Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Deep learning based retinal OCT segmentation

M. Pekala[a], N. Joshi[a], T.Y. Alvin Liu[b], N.M. Bressler[b], D. Cabrera DeBuc[c], Burlina P.[a,b,*]

[a] Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
[b] Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[c] Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Miami, FL, USA

ARTICLE INFO

Keywords:
Fully convolutional networks
Gaussian process regression
OCT segmentation
Neurodegenerative
Retinal and vascular diseases

ABSTRACT

We look at the recent application of deep learning (DL) methods in automated fine-grained segmentation of spectral domain optical coherence tomography (OCT) images of the retina. We describe a new method combining fully convolutional networks (FCN) with Gaussian Processes for post processing. We report performance comparisons between the proposed approach, human clinicians, and other machine learning (ML) such as graph based approaches. The approach is demonstrated on an OCT dataset consisting of mild non-proliferative diabetic retinopathy from the University of Miami. The method is shown to have performance on par with humans, also compares favorably with the other ML methods, and appears to have as small or smaller mean unsigned error (equal to 1.06), versus errors ranging from 1.17 to 1.81 for other methods, and compared with human error of 1.10.

1. Introduction

Optical coherence tomography (OCT) is an important retinal imaging modality as it is a non-invasive, high-resolution imaging technique capable of capturing micron-scale structure within the human retina.

The advent of OCT has transformed ophthalmology and neurology. In ophthalmology, OCT is currently instrumental in the diagnosis, prognostication and treatment management of diabetic retinopathy (DR) and age-related macular degeneration (AMD). This has had far-reaching impact, as diabetic retinopathy is the leading cause of blindness in the developed world among the working-age population [1], while age-related macular degeneration is the leading cause of central vision loss throughout North America [2] and other developed countries in patients aged 50 and over. In terms of embryonic development, the retina is part of the central nervous system (CNS) and is organized into different layers (Fig. 1). In neurology, detailed cross-sectional imaging of the retina by OCT has provided neurologists with new biomarkers, as it has been shown that retinal thinning, in particular thinning of the retinal nerve fiber layer (RNFL), is associated with various neurological disorders such as stroke, Parkinson's disease and Alzheimer's disease [3]. Different ophthalmic and neurological diseases affect different layers of the retina in various forms. For example, diabetic macular edema (DME) typically leads to fluid accumulation within the retina; neovascular AMD is usually characterized by fluid underneath the retina; multiple sclerosis can be associated with alteration of the RNFL thickness. These various pathologies and their resulting effects on the retina highlight the importance of the ability to accurately analyze retinal structure in retinal OCT and in particular the importance of accurate OCT segmentation. Manual segmentation is labor and time intensive. Therefore, automatic, reliable, accurate OCT segmentation is crucial for further expanding the usefulness of the OCT technology.

Work in machine learning, medical image analysis and specifically automated retinal image analysis (ARIA) has steadily progressed in the past two decades, as datasets have become more plentiful, and machine vision and machine learning techniques have improved as well (e.g., Refs. [4–11]). This progress has also benefited automatic OCT segmentation. A substantial body of work in automated segmentation of medical images and OCT segmentation in particular has been directed towards the use of statistical and graph based methods [12–21]. In a recent study [22], a 7 layer OCT segmentation using kernel regression (KR)-based classification was developed to analyze diabetic macular edema (DME) as well as OCT layer boundaries. This method was then combined with an approach using graph theory and dynamic programming (GTDP) and validated on 110 B-scans from ten patients with severe DME, yielding a DICE coefficient of 0.78.

Work in deep learning has had a substantial impact in medical imaging [10,23,24] in general and ARIA in particular, including studies that have demonstrated automated detection of patients with referable

[*] Corresponding author. Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA.
E-mail address: pburlin2@jh.edu (P. Burlina).

https://doi.org/10.1016/j.compbiomed.2019.103445
Received 24 November 2018; Received in revised form 11 September 2019; Accepted 12 September 2019
Available online 17 September 2019

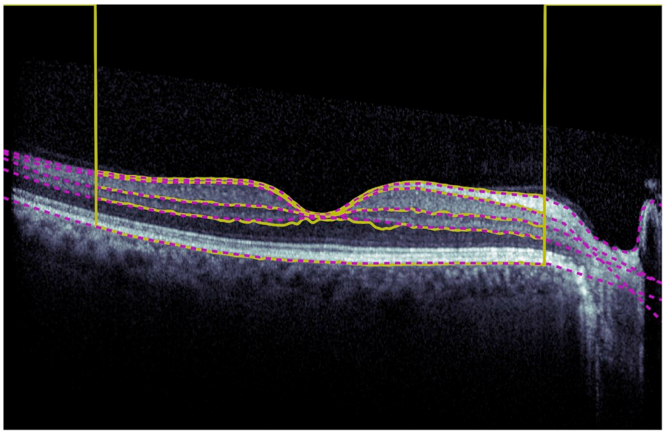

Fig. 1. Example annotations from the dataset of [13]. The yellow lines depict the AURA surface estimate; note the estimate does not span the entire image. Consequently, comparison with AURA must be limited to the regions where estimates are available. Magenta lines denote the estimates generated by human observer #1 which are used as ground truth in this study.

age related macular degeneration from fundus images [25,26], AMD fine grained severity classification [23,27–29], detection of DR [30], and the estimation of cardiovascular risks factors from retinal fundus imagery [31] or AMD in retinal OCT [32].

For OCT segmentation, some recent studies have featured the use of convolutional neural networks (ConvNets) and fully connected networks: [33] used ConvNets and a U-Net architecture to delineate macular edema and obtained DICE of 0.91 with performance on par with humans. Another recent study [34] used a cascaded U-Net architecture [35] and compared performance to that of a classical approach based on random forests, and [36] used hybrid ConvNets and graph based method to identify OCT boundary layers. Recent efforts at the University of Miami [13] led to the development of a publicly available OCT dataset with clinical gold standards for comparing performance among methods, including a number of OCT segmentation algorithms of record. Additionally [37], used a contracting convolutional network (encoder) to learn a hierarchy of contextual features, which was combined with an expansive convolutional network (decoder) for semantic segmentation of OCT, and obtained good performance results when compared to other FCNs using a dataset that included patients with edema. In Ref. [38], a diluted residual U-Net like architecture was used to segment optical nerve head on OCT with good results. For additional reviews of the state of the practice for OCT segmentation, see Refs. [20,21]. More discussions of the above studies are further included in the discussion section.

In contrast to the aforementioned methods, the novelty of our OCT segmentation approach consists of using a fully convolutional network (FCN) using the DenseNet architecture and a Gaussian process regression. We do a performance comparison both with classical methods of record and against human clinicians' performance. Our proposed approach performs on par with human clinicians and also compares favorably against other methods of record. This comparison is done using the publicly available University of Miami dataset [13]. In particular, we show that our method exhibits the smallest unsigned boundary estimation errors. This result is promising for clinical applications, especially for neurological and retinal disorders that manifest with retinal layer thinning on OCT. The proposed method and its results in this publicly available dataset can serve as a benchmark against which future ML-based OCT segmentation algorithms can be compared. We demonstrate the improvements and benefits of using Gaussian Processes, a post-processing technique which can be used also with other DL methods and FCNs.

2. Methods

2.1. Data

For our study we utilize the publicly available University of Miami OCT dataset [13]. This includes 50 OCT of 10 different patients with mild, non-proliferative diabetic retinopathy. Each image consists of 768×496 pixels with transversal and axial resolutions of 11.11μ m/pixel and 3.86μ m/pixel. These images are from a set of volumetric data captured by a Spectralis SD-OCT (Heidelberg Engineering GmbH, Heidelberg, Germany). There are five images available for each patient, which includes one image of the fovea center, two of the perifovea, and two of the parafovea. Two expert graders each annotated five retinal surfaces per image, where a "surface" is defined as the boundary between a pair of adjacent retinal layers (see Fig. 1). The result is a total of 250 annotated surfaces per grader. The annotated surfaces are numbered 1,2,4,6 and 11 (following the convention introduced in Ref. [13]). These surfaces and the associated layers are described in Table 1. Following the approach in Ref. [13], we use the first grader's annotations as ground truth and the second grader's annotations as a measure of inter-operator agreement.

2.2. Segmentation approach

Our approach for estimating retinal surfaces consists of two primary steps. The first solves a per-pixel (or "dense") classification problem of associating each pixel in the image with the most likely corresponding retinal layer. These per-pixel estimates are then post-processed via a regression procedure which models retinal surfaces as smooth functions. Note that, while our current experiments involve two-dimensional images, both steps above extend naturally to three dimensions. Thus, our approach is extensible to settings where labeled volumetric data is available.

2.2.1. Semantic Segmentation

Fully convolutional neural networks (FCNs) provide effective and computationally efficient alternatives to sliding window approaches [40] for image processing problems where per-pixel labeling is desired. FCNs are a subcategory of ConvNets that take tensor-like data as input and produce class estimates having the same spatial dimensions (i.e. per-pixel or per-voxel labels). For this application we elected to use the FCN version of DenseNet [39,41] (Fig. 2). DenseNets are characterized by an extensive use of "skip connections" which permit each layer of the network to directly process the outputs from all previous layers (see Fig. 2). This construction makes richer sets of features available at each layer of the network while also providing a mechanism to alleviate the vanishing gradient problem which can arise during training. This is in contrast to more traditional networks which generate features in a strictly serial fashion (i.e. each layer operates solely upon the output of the previous layer). Other earlier FCN architectures, such as U-Nets [35], also directly propagate a subset of features maps; however these intra-layer connections are less abundant relative to the DenseNet architecture.

For our experiments we adopted the 103 layer DenseNet-FCN architecture described in Ref. [39]; in particular, we used the publicly-available Keras implementation of [42]. We used the

Table 1
Annotated surfaces provided by dataset in Ref. [13].

Surface ID	Upper Layer	Lower Layer
1	Pre-retinal space	Nerve fiber layer
2	Nerve fiber layer	Ganglion cell layer
4	Inner plexiform layer	Inner nuclear layer
6	Outer plexiform layer	Henle's Fiber layer and Outer nuclear layer
11	Bruch's complex	Choriocapillaris

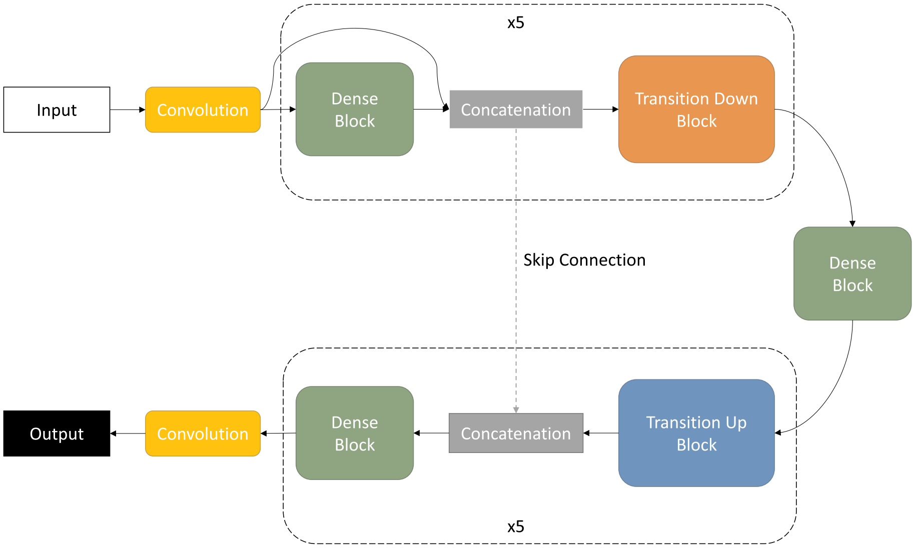

Fig. 2. Simplified depiction of network architecture for the fully convolutional version of DenseNet, summarized in Ref. [39].

fundamental architecture of this network and our modifications consisted in adjustments to the loss function and the synthetic data augmentation methodology described below. During training, we minimized the pixel-wise cross entropy loss using the Adam [43] optimizer, with a learning rate of $1e-3$. Due to memory constraints, each input training example was taken as a vertical slice of 256×512 pixels that had been randomly cropped from the original image. These input slices were then fed in small mini-batches of cardinality 2 (again, due to memory considerations). In addition to random cropping, input batches were further augmented with horizontal flipping, image blurring, image sharpening, and brightness adjustments.

Variations in thickness of retinal layers introduces a non-trivial amount of class imbalance (as there are fewer pixels inside some of the thinner, inner retinal layers). To mitigate the impact of this class imbalance in training we increased the weight in the loss penalty for the pixels associated with minority classes by a factor of 10 (roughly corresponding to the level of class imbalance). The model was trained for 500 epochs and model weights were saved whenever performance on the validation set improved. Training the model took roughly 24 h on an NVIDIA Titan X GPU and inference time on our test set was only a few hundred milliseconds, whereas the processing time of algorithms reported on in Ref. [13] can range from 28 to 152 s.

During training, we used a K-fold cross validation ($K = 10$) process where we used nine sets of five images (resulting in a total of 45 images) from nine patients for training a FCN, and testing was done on the remaining test patient's five images (in essence, a "leave-one-patient-out" procedure). Then the patient used for testing was rotated as is done in conventional K-fold testing approaches, resulting in testing performed on all images. Of the nine patients available for training in a given fold, one patient was reserved as a validation set. This stratification allowed us to train the network on representative data while ensuring that the segmented images for a given patient were not a by-product of training on that patient's images. A few of the images contain regions that consist of all zero pixels; these regions were not used during training (although they were evaluated at test time).

2.2.2. *Post-processing via Gaussian Processes*

With per-pixel estimates in hand, one might attempt to directly extract surfaces from the layer estimates (e.g. by identifying locations where class estimates change along the axial dimension). However, surfaces are defined as a *unique* location in the axial dimension where the layer estimates change and the raw semantic segmentation outputs do not necessarily satisfy this constraint. A potential problem with this direct DL approach using solely DenseNets is that errors in individual pixel estimates can introduce missing or spurious class transitions. This is exemplified in the right panel of Fig. 3 which shows a few small regions indicated by arrows where monotonicity of class estimates is violated or some other undesirable artifacts appear. In this case, the cluster of bright spots towards the top of the image in the left panel has produced a spurious region of estimates. Local heuristics can be used to address such issues; e.g. using statistical methods to resolve duplicate or missing surface estimates.

We explored one heuristic which addresses both spurious and missing estimates. If the classification procedure generates more than one candidate for a layer at a given location, the point which is nearest in Euclidean distance to the prior surface is used (in the case of surface 1, distance to surface 2 is used as the adjudication method). Alternately, if a layer estimate is missing for any given location, an estimate is imputed from the nearest available value for that layer. This particular heuristic coupled with the DenseNet FCN segmentation constitutes a baseline algorithm which we term "SEG".

As hand-crafted heuristics such as described above may be considered somewhat ad-hoc, an alternative to making arbitrary local repairs is to explicitly use prior knowledge that retinal surfaces (in two-dimensional images) can be modeled as scalar-valued functions with an appropriate level of smoothness. This suggests the use of a post-processing procedure that solves a suitable regression problem for each surface. Depending upon the regression procedure, this notion then extends naturally to higher dimensions as well (e.g. for settings where volumetric data is available).

In this study, we employed Gaussian processes (GP) regression with a Radial Basis Function (RBF) kernel [44]. The RBF kernel has two hyper-parameters, a noise variance and a characteristic length scale; we selected both using a leave-one-patient-out cross-validation procedure analogous to what was done when training the CNN. We used the GPy software library [45] to implement the regression and optimized the hyperparameters by random search via candidate hyper-parameters drawn by uniform random sampling from a 2D hypercube. In this study, we observed that using a single kernel for each patient/region pair produced adequate results; however, in settings where there is substantial non-stationarity in the behavior of a patient/region further partitioning methodologies may be of value (e.g. Ref. [46]). The GP is characterized by the combination of a mean and a covariance function; the mean function was used as our best estimate for the corresponding surface while the covariance provides some measure of the confidence of

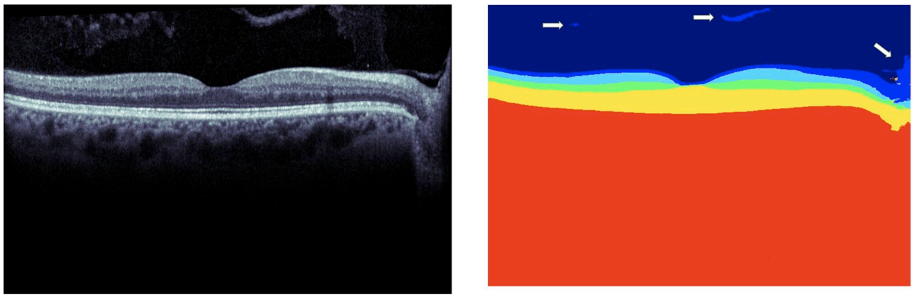

Fig. 3. Example segmentation; original image (left); neural network segmentation output, before post-processing (right). White arrows denote challenging regions where semantic segmentation layer estimates exhibit non-ideal behavior.

the estimate. While these estimates are all 1D functions, we note that GPs extend directly to higher dimensions.

Other post-processing approaches are of course possible; using GPs provides an interesting alternative in that the built-in confidence measure may provide some additional diagnostic value. This could be especially compelling in settings where images are less uniform in quality or surfaces are more structurally diverse. We term this combined FCN and GP approach “SEG + REG”.

2.3. *Comparison with other approaches*

In addition to OCT images and ground truth, the publicly available University of Miami OCT dataset [13] also includes annotations generated by five commonly used OCT segmentation software packages and/or algorithms of record. These reference algorithms/implementations are: Spectralis 6.0 [16], IOWA Reference Algorithm [17], AUtomated retinal analysis tools (AURA) [18], Dufour’s (Bern) algorithm [19], and OCTRIMA3D [20]. We refer the reader to Ref. [13] for a complete description of these algorithms. Note that the algorithms of record do not always produce estimates which span the entire OCT image (e.g., see Fig. 1). For a fair comparison, performance evaluation is restricted to the subset of each image for which all algorithms produced a valid surface estimate. Note that our algorithm produces estimates for the entire image.

2.4. *Evaluation methods and metrics*

As mentioned in sect. 2 we used a K-fold cross validation procedure to evaluate our algorithm’s performance. Following the approach in Ref. [13], we measure the accuracy of surface estimates by computing the per-pixel differences between the estimate and the ground truth annotations generated by the first manual grader. For a fair comparison, metrics calculations are limited to the regions for which all automated algorithms in the dataset had valid estimates. This unfortunately excludes remote/lateral regions where cut artifacts are more prevalent; these are operator-induced artifacts where the edge of the scan is abnormally truncated (a defect which does not affect central retinal thickness measurements). We used mean unsigned errors and mean signed errors as performance metrics for both the proposed algorithms and algorithms of record. For a given surface, the estimate v_{est} and the corresponding ground truth v_{ref} are both vectors (with dimension equal to the width of the evaluation region, in pixels) and the signed error is defined to be

$$e_s = v_{ref} - v_{est};$$

the unsigned error is the absolute value of e_s taken component-wise.

3. Results

We report the performance of both the SEG and SEG + REG, as compared to the performance of other algorithms. Table 2 reports the mean unsigned errors for each algorithm and surface, and the average and max values across all testing data. Values in bold font indicate when an algorithm meets or exceeds human performance (i.e., is within range of human inter-operator error). The table suggests that in aggregate the proposed methods match human performance, and perform favorably when compared to other algorithms of record. These results indicate good performance of the proposed methods on the layers within the neurosensory retina (layers 1, 2 and 4) as well as the outer plexiform layer and Bruch’s complex (6 and 11). Table 3 shows the signed errors for the corresponding regions, from which it appears that our method may be slightly overestimating the support of the retinal layers as evidenced by a relatively large positive error on surface 1 and a relatively large negative error on surface 11. Following [13] we also provide the mean unsigned error broken down by retinal layers and areas within the macula in Table 4.[1]

4. Discussion

We present results demonstrating that semantic segmentation using a fully convolutional network via application of DenseNets together with post-processing based on GP is a promising approach to address the problem of fine-grained automated OCT segmentation, a capability with potential clinical applications, especially in those applications using OCT measures as biomarkers for neurodegenerative diseases. Although this was not tested in our study, it may also help in identifying other clinically important structures (such as drusen) by considering the semantic segmentation component. Our experiments also suggest that the use of DenseNets alone can be improved by the post-processing of GPs.

When compared to other methods of record used in the University of Miami study, our results compare well, often resulting in the smallest mean unsigned errors. Overall, performance is largely comparable with human annotation. Caution should be exercised when interpreting such strict comparisons, since the algorithms of record we compared against were developed and optimized using datasets that may be different from the University of Miami evaluation dataset in terms of resolution, noise characteristics, and artifacts.

Recently, several other methods have been reported in the literature using various deep learning approaches with promising performance as well. Most of these methods rely on FCNs inspired from U-Nets. The application of FCNs in RelayNet for example in Ref. [37] was done to both retinal layers as well as fluid segmentation. That study used an architecture that is related to U-Nets, with modified pooling and un-pooling layers and used a loss function including weighted logistic regression and a DICE overlap. It was evaluated using the Duke OCT

[1] Note there is some minor difference between these results and table 5 of [13] for the algorithms of record which may be attributed to variations in the extent of the macular region that was evaluated; many of the automated methods tend to exhibit greater variation towards the edges of the scans and we evaluate on the largest common intersection across all algorithms.

Table 2
Mean unsigned error aggregated across all OCT layers. Values in bold indicate when an algorithm meets or exceeds inter-observer (I–O) performance.

	SEG	SEG + REG	Spectralis	OCTRIMA	AURA	IOWA	Bern	I–O
surface 1	1.13	1.11	1.09	0.95	1.35	2.03	1.71	0.87
surface 2	**1.14**	**1.07**	1.45	1.18	1.19	1.74	2.77	1.14
surface 4	**0.95**	**0.90**	1.92	**0.99**	1.12	1.79	1.60	1.10
surface 6	**1.23**	**1.18**	**1.19**	1.52	1.54	1.51	1.72	1.29
surface 11	**1.06**	**1.02**	**0.99**	1.20	**0.96**	1.22	1.24	1.12
mean	**1.10**	**1.06**	1.33	1.17	1.23	1.66	1.81	1.10
std	0.10	0.10	0.37	0.23	0.22	0.30	0.57	0.15
min	0.95	0.90	0.99	0.95	0.96	1.22	1.24	0.87
max	1.23	1.18	1.92	1.52	1.54	2.03	2.77	1.29

Table 3
Mean signed error across all OCT layers.

	SEG	SEG + REG	Spectralis	OCTRIMA	AURA	IOWA	Bern	I–O
surface 1	0.90	0.91	−0.82	0.66	1.22	1.99	1.65	0.26
surface 2	−0.12	−0.13	0.76	0.16	0.34	1.47	2.53	0.29
surface 4	0.18	0.18	1.43	0.12	0.41	1.59	1.30	0.29
surface 6	−0.30	−0.29	−0.51	−0.92	−0.51	0.78	1.13	0.09
surface 11	−0.66	−0.65	−0.44	−0.94	−0.58	1.04	0.90	−0.69

Table 4
Mean unsigned error for all retinal layers and macular regions.

	SEG	SEG + REG	Spectralis	OCTRIMA	AURA	IOWA	Bern	I–O
surface1 fovea	1.18	1.14	0.90	0.90	0.90	2.14	1.67	0.85
surface1 parafovea	1.12	1.10	1.14	1.00	1.31	1.98	1.81	0.89
surface1 perifovea	1.12	1.10	1.13	0.92	1.62	2.01	1.62	0.86
surface2 fovea	1.34	**1.25**	1.39	**1.15**	**1.29**	2.42	2.02	1.31
surface2 parafovea	1.03	0.98	**0.92**	1.03	**0.92**	1.59	2.45	0.97
surface2 perifovea	**1.15**	**1.09**	2.02	1.35	1.42	1.54	3.47	1.22
surface4 fovea	**1.10**	**1.03**	1.30	**1.12**	1.25	1.81	1.44	1.13
surface4 parafovea	**0.91**	**0.88**	1.32	**0.91**	**1.02**	1.67	1.52	1.08
surface4 perifovea	**0.92**	**0.86**	2.82	**1.00**	1.14	1.89	1.76	1.11
surface6 fovea	**1.45**	**1.38**	1.79	2.75	2.58	1.58	1.86	1.50
surface6 parafovea	**1.26**	**1.22**	**1.10**	**1.36**	1.42	1.50	1.74	1.36
surface6 perifovea	**1.08**	**1.04**	**0.99**	**1.08**	1.14	1.49	1.62	1.11
surface11 fovea	**0.92**	**0.87**	**0.81**	**1.02**	**0.88**	**1.08**	1.23	1.12
surface11 parafovea	**1.07**	**1.03**	**0.98**	1.19	**0.95**	1.14	1.16	1.12
surface11 perifovea	**1.11**	**1.07**	**1.07**	1.31	**1.02**	1.38	1.32	1.11
mean	**1.12**	**1.07**	1.31	1.21	1.26	1.68	1.78	1.12
std	0.15	0.14	0.53	0.45	0.43	0.37	0.57	0.18
min	0.91	0.86	0.81	0.90	0.88	1.08	1.16	0.85
max	1.45	1.38	2.82	2.75	2.58	2.42	3.47	1.50

dataset with DME patients. This dataset included 110 annotated SD-OCT B-scans of 512 × 740 images from 10 patients. The segmentation was done on 7 layers with an additional class for the fluid labeling. The study showed improvement over baseline versions of FCNs including U-Nets. Distances in pixels from ground truth for the methods ranged from 0.16 to 0.34 pixels for different zones in the retina, which compared favorably to the other methods. DICE metrics on leave one out experiments performed on 8 patients yielded values ranging from 0.77 (for the fluid segmentation) to 0.99 for several other layers. The study demonstrated good performance on images that had edema. The method in Ref. [47] also used a variation on U-net and reported promising results, with a DICE coefficient of about 0.91 on a large clinical dataset of about 100 patients from the Singapore National Eye Center. That dataset included image artifacts and pathologies. Another approach using FCN architectures was reported in Ref. [33] and used ConvNets and also a U-Net-like architecture to delineate macular edema an obtained DICE of 0.91 with performance on par with humans. When compared to other recent and prior methods using automated segmentation, which have used either classical machine learning or FCNs mostly based on U-Nets, the novelty of our approach is in using FCNs based on DenseNets and in using a post-processing step based on Gaussian Process regression.

The Gaussian Process-based post-processing can complement other methods such as those cited earlier. It also comes equipped with an uncertainty estimate that could prove useful in some settings. For example, it could be advantageous in situations where we need to provide a plausible range of uncertainty to process time-varying patterns in longitudinal clinical data. However, a potential limitation of our post-processing approach is that, by estimating surfaces independently, there is no theoretical guarantee that the resulting collections of surfaces do not intersect. Last, while our study focused on estimating OCT surfaces in 2D images, the method can also be extended to 3D volumetric data (e.g. see Ref. [48]).

While results are promising, there are other future directions along which this study could evolve moving forward, to address possible current limitations. Image quality is one such factor, and is an important factor to consider when assessing performance. The Spectralis SDOCT we used includes an image averaging function and is able to provide averaged images based on the scanning protocol selection. The type of scan used in this study represents a compromise between the time required to obtain the scan, the field of view included, and the density of the A- and B-scans. Our dataset was acquired following the standard clinical protocol used as part of the imaging routine at the clinic, which is defined using an automated real time averaging (ART) of 5X to facilitate a timely flow of patients in our imaging clinic and avoid a prolonged acquisition time that could be more burdensome for the patient. When compared with other aforementioned studies (e.g. Ref. [47] used 48X averaging) our methods used representative images with regard to quality and therefore may be expected to perform equally on clinical-grade images with similar characteristics with regard to disease and artifacts. However, more experiments to validate this are needed and are left for future studies. Future work should also involve the use of datasets with greater sample size and more variations in severity of disease. While there are other datasets of greater size and including more variations and severity (e.g., Ref. [22] has 110 B-scans) in pathologies, the benefit of the dataset used in this study is that it is presently the only one – known to us – that includes annotations and other algorithms' segmentation results. This allows for an apples-to-apples performance comparison with other widely used ML-based methods. This dataset also includes annotations from two humans from which interoperator error can be computed and compared. Finally, we note that the sample size of 50 image scans is adequate for training FCNs considering that individual samples for semantic segmentation are measured in pixels and not in images. This is in line with observations made in other studies (e.g. U-Net) where similar numbers of training annotated images were used [35,48]. It also aligns with reports which discuss the apparent misconception related to the need of a significant training dataset in semantic segmentation [47], a point which could appear at first counterintuitive given that it is a well known fact it takes a much larger number of images in training to perform full image classification.

While the presence of fluid, such as intraretinal fluid in DME or subretinal fluid in neovascular AMD, is perhaps the most commonly encountered pathology on OCT, it is now increasingly recognized that retinal thinning in the absence of fluid is also an important pathologic biomarker in ophthalmic and neurologic diseases. For example, structural OCT changes can be detected in diabetic patients without clinical diabetic retinopathy and can include: thinning within the retinal nerve fiber layer (RNFL) [49], inner plexiform layer (IPL) [50], ganglion cell-inner plexiform layer (GCIPL) [50,51], ganglion cell complex (RNFL + GCIPL) [52], and photoreceptor layer [53,54]. These studies sought to determine which retinal layer was selectively thinned in patients with pre-clinical diabetic retinopathy and arrived at different conclusions, partly because these studies used different OCT segmentation algorithms and analyses, again highlighting the fact that accurate, reproducible segmentation algorithms even in the absence of fluid are of immense clinical and research utility. In sum, future extensions of this work will involve the analysis on more severe cases, including DME, and should involve developing and/or testing additional datasets that are reflective of broader pathologies and permit more comprehensive comparison with other recent methods. This is especially important since the space of deep learning-based approaches is growing rapidly.

5. Conclusion

We proposed a novel method for automated segmentation of OCT, using deep learning, and particularly combining fully convolutional networks with Gaussian processes. Performance evaluation shows that DenseNet-based semantic segmentation, when coupled with regression-based post-processing using GP, can effectively address the automated OCT segmentation problem in specific contexts. We demonstrate that the method is on par with human performance and compares favorably with ML-based segmentation algorithms of record when evaluated on an OCT dataset developed by the University of Miami. The proposed method appears to have as small or smaller mean unsigned error (equal to 1.06), versus errors ranging from 1.17 to 1.81 for other methods, and versus a human error of 1.10.

Conflicts of interest

Some co-authors acknowledge patents for retinal image diagnostics for AMD (P. Burlina and N. Bressler). The authors have otherwise no other conflicts of interest.

Acknowledgement

The support of JHU APL internal research and development is gratefully acknowledged. This work was supported in part by the James P. Gills Professorship, and unrestricted research funds to the Johns Hopkins University School of Medicine Retina Division for Macular Degeneration and Related Diseases Research. We thank William Gray-Roncal and Nathanael Kuo at JHU/APL for useful discussions regarding the use of OCT in MS studies and OCT segmentation.

References

[1] N. Wong TY Cheung, P. Mitchell, Diabetic retinopathy, Lancet 376 (9735) (2010) 124–136.

[2] Neil M. Bressler, Age-related macular degeneration is the leading cause of blindness, J. Am. Med. Assoc. 291 (15) (2004) 1900–1901.

[3] Anat London, Inbal Benhar, Michal Schwartz, The retina as a window to the brain: from eye research to cns disorders, Nat. Rev. Neurol. 9 (1) (2013) 44–53.

[4] P. Burlina, D.E. Freund, B. Dupas, N. Bressler, Automatic screening of age-related macular degeneration and retinal abnormalities, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 3962–3966.

[5] Frank G. Holz, Erich C. Strauss, Steffen Schmitz-Valckenberg, Menno van Lookeren Campagne, Geographic atrophy: clinical features and potential therapeutic approaches, Ophthalmology 121 (5) (2014) 1079–1091.

[6] Freerk G. Venhuizen, Bram van Ginneken, Freekje van Asten, Mark J.J.P. van Grinsven, Sascha Fauser, B. Hoyng Carel, Thomas Theelen, Clara I. Sánchez, Automated staging of age-related macular degeneration using optical coherence tomography, Investig. Ophthalmol. Vis. Sci. 58 (4) (2017) 2318–2328.

[7] Philippe Burlina, David E. Freund, Neil Joshi, Y. Wolfson, Neil M. Bressler, Detection of age-related macular degeneration via deep learning, in: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, IEEE, 2016, pp. 184–188.

[8] David E. Freund, Neil Bressler, Philippe Burlina, Automated detection of drusen in the macula, in: Biomedical Imaging: from Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, IEEE, 2009, pp. 61–64.

[9] Albert K. Feeny, Mongkol Tadarati, David E. Freund, Neil M. Bressler, Philippe Burlina, Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images, Comput. Biol. Med. 65 (2015) 124–136.

[10] Philippe Burlina, Seth Billings, Neil Joshi, Jemima Albayda, Automated diagnosis of myositis from muscle ultrasound: exploring the use of machine learning and deep learning methods, PLoS One 12 (8) (2017), e0184059.

[11] Saurabh Vyas, Amit Banerjee, Philippe Burlina, Estimating physiological skin parameters from hyperspectral signatures, J. Biomed. Opt. 18 (5) (2013), 057008.

[12] Radford Juang, Elliot R. McVeigh, Beatrice Hoffmann, David Yuh, Philippe Burlina, Automatic segmentation of the left-ventricular cavity and atrium in 3d ultrasound using graph cuts and the radial symmetry transform, in: Biomedical Imaging: from Nano to Macro, 2011 IEEE International Symposium on, IEEE, 2011, pp. 606–609.

[13] Jing Tian, Boglarka Varga, Erika Tatrai, Palya Fanni, Gabor Mark Somfai, William E. Smiddy, Delia Cabrera DeBuc, Performance evaluation of automated segmentation software on optical coherence tomography volume data, J. Biophot. 9 (5) (2016) 478–489.

[14] Amit Banerjee, Philippe Burlina, Fady Alajaji, Image segmentation and labeling using the polya urn model, IEEE Trans. Image Process. 8 (9) (1999) 1243–1253.

[15] Delia Cabrera DeBuc, A review of algorithms for segmentation of retinal image data using optical coherence tomography, in: Image Segmentation, InTech, 2011.

[16] Heidelberg Engineering GmbH, Spectralis HRA+OCT User Manual Software, 2014.

[17] K. Lee, M.D. Abramoff, M. Garvin, M. Sonka, The Iowa Reference Algorithms (Retinal Image Analysis Lab, iowa institute for biomedical imaging, IA), 2014.

[18] Andrew Lang, Aaron Carass, Matthew Hauser, Elias S. Sotirchos, Peter A. Calabresi, S Ying Howard, Jerry L. Prince, Retinal layer segmentation of macular OCT images using boundary classification, Biomed. Opt. Express 4 (7) (2013) 1133–1152.

[19] Pascal A. Dufour, Ceklic Lala, Hannan Abdillahi, Simon Schroder, Sandro De Dzanet, Ute Wolf-Schnurrbusch, Jens Kowal, Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints, IEEE Trans. Med. Imaging 32 (3) (2013) 531–543.

[20] Jing Tian, Boglárka Varga, Gábor Márk Somfai, Wen-Hsiang Lee, William E. Smiddy, Delia Cabrera DeBuc, Real-time automatic segmentation of optical coherence tomography volume data of the macular region, PLoS One 10 (8) (2015), e0133908.

[21] A. Breger, M. Ehler, H. Bogunovic, S.M. Waldstein, A.M. Philip, U. Schmidt-Erfurth, B.S. Gerendas, Supervised learning and dimension reduction techniques for quantification of retinal fluid in optical coherence tomography images, Eye 31 (8) (2017) 1212.

[22] J Chiu Stephanie, Michael J. Allingham, S Mettu Priyatham, Scott W. Cousins, Joseph A. Izatt, Sina Farsiu, Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema, Biomed. Opt. Express 6 (4) (2015) 1172–1194.

[23] Philippe M. Burlina, Neil Joshi, Katia D. Pacheco, David E. Freund, Jun Kong, Neil M. Bressler, Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration, JAMA Ophthalmol. 136 (12) (2018) 1359–1366.

[24] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[25] Philippe Burlina, Neil Joshi, Michael Pekala, Katia Pacheco, David E. Freund, Neil M. Bressler, Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks, JAMA Ophtalmol. (2017).

[26] Philippe Burlina, Katia D. Pacheco, Neil Joshi, David E. Freund, Neil M. Bressler, Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis, Comput. Biol. Med. 82 (2017) 80–86.

[27] Phillippe Burlina, Neil Joshi, Katia D. Pacheco, David E. Freund, Jun Kong, Neil M. Bressler, Utility of deep learning methods for referability classification of age-related macular degeneration, JAMA Ophthalmol. 136 (11) (2018) 1305–1307.

[28] SW Ting Daniel, Yong Liu, Philippe Burlina, Xinxing Xu, Neil M. Bressler, Tien Y. Wong, Ai for medical imaging goes deep, Nat. Med. 24 (5) (2018) 539.

[29] Philippe M. Burlina, Neil Joshi, Katia D. Pacheco, TY Alvin Liu, Neil M. Bressler, Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration, JAMA Ophthalmol. 137 (3) (2019) 258–264.

[30] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, D Quang Nguyen, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al., Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, Jama 318 (22) (2017) 2211–2223.

[31] Poplin Ryan, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng, Dale R. Webster, Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning, Nat. Biomed. Eng. 2 (3) (2018) 158.

[32] Cecilia S. Lee, Doug M. Baughman, Aaron Y. Lee, Deep learning is effective for classifying normal versus age-related macular degeneration oct images, Ophthalmol. Retin. 1 (4) (2017) 322–327.

[33] Cecilia S. Lee, Ariel J. Tyring, Nicolaas P. Deruyter, Yue Wu, Ariel Rokem, Aaron Y. Lee, Deep-learning Based, Automated Segmentation of Macular Edema in Optical Coherence Tomography, bioRxiv, 2017, 135640.

[34] Yufan He, Aaron Carass, Yeyi Yun, Can Zhao, Bruno M. Jedynak, Sharon D. Solomon, Shiv Saidha, Peter A. Calabresi, Jerry L. Prince, Towards topological correct segmentation of macular oct from cascaded fcns, in: Fetal, Infant and Ophthalmic Medical Image Analysis, Springer, 2017, pp. 202–209.

[35] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[36] Leyuan Fang, David Cunefare, Chong Wang, Robyn H. Guymer, Shutao Li, Sina Farsiu, Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative amd patients using deep learning and graph search, Biomed. Opt. Express 8 (5) (2017) 2732–2744.

[37] Abhijit Guha Roy, Sailesh Conjeti, Sri Phani Krishna Karri, Debdoot Sheet, Amin Katouzian, Christian Wachinger, Nassir Navab, Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks, Biomed. Opt. Express 8 (8) (2017) 3627–3642.

[38] Sripad Krishna Devalla, Prajwal K. Renukanand, K Sreedhar Bharathwaj, Giridhar Subramanian, Liang Zhang, Shamira Perera, Jean-Martial Mari, Khai Sing Chin, Tin A. Tun, Nicholas G. Strouthidis, et al., Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images, Biomed. Opt. Express 9 (7) (2018) 3244–3265.

[39] Jégou Simon, Michal Drozdzal, David Vazquez, Adriana Romero, Yoshua Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE, 2017, pp. 1175–1183.

[40] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[41] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, Laurens van der Maaten, Densely Connected Convolutional Networks, 2016 arXiv preprint arXiv:1608.06993.

[42] Fariz Rahman, keras-contrib. https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/applications/densenet.py, 2018.

[43] P. Diederik, Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2014, p. 6980. *CoRR*, abs/1412.

[44] Carl Edward Rasmussen, Christopher KI. Williams, Gaussian Processes for Machine Learning, 1, MIT press Cambridge, 2006.

[45] GPy GPy, A Gaussian process framework in python, since 2012, http://github.com/SheffieldML/GPy.

[46] Robert B. Gramacy, Herbert K.H. Lee, Bayesian treed Gaussian process models with an application to computer modeling, J. Am. Stat. Assoc. 103 (483) (2008) 1119–1130.

[47] Sripad Krishna Devalla, Khai Sing Chin, Jean-Martial Mari, Tin A. Tun, Nicholas G. Strouthidis, Tin Aung, Alexandre H. Thiéry, Michaël JA. Girard, A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head, Investig. Ophthalmol. Vis. Sci. 59 (1) (2018) 63–74.

[48] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, Olaf Ronneberger, 3d U-Net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 424–432.

[49] Stela Vujosevic, Edoardo Midena, Retinal layers changes in human preclinical and early clinical diabetic retinopathy support early retinal neuronal and müller cells alterations, J. Diabetes Res. (2013), 2013.

[50] Shu-ting Li, Xiang-ning Wang, Xin-hua Du, Qiang Wu, Comparison of spectral-domain optical coherence tomography for intra-retinal layers thickness measurements between healthy and diabetic eyes among Chinese adults, PLoS One 12 (5) (2017), e0177515.

[51] Jay Chhablani, Apoorva Sharma, Abhilash Goud, Hari Kumar Peguda, Harsha L. Rao, Viquar Unnisa Begum, Giulio Barteselli, Neurodegeneration in type 2 diabetes: evidence from spectral-domain optical coherence tomography, Investig. Ophthalmol. Vis. Sci. 56 (11) (2015) 6333–6338.

[52] Ahmed I. Hegazy, Rasha H. Zedan, Tamer A. Macky, Soheir M. Esmat, Retinal ganglion cell complex changes using spectral domain optical coherence tomography in diabetic patients without retinopathy, Int. J. Ophthalmol. 10 (3) (2017) 427.

[53] Joana Tavares Ferreira, Marta Alves, Arnaldo Dias-Santos, Lívio Costa, Bruno Oliveira Santos, João Paulo Cunha, Ana Luísa Papoila, Luís Abegão Pinto, Retinal neurodegeneration in diabetic patients without diabetic retinopathy, Investig. Ophthalmol. Vis. Sci. 57 (14) (2016) 6455–6460.

[54] Laxmi Gella, Rajiv Raman, Tarun Sharma, Quantitative spectral domain optical coherence tomography thickness parameters in type ii diabetes, Oman J. Ophthalmol. 9 (1) (2016) 32.