk before inference. Similarly, [11] follows the same pre-processing and training routine as [7], while adding a bias and regularization

term to the objective.

In recent years, deep convolutional neural networks [12] (CNN) have shown state of the art performance on several challenging tasks, such as, image classification and semantic segmentation. A popular architecture used for semantic segmentation is fully convolutional network (FCN) [13], which uses a VGG16 based encoder-decoder network with skip connections to perform one-shot inference on full-scale images. Thus we choose the FCN architecture as the basis of our solution.

In [4], the authors train a shallow CNN for cup and disk segmentation on 48x48 patches from MESSIDOR and SEED database. Deep Retinal Image Understanding [14] uses features extracted from a CNN pre-trained on ImageNet [15] along with specialized convolution layers to predict blood vessels and OD from CFI. Unlike [4] and [14], we focus on the problem of OD and cup segmentation instead of blood vessels from the Drishti database, leverage multiple annotations, and seek to perform one-shot segmentation on full scale images. In [16], the author trains two UNets [17], for cup and disk segmentation separately. In this two-stage approach, the input image is resized to 256x256, heavily pre-processed and inferenced twice to predict cup and disk boundary. Instead we propose a one-shot segmentation pipeline that works on full-scale images and uses a simple post-processing method. Both [14] and [16] have acknowledged that segmentation of OD can be automated to human level accuracy but that of cup is non trivial.

Our main contributions are summarized below.

1. Present a system using FCN8s architecture that generates cup and disk segmentation in a single shot using one deep neural network on full resolution images.

2. Propose various strategies of utilizing multiple expert annotations and prioritizing certain regions during training for optimal boundary retrieval.

We also use a post-processing technique (section 3.2) that filters out network predictions to reduce false positives. Comparison on the Drishti dataset against prior work [9, 6, 3, 11, 16] shows comparable and superior F-scores in disk and cup segmentation, respectively.

Paper Organization. In Section 2 we present the problem statement and a comprehensive description of Drishti-GS dataset. Section 3 presents system overview along with detailed description of underlying components and strategies. Experimental results are discussed in section 4. Finally, section 5 concludes the paper with future work.

2. PROBLEM STATEMENT AND DATASET

The problem of optic disk (OD) and cup segmentation is posed as a ternary classification problem, where each pixel of an input CFI image is classified as cup, disk, or background.

Though there are datasets available for OD segmentation [1, 18], only Drishti-GS dataset [3] has both OD and cup annotations by multiple experts. It consists of two sets of CFI images: 50 for training and 51 for testing. A subset of 10 images from training set is kept aside for validation purposes. Groundtruth (GT) for each OD and cup is provided by four ophthalmologists as a softmap shown in Figures 1d and 1e, respectively. To be comparable with prior work, we set the groundtruth by taking the consensus of three or more doctors. In order to leverage CNNs effectively and introduce invariance, we augment the training data by horizontal and vertical flips, yielding a total of 160 training images. Instead of downsampling images, we take a center crop of 1200 x 1600, thus preserving the fidelity of the original image. On an average, each image in the dataset consists of 2.84% cup pixels, 2.86% disk pixels (after subtracting cup pixels inside disk region), and 94.3% background pixels.

3. SYSTEM OVERVIEW

A fully convolutional network (FCN) [13] with a stride of 8 pixels is chosen as an underlying architecture due to consistent semantic segmentation performance on other public datasets. It consists of a VGG16 encoder along with a decoder comprising upsampling layers and skip-connections that can be trained in an end-to-end fashion using a pixel-wise cross-entropy loss. Due to limited number of training images, the fcn encoder is initialized with weights from a model trained on ImageNet [15]. Additionally, as shown in Figure 2g, we use weighted loss based on a weight mask to prioritize pixels during the training process. During inference, the predicted segmentation is passed through a filtering module that cleans up noisy regions to generate clean OD and cup boundaries.

3.1. Strategies

In order to rightly capture the disagreement in annotations amongst experts and to account for the relatively small area occupied by OD and cup regions in comparison to the background, we design the following five experiments.

1. **Exp1:** This is a vanilla experiment which trains a FCN using a CFI image and the GT generated as in Section 2. Each pixel contributes equally to the loss.

2. **Exp2:** To account for the imbalance between the number of background pixels and cup/OD pixels, we use a weighted loss that assigns a 10x higher loss to pixels belonging to OD/cup regions in this experiment. The weight mask in Figure 2c is a grayscale image that encodes this weighing scheme and is passed as an input to a weighted cross-entropy loss layer.

3. **Exp3:** The accuracy of CDR depends upon the quality of the boundaries generated by the neural network. In
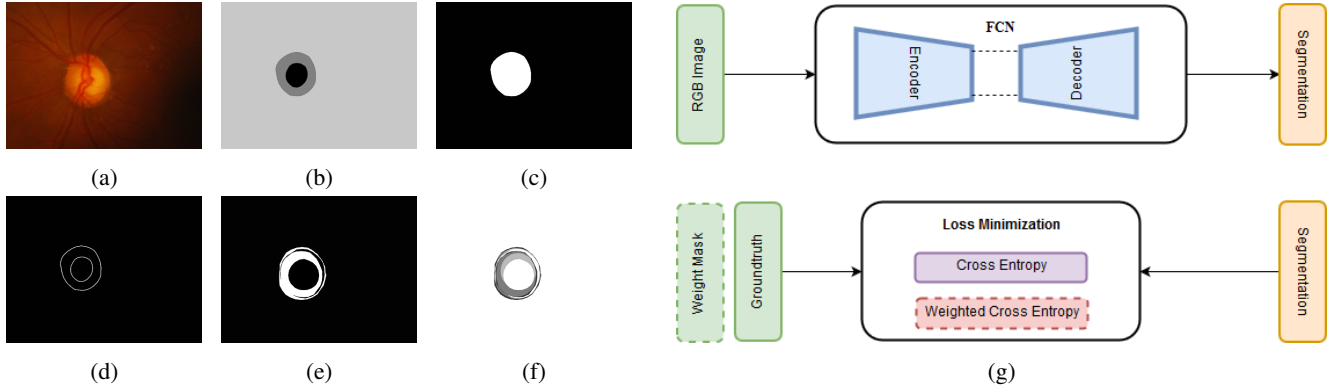
Fig. 2: Data and System overview. (a) input image, (b) groundtruth (inner to outer:cup, disk, and background), (c) weight mask (pixel values of OD and cup:10, background:1), (d) boundary weight mask (pixel values of OD and cup boundaries of 5 pixels width:10, rest:1), (e) ignore disagreement (white) pixels in groundtruth, and (f) soft disagreement weights (weight decreases with pixel brightness). (g) Training of FCN involves minimization of cross entropy loss between groundtruth and the network output through backpropagation. Inputs (shown in green) are the RGB image, groundtruth segmentation and an optional weight mask (dotted box).
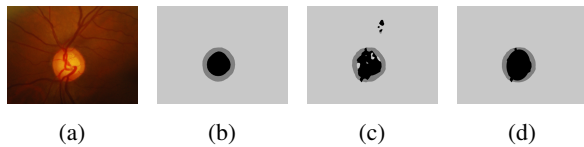

Fig. 3: Post processing. (a) input image, (b) groundtruth, (c) network output, and (d) filtered output.

order to bake this constraint into optimization problem, we use a weighted loss like Exp2, however, we assign a 10x higher loss to pixels belonging only to the OD and cup boundaries as shown in Figure 2d. Our reasoning behind this masking is that the network will now try to learn features to yield better boundaries.

4. **Exp4:** As shown in Figures 1e and 1d, annotators often disagree on the exact cup and disk boundaries. Since this disagreement might lead to poor convergence, we set their GT value to 255 as shown in Figure 2e so as to ignore the loss due to these regions. We now expect our network to learn distinctive features that are commonly looked at by all experts, rather than overfitting to an individual.

5. **Exp5:** To bridge the ideas of Exp3 and Exp4, in this experiment, we use a mask that is weighted by the amount of disagreement among experts, i.e., less weight for pixels with high disagreement as shown in Figure 2f.

3.2. Post processing

In order to reduce false positives and generate clean cup and disk boundaries, we use a filtering module on top of the network predictions. Figure 3 shows the impact of this module.

1. Remove regions with area less than 100 pixels to filter out small false positives.

2. Filter out regions where cup is not inside disk and keep the pair with maximum disk area.

3. (optional) Approximate contour points by an ellipse to obtain smooth boundaries.

4. EXPERIMENTAL RESULTS

Experiments are conducted on a system running Ubuntu 16.04 with a Nvidia K80 having 12GBs of GPU RAM. The training set of 40 images is augmented to 160 using image flips. Validation and test set contain 10 and 51 images, respectively. Each experiment in section 3.1 is trained for 100 epochs using an Adam optimizer with an initial learning rate of 1e-9. The FCN8s network with 134M parameters is initialized with ImageNet weights. We also apply a dropout of 0.5 after fc6 and fc7 layers to help with generalization on our limited dataset. We validate using the intersection over union metric as defined in Equation 1, where TP, FP, FN denote true positives, false positives, and false negatives, respectively. In order to compare with previous work on Drishti dataset, we also report F-scores as defined in Equation 2 where cup is considered as part of disk while computing disk F-score.

$$IoU = \frac{TP}{TP + FP + FN} \qquad (1)$$

Table 1: IoU evaluation

(a) Best validation set checkpoint

IoU	Segmentation Strategies				
	Exp1	Exp2	Exp3	Exp 4	Exp5
Mean	0.8509	0.8559	0.8660	0.8823	0.8055
Cup	0.8055	0.8240	0.8284	0.8450	0.7640
Disk	0.7507	0.7482	0.7743	0.8030	0.6591
Void	0.9964	0.9957	0.9962	0.9970	0.9926

(b) Test Set

IoU	Segmentation Strategies				
	Exp1	Exp2	Exp3	Exp4	Exp5
Mean	0.8346	0.8341	0.8266	0.7940	0.7795
Cup	0.8122	0.8139	0.7920	0.7720	0.7660
Disk	0.6958	0.6931	0.6911	0.6160	0.5802
Void	0.9959	0.9952	0.9959	0.9940	0.9917

Table 2: F-score comparison with prior state of the art

	cup	disk
G. D. Joshi et al. [9]	0.84	0.97
G. D. Joshi et al. [6]	0.85	–
J. Sivaswamy et al. [3]	0.79	0.96
J. Zilly et al. [11]	0.871	0.973
A. Sevastopolsky [16]	0.85	–
Proposed Exp1	**0.897**	**0.967**

$$F - score = \frac{2 * TP}{2 * TP + FP + FN} \qquad (2)$$

Table 1a reports cup, disk, background, and mean IoU values of the best epoch for each experiment on validation set. It is observed that Exp4 which ignores regions of disagreement among experts outperforms all other experiments and Exp5 which is an application of soft disagreement weights is an under performer. An examination of training curves reveal that Exp4 demonstrates less fluctuation over epochs while maintaining a higher mean IoU value. It is also observed that Exp2 exhibits quick learning during initial epochs due to the prioritization of regions. For large datasets where compute is limited, Exp2 could yield a reasonable accuracy in fewer epochs.

The best model is applied to the test set and the results are documented in Table 1b. It is observed that Exp4 which performed well while validating does not generalize well to the test set, however, Exp1 performs remarkably well and gives the best results, closely trailed by Exp2. A visual inspection of the dataset confirms that there is a difference between the distribution of training and test sets, with images in the test set being taken from a different pool of participants. However, to compare with prior work, we did not alter the data split.

Table 2 lists comparison of Exp1 with prior work. F-score of cup on test set is improved by 3.1% while that of disk is competitive. As mentioned earlier, reason for high disk F-scores across the board is due the low ambiguity in expert markings and the clear visual demarcation of the nerve head. Cup boundaries are harder to distinguish as shown in Figure 1e due to large disagreement among experts.

Figure 4 shows visualizations of a good and a bad prediction on the test set. A good prediction in Figure 4a has

predicted boundaries (blue) that are close to groundtruth annotations (green). This is due to a visible boundary between cup, disk, and background regions. The bad prediction from in Figure 4b shows that the performance of the network deteriorates when there is no visible transformation from OD to cup.

(a) (b)
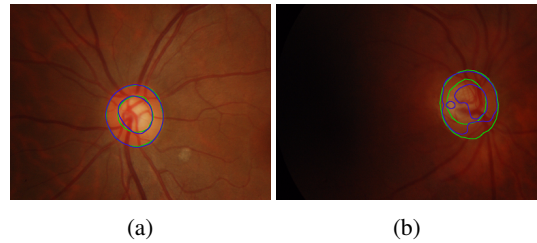
Fig. 4: Visualizations of results (a) well performing sample, (b) under-performing sample. GT boundary is in green and network predictions are in blue.

5. CONCLUSIONS AND FUTURE WORK

An automatic system to segment optic disk and cup from color fundus images is presented using deep learning. Unlike previous methods, the system predicts OD and cup boundaries in a single pass using fully convolutional networks and does not require resizing of original image or cropping to OD region for cup prediction. A brief discussion of various strategies on how to leverage multiple expert annotations and prioritize pixels belonging to different regions while training the neural network is presented along with extensive experimental results. Experimental evaluations on Drishti-GS dataset have shown comparable and superior F-score to prior art for optic disk and cup segmentation, respectively. Due to the complexity of the network, a focus of future work is on using network pruning techniques for parameter reduction, inference acceleration, and coming up with a compact architecture.

6. REFERENCES

[1] M Usman Akram, Anam Tariq, Shehzad Khalid, M Younus Javed, Sarmad Abbas, and Ubaid Ullah Yasin, "Glaucoma detection using novel optic disc localization, hybrid feature set and classification techniques," *Australasian physical & engineering sciences in medicine*, vol. 38, no. 4, pp. 643–655, 2015.

[2] Ahmed Almazroa, Ritambhar Burman, Kaamran Raahemifar, and Vasudevan Lakshminarayanan, "Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey," *Journal of ophthalmology*, vol. 2015, 2015.

[3] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish, "Drishtigs: Retinal image dataset for optic nerve head (onh) segmentation," in *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*. IEEE, 2014, pp. 53–56.

[4] Gilbert Lim, Yuan Cheng, Wynne Hsu, and Mong Li Lee, "Integrated optic disc and cup segmentation with deep learning," in *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*. IEEE, 2015, pp. 162–169.

[5] Arunava Chakravarty and Jayanthi Sivaswamy, "Coupled sparse dictionary for depth-based cup segmentation from single color fundus image," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 747–754.

[6] Gopal Datt Joshi, Jayanthi Sivaswamy, and SR Krishnadas, "Depth discontinuity-based cup segmentation from multiview color retinal images," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1523–1531, 2012.

[7] Julian G Zilly, Joachim M Buhmann, and Dwarikanath Mahapatra, "Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2015, pp. 136–143.

[8] A. Aquino, M. E. Gegundez-Arias, and D. Marin, "Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques," *IEEE Transactions on Medical Imaging*, vol. 29, no. 11, pp. 1860–1869, Nov 2010.

[9] G. D. Joshi, J. Sivaswamy, and S. R. Krishnadas, "Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment," *IEEE Transactions on Medical Imaging*, vol. 30, no. 6, pp. 1192–1205, June 2011.

[10] G. D. Joshi, J. Sivaswamy, K. Karan, P. R., and S. R. Krishnadas, "Vessel bend-based cup segmentation in retinal images," in *2010 20th International Conference on Pattern Recognition*, Aug 2010, pp. 2536–2539.

[11] Julian Zilly, Joachim M Buhmann, and Dwarikanath Mahapatra, "Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation," *Computerized Medical Imaging and Graphics*, vol. 55, pp. 28–41, 2017.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[14] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool, "Deep retinal image understanding," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 140–148.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[16] Artem Sevastopolsky, "Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network," *arXiv preprint arXiv:1704.00979*, 2017.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[18] Enrique J Carmona, Mariano Rincón, Julián García-Feijóo, and José M Martínez-de-la Casa, "Identification of the optic nerve head with genetic algorithms," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243–259, 2008.