J. Biophotonics **9**, No. 5, 478–489 (2016) / **DOI** 10.1002/jbio.201500239

Journal of BIOPHOTONICS

REVIEW ARTICLE

# Performance evaluation of automated segmentation software on optical coherence tomography volume data

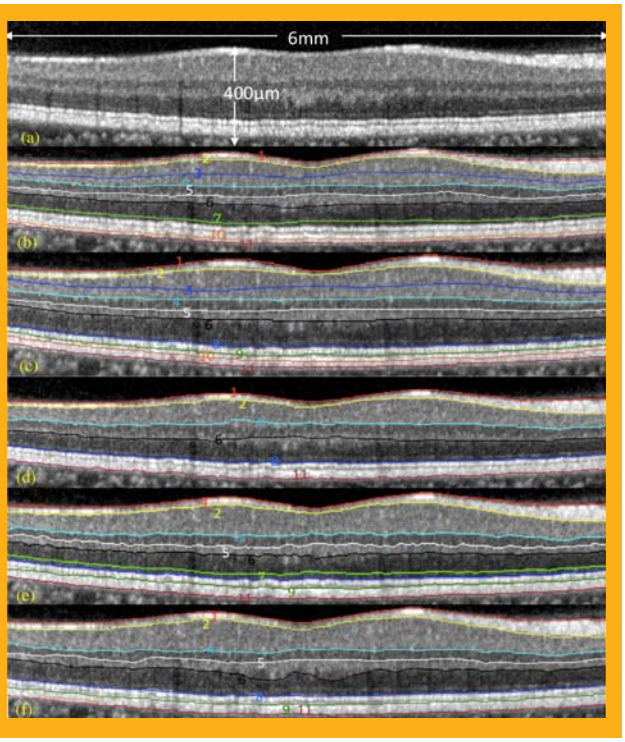
*Jing Tian***, [1], *Boglarka Varga*[2], *Erika Tatrai*[2], *Palya Fanni*[2], *Gabor Mark Somfai*[2], *William E. Smiddy*[1], *and Delia Cabrera Debuc**, **, [1]

[1] Bascom Palmer Eye Institute, University of Miami, 900 NW 17th Street, Miami, FL 33136, United States
[2] Semmelweis University, 39 Maria Street, 1085 Budapest, Hungary

Received 8 September 2015, revised 10 February 2016, accepted 10 February 2016
Published online 12 March 2016

**Key words:** Optical coherence tomography, automated segmentation software, Spectralis SD-OCT, performance evaluation, ground truth

Over the past two decades a significant number of OCT segmentation approaches have been proposed in the literature. Each methodology has been conceived for and/or evaluated using specific datasets that do not reflect the complexities of the majority of widely available retinal features observed in clinical settings. In addition, there does not exist an appropriate OCT dataset with ground truth that reflects the realities of everyday retinal features observed in clinical settings. While the need for unbiased performance evaluation of automated segmentation algorithms is obvious, the validation process of segmentation algorithms have been usually performed by comparing with manual labelings from each study and there has been a lack of common ground truth. Therefore, a performance comparison of different algorithms using the same ground truth has never been performed. This paper reviews research-oriented tools for automated segmentation of the retinal tissue on OCT images. It also evaluates and compares the performance of these software tools with a common ground truth.




* Corresponding author: e-mail: dcabrera2@med.miami.edu
** Authors contributed equally.

## 1. Introduction

Optical coherence tomography (OCT) provides a non-invasive, high-speed and high-resolution approach to visualize the cross-sectional or even three-dimensional tissue structures in vivo. OCT has added significant contributions to many fields of clinical research since its invention in 1991 and since then it has possibly become the most commonly used ophthalmic decision-making technology [1–4]. During the last decade, OCT technology has advanced drastically in terms of speed, resolution and sensitivity and has become a key diagnostic tool in the areas of retinal and optic nerve pathologies [5]. Recent advancements in OCT imaging allow the visualization of retinal structures in a few seconds with an axial resolution of ~2 microns [6–7]. The upgrade of both scanning speed and resolution has significantly increased the potential of OCT to visualize more detailed retinal structures. However, the amount of data to be analyzed has also increased significantly. Automatic analysis algorithms or software are therefore essential to the clinical applications because the huge amount of volumetric data is no longer possible to be analyzed by visual identification or manual labeling.

The OCT segmentation algorithms have evolved with hardware and software improvements within and between manufacturers, frequently with differences in resultant measurements, so the straightforward comparability of measurements is not guaranteed when clinicians and researchers use different generations of OCTs within studies or within individual patients. In addition, major developments of both hardware and software across the years have improved the capabilities of the technology to investigate the multi-layered structure of the retina in more detail. For example, the commercial software of Bioptigen Inc. (Envisu SD-OCT, Bioptigen Inc., Morrisville, North Carolina, USA), Canon (OCT-HS100 SD-OCT, Canon Europe, N.V.), Optovue, Inc. (i-VUE SD-OCT, Optovue, Fremont, California, USA) and Spectralis SD-OCT (Heidelberg Engineering, Germany) could segment 8, 11, 3 and 11 retinal surfaces, respectively [8–11]. Therefore, differences in automatic segmentation algorithm results based on either OCT device or quantitative software selection would have important consequences in clinical practice and research.

As the retina is a multi-layered tissue, it is important to segment the various layers or surfaces in order to fully explore the retinal structure and function. The development of OCT segmentation software has progressed extensively during the last decade. It was originally a proprietary software solution of individual manufacturers of OCT but it has become a generic software solution of various research groups that have developed algorithms to automatically detect retinal surfaces [12–26]. A review of the early methods can be found in [27]. There is also a fully automated research tool that offers commercial software with an independent platform for processing data from different OCT scanners [28].

The earliest image analysis software was mainly developed for Time-Domain OCT, i.e. Stratus OCT (Carl Zeiss Meditec, United States), using features in each A-scans to form smooth boundaries [12–17]. The OCT scan pattern was mainly a single line or radial lines across the fovea. The significant advancement of imaging technology that occurred since then in the commercial devices in terms of scanning speed and resolution has facilitated a dense raster scan of the entire retinal structure. The recently developed OCT segmentation algorithms by research groups are able to segment the retinal surfaces in OCT volume data using the graph-based method [18–22], active contour [23] and texture models [24]. The automatic algorithms were also developed to segment the drusens [25] and retinal fiber layer in optic nerve head region [26]. Our recent work on *OCT Retinal IMage Analysis 3D* (OCTRIMA3D) has been also able to delineate 8 retinal surfaces fast and accurately [29]. Some of these published works even provided standalone software tools developed for Spectralis SD-OCT's volume data that are freely available to be used for research purposes [30–32]. However, the validation process of OCT segmentation algorithms has been usually performed by comparison of segmentation results with manual labelings from each study, and there has been a lack of common ground truth. Of note, the first intent to use a large, manually segmented data set consisting of 466 B-scans from 17 healthy eyes was segmented twice by different operators and compared to the automated segmentation algorithm [24]. However, up to the authors' knowledge, a performance comparison of different algorithms using the same ground truth has never been accomplished.

It is worthwhile to mention that the ground truth forms the foundation for all comparisons with the output of any automatic segmentation method to be evaluated. The complex structure of the retinal tissue and the variety of research tools lately developed points to an in-depth analysis of results that can provide the users with a widespread range of evaluation scenarios and anticipated future needs (as evidenced by current developments). Particularly, Heidelberg Engineering offers a widely available platform with the latest software upgrade being currently available to all users. An *all layer segmentation* function was recently included in the latest software version of Spectralis 6.0, which is able to detect 11 retinal surfaces and measure the thickness of 10 layers [11]. However, only minor information regarding the performance of this proprietary software has been revealed. The Spectralis platform also al-

lows access to the imaging data and therefore publicly available segmentation tools have been mostly developed for the OCT volume data obtained with the Spectralis device [30–32]. This calls for the need to evaluate and compare the performance of all these research tools including ours along with the Spectralis's own segmentation software with a common ground truth. In this context, the aim of this study is to review research oriented tools for automated segmentation of the retinal tissue on OCT images [29–32] including Spectralis 6.0 [11], as well as our own custom-built software [29] and compare their performances using the same set of input data and evaluation criteria. Although other commercial OCT devices offer the capability of standalone software for data review, the evaluations performed were based on the predominant OCT volume data (i.e., Spectralis SD-OCT) used by the publically freely available software. It is worthwhile to mention that outcomes are not presented in terms of better or worse or as more or less accurate. The major intent of this review is to show a number of important issues surrounding the lack of a representative and practical dataset that could be used as a common ground truth for the evaluation of the performance of OCT quantitative methods.

## 2. Materials and method

### 2.1 SD-OCT volume data collection

In this study, the SD-OCT volume dataset used in the performance evaluations was obtained from subjects enrolled in studies involving patients with diabetes mellitus and complications in the eye. This dataset consisted of 610 B-scans with the size of $768 \times 496$ pixels collected from 10 eyes of patients with mild non-proliferative diabetic retinopathy and relatively latent segmentation difficulty. We note that in this paper, the term "surface" refers to a set of pixels that fall on the interface between two layered structures.

### 2.2 Subjects

A total of ten eyes (8 OD, 2 OS) from 10 diabetic subjects (6 male, 4 female) with mild non-proliferative diabetic retinopathy ($53 \pm 6$ years old) were scanned by Spectralis SD-OCT at the Bascom Palmer Eye Institute, Miami, USA. This study was approved by an Institutional Review Board at the University of Miami. Prior to enrollment, the research study was explained to the subject and informed consent was obtained according to the tenets of the Declaration of Helsinki. Patients with any medical condition that might affect visual function other than diabetes were excluded from the study.

### 2.3 Image acquisition and retinal layer classification

The subjects were dilated and then scanned by Spectralis SD-OCT with the IR+OCT protocol using the setting of 30° IR scan angle and $30° \times 25°$ ($8.5 \times 7.1$ mm) OCT pattern. To reduce the speckle noise and enhance the image contrast, every B-scan

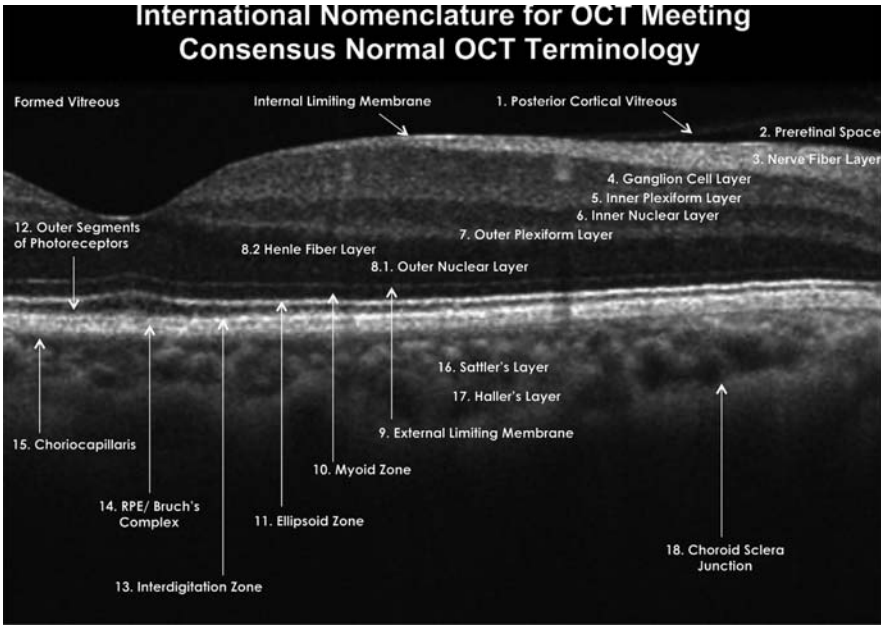



**Figure 1** International Nomenclature for the classification of retinal layers on OCT images. Image was taken from [33] with permission of the author.

**Table 1** Classification of retinal layers used in our study according to the international OCT consensus nomenclature [33].

| Consensus Number | Layer Abbreviations | Layer Full name |
|---|---|---|
| 2 | PRS | Pre-retinal space |
| 3 | NFL | Nerve fiber layer |
| 4 | GCL | Ganglion cell layer |
| 5 | IPL | Inner plexiform layer |
| 6 | INL | Inner nuclear layer |
| 7 | OPL | Outer plexiform layer |
| 8.1 + 8.2 | HFLONL | Henle's Fiber layer and Outer nuclear layer |
| 9 + 10 | ELMMYZ | External Limiting Membrane and Myoid zone of the photoreceptors |
| 11 + 12 | ELZOS | Ellipsoid zone and outer segment of the photoreceptors |
| 13 | IDZ | Interdigitization zone with retinal pigment epithelium |
| 14 | RPE | Retinal pigment epithelium or Bruch's complex |
| 15 | CRC | Choriocapillaris |

was the average of five aligned images using the TruTrack™ active eye tracking technology (ART = 5). The size of the OCT volume collected from the study subjects was $768 \times 61 \times 496$ pixels (width × length × depth) and the resolution of each pixel in the B-scan was 11.11 μm/pixel and 3.87 μm/pixel in the transversal and axial directions, respectively.

To facilitate the communication of anatomy and disease pathophysiology, a consensus nomenclature (as shown in Figure 1) for the classification of retinal layers in Spectralis SD-OCT images developed by an international panel with expertise in retinal imaging [33] was applied throughout the paper and the abbreviations are defined in Table 1. Note that the following anatomical structures as described in [19] were not used in the current study as they are not part of the neuro-retina that is classically the target of image segmentation: 1. Posterior cortical vitreous, 16. Sattler's layer, 17. Haller's layer, and 18. Choroid sclera junction. The transitions from 8.1 Henle's fiber layer to 8.2 Outer nuclear layer, from. 9. External limiting membrane to 10. myoid zone of photoreceptors, and 11. ellipsoid zone to 12. outer segment of photoreceptor are too smooth to be segmented by any existing algorithms, hence they are merged as HFLONL, ELMMYZ, and ELZOS, respectively, in this study.

## 2.4 Image analysis software

In this study, we evaluated the following five automated quantification software for segmentation of macular volume data:

### 2.4.1 Spectralis 6.0

In this study, the Spectralis SD-OCT standalone software for data review is the commercial software used based on the predominant OCT volume data (i.e., Spectralis SD-OCT) employed by the publically freely available software evaluated. The built-in software of Spectralis SD-OCT offers the segmentation of 11 surfaces in the latest software version of 6.0. The users can easily obtain the thickness or volume of three composite retinal segments including *Retina*, *Inner Retinal Layer and Photoreceptors* as well as 7 retinal layers, including *RNFL*, *Ganglion Cell Layer*, *Inner Plexiform Layer, Inner Nuclear Layer, Outer Plexiform layer, Outer Nuclear Layer, Retinal Pigment Epithelium Layer* in each ETDRS grid. However, no technical details have been revealed about this proprietary software in the user manual [11] and the validation of the software was also not found in the literature by a detailed search in PubMed and Google. As the results of surface positions is not provided by Spectralis 6.0 directly, we extracted the layer boundaries by detecting color annotation in the exported video and connected the detected pixels by using Dijkstra's algorithm [34]. An example of boundary extraction is shown in supporting information S1.

### 2.4.2 IOWA reference algorithm

A graph theoretical approach was developed to segment optimal surface in volumetric images [35] and was applied to segment multiple surfaces on the volumetric OCT volumes [20] by Retinal Image Analysis Laboratory in the Iowa Institute for Biomedical Imaging [30]. The standalone software, IOWA Reference Algorithm, could be downloaded free for research use [30]. This algorithm is able to segment eleven surfaces for each OCT volume.

### 2.4.3 AUtomated retinal analysis tools (AURA)

Lang et al. built a random forest classifier to segment 9 retinal surfaces. In this algorithm, the results of a probability map for each surface are used to find the optimal retinal surface by the graph-search method [21]. A set of 27 features covering spatial awareness, local and context-ware information is selected and the manual labelings of 56 B-scans are used to train the classifiers. The author has made both the code of AUtomated Retinal Analysis tools (AURA) and the trained classifier for Spectralis SD-OCT data publicly available [31]. In our study the default setting was used and final set of parameters (FSP) [21] was applied to segment our data.

### 2.4.4 Dufour's algorithm

Another modification of optimal graph search in retinal surface segmentation was provided in the work of Dufour. P [22], which improves the accuracy and robustness of the original framework by using soft constraints to add prior information from a learned model. Six retinal layers could be segmented automatically in both healthy and macular edema subjects. The standalone software could be downloaded at [32] and is abbreviated as Dufour's algorithm in this paper.

### 2.4.5 OCTRIMA3D

Our previous work deployed a shortest-path graph search based on the original development of Chiu et al. [19] for OCT volume data obtained with a Bioptigen device (Bioptigen Inc., Morrisville, North Carolina, USA). OCTRIMA3D is able to segment 8 retinal surfaces in the macular area and its accuracy was found to be comparable with the other automated quantification software [29]. By adding inter-frame flattening, inter-frame search region refinement, masking and biasing the spatial dependency between adjacent frames we were able to reduce the processing time compared to the original study by Chiu et al. [19].

## 2.5 Ground truth description and evaluation criteria

A suitable dataset with ground truth that reflects the realities of everyday OCT imaging in clinical settings is a demanding need for objective performance evaluation of both commercial and research oriented software for segmentation of OCT volume data. However, the creation of an OCT ground truth is not a trivial matter because of the significant cost associated to its creation, as it impacts on both the design and the maintainability of the OCT volume dataset. This cost is due to the circumstance that the construction of an OCT ground truth cannot be fully automated because it is a time-consuming and laborious process that has been reported as a limitation in the majority of previous studies [11, 29–32]). Typical times for creating an OCT ground truth can run in the hours for a single volume dataset. In addition, the effectiveness of the ground truth depends of the following requirements:

1. Richness of information, to facilitate different clinical evaluation settings.
2. Accuracy, both in terms of nonexistence of human errors and in the intrinsic capability to characterize complex information.
3. Ease of design and use, in terms of the capacity to enable the evaluation of large datasets.
4. Ease of understanding, in terms of organization to facilitate use and maintenance.
5. Efficiency of comparison, to allow evaluation using large datasets.
6. Anticipation of future needs, in terms of extensibility to prevent uselessness.

In this study, a ground truth obtained from a pathologic dataset was designed using manual grading from macular OCT volume data that consisted of 50 B-scans with a dimension of $768 \times 496$ pixels and the OCT volume outside the $6 \times 6$ mm$^2$ area around the foveola was discarded. Each volume was represented by 5 systematically selected B-Scan, including 1 from the fovea center, 2 from the perifovea and 2 from the parafoveal regions.

Therefore, a total of 250 (5 surfaces per B-scan) surfaces were manually outlined by each grader in the pathologic dataset. We note that each of the automated quantification software for segmentation of macular volume data has its own advantages and drawbacks and choice has to be made according to the needs and knowledge of users. In addition, some of them have the capability of handling macular edema. However, our ground truth obtained from OCT data collected from subjects with non-proliferative mild diabetic retinopathy is not designed to evaluate the performance when severe pathology is present. Our study is conducted to evaluate the performance of the research-oriented software against the ground truth dataset constructed using the following four criteria:

### 2.5.1 Target surfaces and region of interest

We evaluated the output surfaces provided by the five automated quantification software and estab-

**Table 2** Retinal surface notations and sequence in our study. The abbreviations of the layers are shown in Table 1.

| Surface Sequence | Surface Notation |
|---|---|
| 1 | PRS-NFL |
| 2 | NFL-GCL |
| 3 | GCL-IPL |
| 4 | IPL-INL |
| 5 | INL-OPL |
| 6 | OPL-HFLONL |
| 7 | HFLONL-ELMMYZ |
| 8 | ELMMYZ-ELZOS |
| 9 | ELZOS-IDZ |
| 10 | IDZ-RPE |
| 11 | RPE-CRC |

lished a common notation of the surfaces to provide guidance to the ophthalmic researchers to make a choice. There was a lack of common notations in the various image analysis software tools and hence it may cause confusion to the users. For example, the name of the retinal layers, i.e. RNFL, IPL and OPL were often used to denote the outer surface of layers in Spectralis 6.0 but other software adopted the convention of A–B, where A and B are the adjacent layer structure in the volume OCT (e.g. RNFL-GCL). Surfaces are denoted as A–B, where A and B are the adjacent layers shown in Table 1. To shorten the notation, a unique surface number is defined in Table 2.

Besides the differences in the detected surface, the users need to know the area of interest for each automated segmentation software as well. For example, some software only segment the OCT volume within an area of $6 \times 6$ mm$^2$ around the fovea while others can detect the retinal surfaces in the entire input volume data.

### *2.5.2 Input data format, prerequisite software, and output format of surface location*

The usefulness of the software was limited by the following three obstacles: (a) Input data format: input data in the format of *.vol was only available to limited research groups authorized by Heidelberg Engineering; (b) Prerequisite software: Matlab or C++ compiler may be needed to run the segmentation successfully; (c) The exact locations of each surface are provided in different file formats, which may be difficult to be extracted.

### *2.5.3 Accuracy of surface detection*

The accuracy of the automated software was evaluated using the ground truth datasets. As the line width of the surface location plotted in each B-scan was usually 2–3 pixels wide, the segmentation errors of the automated algorithms were hard to be revealed by visual inspection. Hence, ground truth from manual grading was used to evaluate the accuracy of the five automated segmentation methods. Using a customized tool implemented with Matlab 2014a, two expert manual graders labeled surfaces 1, 2, 4, 6 and 11 in 50 representative B-scans (with a total of 250 surfaces outlined by each grader) collected from a set of 10 macular SD-OCT volumes. Particularly, once the grader clicked on the points along each border, the manual tracing resulting from linear interpolation between the clicked points was taken as the final ground truth for comparison. The grader could also move, add and delete the clicked points to modify the boundary tracings. The labeling task was performed with extreme carefulness by the two observers, Observer 1 and Observer 2 without seeing any segmentation results from any automated software or each other. On average, it took about 20 minutes to label one frame. The delineated results from Observer 1 were taken as the ground truth and the inter-observer difference were used as a benchmark to evaluate the accuracy. The segmentation error of surfaces outlined by the five automated software were compared with inter-observer differences using one-tailed paired *T*-Test with setting the significance level at $p < 0.001$ due to the large number of comparisons. The null hypothesis was that inter-observer difference is significantly higher than the automated software. If the null hypothesis was rejected, the alternative hypothesis "automated method has smaller or equal mean unsigned error" was concluded.

### *2.5.4 Processing time*

Another criteria to consider is the processing time, which is important for the applications in large population studies. The software were all run on a computer with a CPU of Intel® Core™ i7-2600@ 3.4 GHz 3.4 GHz.

## 3. Results

### *3.1 Target surfaces and regions of interest*

The retinal surfaces segmented by five methods are reported in Table 3 and the overlay of the target sur-

**Table 3** Target surfaces and region of interest of five research-oriented quantification software (check mark means the segmented surface result can be exported, cross mark means the surface is not segmented, triangle mark means the surface is segmented but results can not be exported).

| Software Name | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | ROI (mm) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Spectralis 6.0 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Δ | Δ | ✓ | ✓ | Full |
| IOWA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | Full |
| AURA | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ | 6 × 6 |
| Dufour's | ✓ | ✓ | × | ✓ | × | ✓ | × | ✓ | × | × | ✓ | Full |
| OCTRIMA3D | ✓ | ✓ | × | ✓ | ✓ | ✓ | × | ✓ | ✓ | × | ✓ | 6 × 6 |

faces on one sample B-Scan is shown in Figure 2. Although Spectralis 6.0 could segment all 11 surfaces, the surface 8 and 9 could not be extracted and they were not used in the analysis. The IOWA reference algorithm could generate 11 surfaces, but the hyper-reflective surface BMEIS (shown as the dashed line in Figure 2c) was not detected by the other software evaluated and hence it was not considered in this study. AURA, Dufour's algorithm and OCTRIMA3D could segment 9, 6 and 8 surfaces, respectively. Because existing quantitative studies are mostly based on the ETDRS grid, the volume data outside of the $6 \times 6\ mm^2$ area is not considered by AURA and OCTRIMA 3D.

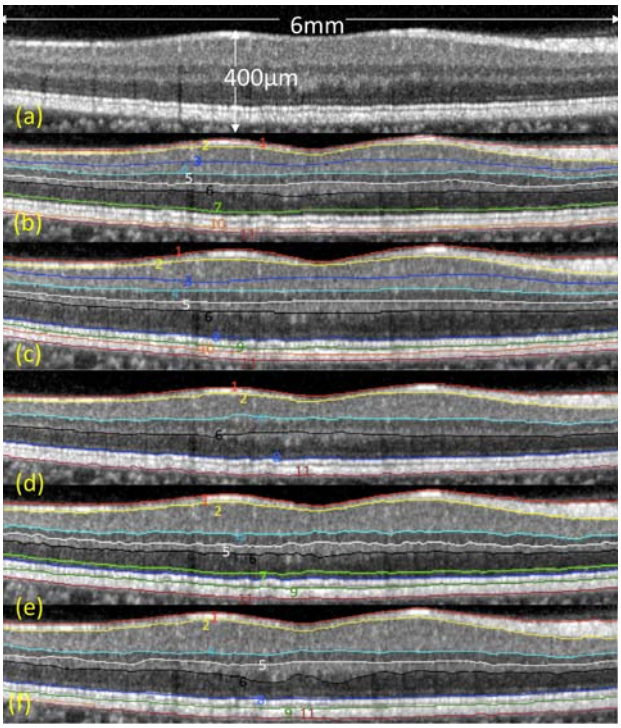



**Figure 2** Targeted retinal surfaces in automated segmentation software. The colored numbers represents the surface sequence shown in Table 2. (**a**) An example raw Spectralis SD-OCT image, which is 369 μm inferior to the foveola. A parafoveal scan is chosen in order to show all the surfaces segmented. The retinal surfaces in automated segmentation software are shown from (**b**)–(**f**). (**b**) Spectralis 6.0 (**c**) IOWA Reference algorithm (**d**) AURA (**e**) Dufour's algorithm (**f**) OCTRIMA3D. Note that the IOWA reference algorithm could generate 11 surfaces, but the hyper-reflective surface BMEIS (shown as the green dashed line in (**c**)) was not detected by the other software and hence it was not evaluated in this study.

### *3.2 Input data format, prerequisite software, and output format of surface location*

The review of the input data format, prerequisite software, output format of surface location and other modality support is shown in Table 4 and discussed briefly as following.

- **Input data format:** The segmentation tool of Spectralis 6.0 was integrated with the database and the surface segmentation of the whole volume could be performed by just one click without the need of providing input data. Other software required the input of volumetric data in *.vol or *.xml format. The export of data in *.vol is a special raw format provided by Heidelberg Engineering to the collaborating institutes and not obtainable to the public. Standard built-in software supports *.xml format and is hence easier to use for the clinicians. In this study, we converted the volume in the *.xml format using the template provided in [36].
- **Prerequisite software:** The prerequisite of AURA is the Matlab and C++ compiler, which required additional license fee to use the software. The OCTRIMA3D application needed the Matlab Compiler Runtime (MCR), which was freely downloadable from [37]. The other software tools could be run as standalone tools on any computer running on Windows operating system.
- **Output format of surface locations:** The output of surface locations was not provided by Spectralis 6.0. The other software provided the output in different formats. AURA and OCTRIMA3D records the location of surfaces using

**Table 4** Review of input data format, prerequisite software, output data format

| Algorithm | Input | Prerequisite | Output |
|---|---|---|---|
| Spectralis 6.0 | – | No | Not provided |
| IOWA | *.vol | No | Surface.xml |
| AURA | *.vol | Matlab and C++ compiler | *.mat |
| Dufour's | *.xml | No | *.raw and *.csv |
| OCTRIMA3D | *.xml | MCR | *.mat |

*.mat format, where the indices ($x$, $y$, $i$) in the output matrix denoted transversal coordinate, frame number, and surface indices, respectively. The format *.raw and .csv represents the surface location as a one-dimensional array, which represents the location of each surface as a raster scan. The surface.xml file records the surface locations using tabs ⟨z⟩ in each surface, but the extraction of the value is hard to handle due to the large size.

### 3.3 Accuracy of surface detection

The results of surface location errors from five automated algorithms were compared with inter-observer differences on surfaces, 1, 2, 4, 6 and 11. The mean unsigned errors in the scans from center foveal, perifoveal and parafoveal regions are presented in Tables 5 and the histogram is shown in Figure 3. The systematic error was calculated as the average signed error in all scans. The positive and negative values of signed errors in the tables indicated that the detected locations were below or above the ground truth, respectively. The segmentation error of various surfaces showed the following trends:

The unsigned segmentation errors on the surface 1 by OCTRIMA3D were smaller than the inter-observer difference ($p < 0.001$) in all regions. AURA had smaller deviation from the ground truth than the inter-observer difference ($p < 0.001$) in the center fovea region. The surface 2 (NFL-GCL) was better delineated by Spectralis, AURA and OCTRIMA in the center fovea region ($p < 0.001$). Both AURA and OCTRIMA 3D detected the surface 4: IPL-INL more reliable than the manual labeling in all regions ($p < 0.001$). The unsigned error of Spectralis and AURA in the perifoveal and parafoveal locations was smaller ($p < 0.001$) than the inter-observer difference on surface 6: OPL-HFLONL. Moreover, RPE-CRC surfaces in all scans were delineated more accurately in Spectralis, AURA and OCTRIMA. The IOWA reference algorithm and Dufour's algorithm tend to have larger systematic error than Spectralis, AURA and OCTRIMA3D. Hence the mean unsigned error was not significantly smaller than the inter-observer differences ($P < 0.001$). It is worth noting that the performance comparison does

**Table 5** Mean unsigned errors (pixels) of five automated segmentation algorithms as compared to the inter-observer differences obtained for the pathological dataset. The values in bold indicate that the errors were smaller than or equal to the inter-observer difference.

| Surface | | Spectralis | IOWA | Dufour | AURA | OCTRIMA | Inter-Observer |
|---|---|---|---|---|---|---|---|
| 1 | Fovea | 1.07 | 1.76 | 1.32 | **0.73** | **0.75** | 0.86 |
| | Perifovea | 1.28 | 1.7 | 1.42 | 0.94 | **0.77** | 0.95 |
| | Parafovea | 1.25 | 1.8 | 1.52 | 1.35 | **0.8** | 0.83 |
| | Signed error | 1.1 | –1.72 | –1.36 | –0.4 | –0.36 | 0.28 |
| 2 | Fovea | **1.11** | 2 | 1.85 | **1.11** | **1.16** | 1.24 |
| | Perifovea | 1.64 | 1.42 | 2.63 | 1.01 | 1.19 | 1.05 |
| | Parafovea | 1.22 | 1.53 | 3.25 | 1.17 | 1.04 | 1.11 |
| | *Signed error* | *–0.46* | *–1.21* | *–2.42* | *–0.19* | *0.42* | *0.2* |
| 4 | Fovea | 1.14 | 1.68 | 1.45 | **1.01** | **1.07** | 1.1 |
| | Perifovea | 1.88 | 1.36 | 1.21 | **0.87** | **0.9** | 1.13 |
| | Parafovea | 1.66 | 1.64 | 1.63 | **0.92** | **0.93** | 1.08 |
| | *Signed error* | *–1.11* | *–1.28* | *–1.06* | *–0.01* | *0.27* | *0.3* |
| 6 | Fovea | 1.55 | 1.56 | 1.75 | 2 | 2.45 | 1.42 |
| | Perifovea | **1** | 1.53 | 1.5 | **1.06** | 1.32 | 1.22 |
| | Parafovea | **0.99** | 1.39 | 1.61 | **1.03** | 1.24 | 1.19 |
| | *Signed error* | *0.54* | *–0.73* | *1.1* | *1.68* | *2* | *0.08* |
| 11 | Fovea | **0.87** | 1.71 | 1.77 | **0.75** | **0.87** | 1.08 |
| | Perifovea | **0.83** | 1.7 | 1.6 | **0.76** | **0.91** | 1.14 |
| | Parafovea | **0.95** | 1.84 | 1.69 | **0.86** | **0.84** | 1.08 |
| | *Signed error* | *–0.22* | *–1.73* | *–1.55* | *–0.14* | *0.06* | *–0.65* |

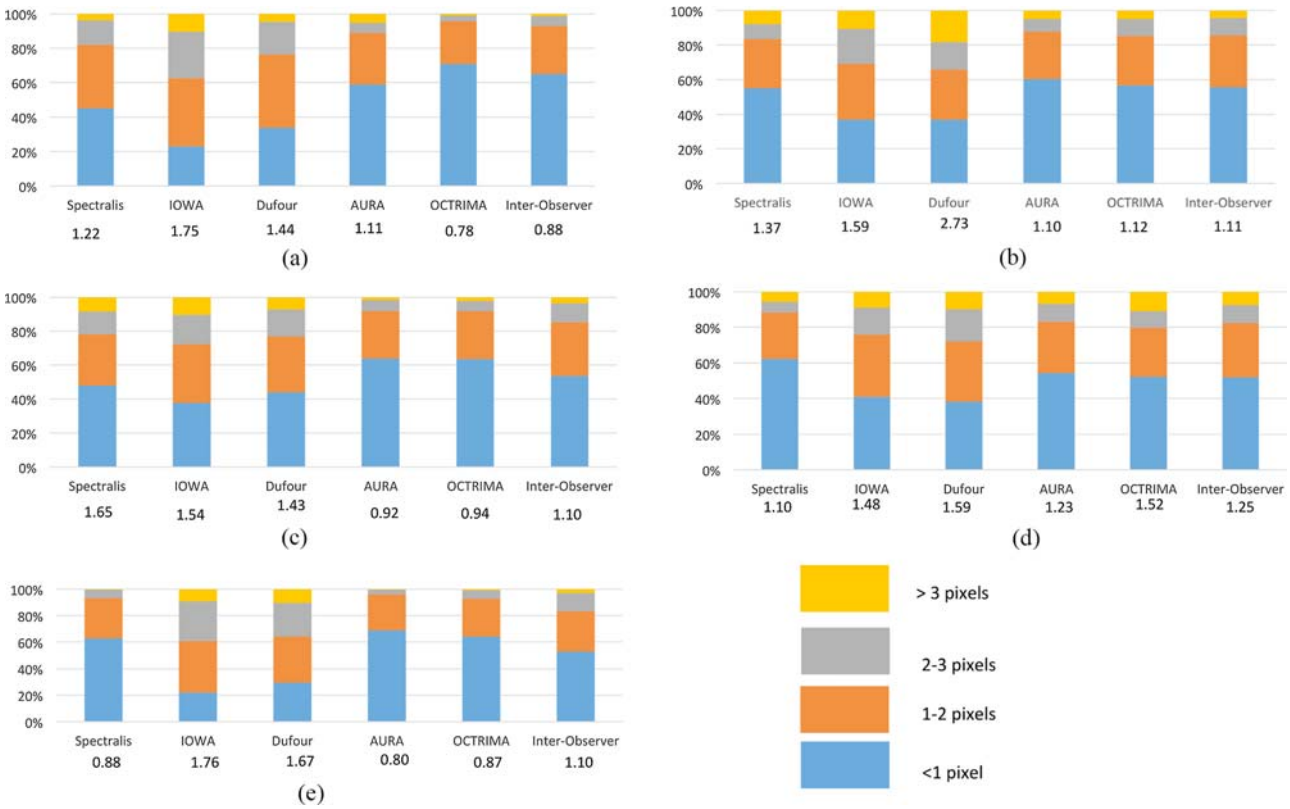



**Figure 3** Histogram of surface detection errors from 5 automatic algorithms as compared to inter-observer differences. The average unsigned errors are shown below each histogram. (**a**)–(**e**) surface 1, 2, 4, 6, and 11.

not imply the quality of the algorithm. Rather, the deviation should be interpreted as the ground truth difference from each research group. In general, automated algorithms use specific optimum settings as well as training ground truth datasets. Smaller unsigned error indicates the training ground truth is closer to our testing ground truth created by the Observer 1. As illustrated in Figure 4, the definition of OPL-HFL-ONL is vague in certain regions. The results from each automated software are a consequence of the difference in the delineated boundaries during each particular software training stage.

### *3.4 Processing time of the five automated software as compared to manual labeling*

The processing time of Spectralis 6.0, the IOWA reference algorithm, AURA, Dufour's algorithm and

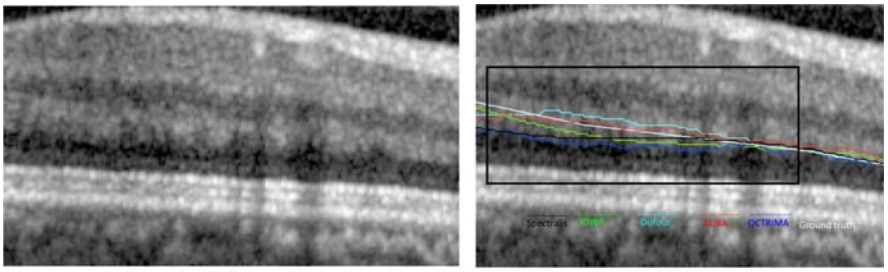



**Figure 4** Effect of training ground truth choices on surface detection: (**a**) A sample OCT raw image section. (**b**) The detection results of the surface 6 by five automatic software and ground truth obtained for the sample image shown in (**a**). The area with major discrepancy is enclosed with a black rectangle box. The difference in the surface detection is due to the training ground truth used by different groups. A common ground truth is necessary to have comparable results in the measurement of retinal layer thickness and topography.

OCTRIMA 3D were 65, 93, 152, 75 and 28 seconds, respectively. For manual labeling, each frame was labeled in about 20 minutes and thus the whole dataset required about 16.7 hours of work from each expert.

## 4. Discussion

This study evaluated a total of five research-oriented software for the automated segmentation of OCT volume data in the macular region. The advancement of OCT imaging technology has allowed the high-speed, non-invasive and high-resolution visualization of three-dimensional retinal structure in vivo but the processing of the volumetric data is still challenging. Manual labeling of the entire volume data is infeasible; therefore automated segmentation of the OCT volume data plays an essential role in exploiting the diagnostic capability of OCT imaging. Various research groups have developed automated segmentation algorithms but there has been a lack of evaluation of the different automated solutions using the same ground truth. Our work evaluated five standalone automated software tools that are able to segment Spectralis SD-OCT images taken at the macular region. The evaluation was performed using an OCT volume dataset collected from eyes of subjects with diabetes and non-proliferative mild diabetic retinopathy. A common surface terminology was established by using the international nomenclature consensus developed recently for structures revealed on OCT images. The evaluation was conducted by using the following three criteria: (a) detected surfaces and area of interest; (b) input data format, prerequisite software and output surface location format; (c) surface detection accuracy (d) processing time. Two experienced graders labeled 5 retinal surfaces in representative images of the OCT volume dataset carefully using a custom-built program and the inter-observer difference was used to benchmark the accuracy of software. Linear interpolation was used to draw the boundaries in each B-scans and each boundary was interpolated from 30–40 manually clicked points. On average, each observer took about 16.7 hours to label the dataset.

Spectralis 6.0 provides the segmentation of 11 surfaces in the entire volume but results for the surfaces 8 and 9 could not be exported. The segmentation module is incorporated into the built-on software and hence no input is needed from the users. However, the standard version of the software does not allow the export of surface locations. The mean unsigned error was significantly smaller than the inter-observer differences in surface 2: NFL-GCL (center foveal region), surface 6 (perifoveal and parafoveal regions) and surface 11 (all regions) ($P < 0.001$). The segmentation process was completed in 65 seconds.

IOWA reference algorithm is able to detect 11 retinal surfaces in the entire volume dataset. The input data requires *.vol format and the output of the surfaces location is written in the surface_xml file. The average signed errors and unsigned errors in the detected surface were larger than the inter-observer difference. This is most probably caused by the disagreement between the manual labeling from the training set in IOWA and our ground truth. A bias correction step may be needed to correct the disagreement. Additional feature of the software was the support of wide range of OCT devices, including Cirrus HD-OCT, Bioptigen and Topcon. The segmentation process was completed in 93 seconds.

As a standalone software that supports the standard output function from Spectralis 6.0, the Dufour's algorithm is able to segment six retinal surfaces in the entire volume scan. It has the capability to also segment the volume data from patients with macular edema. The mean unsigned errors were significantly greater than the inter-observer differences in all surfaces. Again, the disagreement is mainly due to the difference in the training data. The segmentation process was completed in 75 seconds.

AURA is able to segment 8 surfaces in the area $6 \times 6\,mm^2$ around the fovea. It requires the input data in the *.vol format, Matlab and C++ compiler to run the program. The mean unsigned error was smaller than or equal to the inter-observer differences in the surfaces 1 (center fovea), 2 (center foveal location), 4 (all regions), 6 (perifoveal and parafoveal regions) and 11 (all regions). The segmentation process was completed 152 seconds. OCTRIMA3D is able to segment 8 retinal surfaces in the $6{\times}6mm^2$ region of the macula. It requires the installation of the compatible MCR to run the program and also supports the standard Spectralis output export format. The unsigned errors of surface 1 (all region), 2 (center foveal location), 4 (all regions), 11 (all regions) were smaller than or equal to the inter-observer differences. The whole segmentation process was finished in 28 seconds.

The development of automated segmentation software is essential in exploiting the diagnostic capability of optical coherence tomography. The clinical segmentation reality of common pathologies could vary across retinal regions and diseases. Therefore, the segmentation accuracy of the retinal structure is critical for the proper assessment of retinal pathology and current treatment practice. However, the optimal automated segmentation software for OCT volume data remains to be established. This study reveals that the built-in software from Spectralis as well as the research-oriented software evaluated could achieve accuracy close to the inter-observer

differences. It also revealed that most of the automated segmentation software failed to segment some retinal surfaces mostly in regions where a confident unbiased decision by humans was also hard to obtain. However, these results were obtained with the use of an OCT volume dataset collected from 10 diabetic subjects; which is not the actual representative ground truth needed in OCT imaging. A previous study used for the first time a large manually segmented dataset (466 B-scans) but it was obtained from healthy subjects (17 eyes) using an OCT custom-built device and the Amazon Mechanical Turk (AMT) machine [24]. This human intelligence task required the supervision of the manual segmentation relatively often because workers performing the segmentation tasks were not retinal specialists. While the need for unbiased performance evaluation of automated segmentation algorithms is obvious, there does not exist a suitable dataset with ground truth that reflects the realities of everyday retinal features observed in clinical settings (e.g. pathologic cases which contain discontinues surfaces and additional abnormalities disrupting the retinal structure).

In addition to the lack of a common ground truth in OCT imaging, minor information of the retinal tissue from the OCT volume data is commonly revealed besides the thickness of retinal layers [38–41]. Recent advances in OCT technology are adding the capability to extract information on blood flow and perfusion status of the retinal tissue as well as on changes in the polarization state of the probing light beam when interacting with the retinal tissue [42]. Therefore, it is expected that a more complete characterization of the retinal tissue could potentiate the diagnostic capability of the OCT technology. However, more active communication and multi-disciplinary collaboration between research groups would be valuable. One way is to create an OCT volume dataset annotated by expert graders. In this context, the integration of efficient proofreading and editing tools is of relevance when using automated analysis techniques. This dataset needs to be large and representative of both healthy eyes and pathological cases. It should also be available to the public with the hope that developers and researchers will use it to evaluate and verify quantitative algorithms for efficiency, effectiveness, robustness and reliability. Such dataset is crucial for the development of surface detection software, especially for the training based algorithms. This will facilitate procedures to be developed using realistically challenging OCT data, make it feasible to compare algorithms quantitatively by running them on the same dataset, and speed biomarker identification by providing clinicians and industry with metrics for comparing algorithm performance and clinical assessment of therapeutic treatments. It will also assist with OCT technology development by highlighting areas of strength and weakness of current developments. The dataset and ground truth used in this work are available in the supporting information S2. Of note, another potential solution to speed up the standardization of OCT data analysis for clinical use would be a free and open-source software initiative for optimization and wider use of OCT segmentation software with an independent platform for processing data from different OCT scanners [43].

This paper has introduced and discussed a number of important issues surrounding ground truth for the evaluation of the performance of automated segmentation software of OCT volume data. The focus was on the review and comparison of five research-oriented software with a common ground truth. The reported performance differences of the software may depend on the particular OCT device used, differences in the training data used by each software and their effectiveness to segment pathological features of the diseased retina. Therefore, further investigation should be considered as the field evolves. In summary, the use of OCT technology in clinical settings is of great value, but reliable data analysis and proper diagnosis of the various retinal diseases requires a joint effort to create a large and representative repository of OCT information with free access to help advance the judgment and decision making processes of OCT developments and clinical applications.

## Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website.

Video S1: The example of exported video from Spectralis. The annotated color lines were extracted to form the retinal surfaces.

Data S2: The OCT dataset, segmentation results from five software tools and the manual labeling from two observers.

**Acknowledgements** This study was supported in part by a NIH Grant No. NIH R01EY020607, a NIH Center Grant No. P30-EY014801, by an unrestricted grant to the University of Miami from Research to Prevent Blindness, Inc., and by an Eotvos Scholarship of the Hungarian Scholarship Fund. Thanks to Sandra Pineda, B.S. for her assistance with the recruitment of healthy subjects and clinical coordination.

**Author biographies** Please see Supporting Information online.

## References

[1] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fujimoto, Science **80**, 1178–1181 (1991).
[2] A. M. Zysk, F. T. Nguyen, A. L. Oldenburg, D. L. Marks, and S. A. Boppart, J. Biomed. Opt. **12**, 051403 (2007).
[3] R. Hamdan, R. G. Gonzalez, S. Ghostine, and C. Caussin, Archives of Cardiovascular Diseases **105**, 529–534 (2012).
[4] C. A. Puliafito, Ophthalmic Surg. Lasers Imaging **41**, S5 (2010).
[5] J. S. Schuman, C. A. Puliafito, J. G. Fujimoto, and S. D. Jay, Optical Coherence Tomography of Ocular Diseases, 3rd ed. Thorofare: Slack, Inc, 2004.
[6] W. Drexler, Journal of Biomedical Optics **9**, 47 (2004).
[7] A. F. Fercher, Z. Med. Phys. **20**, 251–276 (2010).
[8] Biopigen Inc, Envisu C-Class SDOCT System | Bioptigen, Inc. [Online]. Available: http://www.bioptigen.com/products/c-class/ [Accessed: 06-Sep-2015].
[9] Canon, Canon OCT-HS100 – Eye Care – Canon Europe, 02-Jan-2012. [Online]. Available: http://www.canon-europe.com/medical/eye_care/oct-hs100/ [Accessed: 06-Sep-2015].
[10] Optovue Inc, iVUE SD-OCT. [Online]. Available: http://optovue.com/wp-content/uploads/2013/08/iVue-Brochure.pdf [Accessed: 06-Sep-2015].
[11] Heidelberg Engineering GmbH, Spectralis HRA +OCT User Manual Software Version 6.0, 2014.
[12] D. Koozekanani, K. Boyer, and C. Roberts, IEEE Trans. Med. Imaging **20**, 900–916 (2001).
[13] D. Cabrera Fernández, H. M. Salinas, and C. A. Puliafito, Opt. Express **13**, 10200–10216 (2005).
[14] T. Fabritius, S. Makita, M. Miura, R. Myllylä, and Y. Yasuno, Opt. Express **17**, 15659–15669 (2009).
[15] M. Shahidi, Z. Wang, and R. Zelkha, Am. J. Ophthalmol. **139**, 1056–1061 (2005).
[16] H. Ishikawa, D. M. Stein, G. Wollstein, S. Beaton, J. G. Fujimoto, and J. S. Schuman, Invest. Ophthalmol. Vis. Sci. **46**, 2012–2017 (2005).
[17] G. Gregori and R. W. Knighton, Invest. Ophthalmol. Vis. Sci. **45**, 3007 (2004).
[18] M. D. Abràmoff, M. K. Garvin, and M. Sonka, IEEE Rev. Biomed. Eng. **3**, 169–208 (2010).
[19] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, Opt. Express **18**, 19413–19428 (2010).
[20] M. K. Garvin, M. D. Abràmoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, IEEE Trans. Med. Imaging **28**, 1436–1444 (2009).
[21] A. Lang, A. Carass, M. Hauser, E. S. Sotirchos, P. A. Calabresi, H. S. Ying, and J. L. Prince, Biomed. Opt. Express **4**, 1133–1152 (2013).
[22] P. A. Dufour, L. Ceklic, H. Abdillahi, S. Schroder, S. De Dzanet, U. Wolf-Schnurrbusch, and J. Kowal, IEEE Trans. Med. Imaging **32**, 531–543 (2013).
[23] A. Yazdanpanah, G. Hamarneh, B. R. Smith, and M. V. Sarunic, IEEE Trans. Med. Imaging **30**, 484–496 (2011).
[24] V. Kajić, B. Povazay, B. Hermann, B. Hofer, D. Marshall, P. L. Rosin, and W. Drexler, Opt. Express **18**, 14730–14744 (2010).
[25] Q. Chen, T. Leng, L. Zheng, L. Kutzscher, J. Ma, L. de Sisternes, and D. L. Rubin, Med. Image Anal. **17**, 1058–1072 (2013).
[26] M. A. Mayer, J. Hornegger, C. Y. Mardin, and R. P. Tornow, Biomed. Opt. Express **1**, 1358–1383 (2010).
[27] D. C. DeBuc, A review of algorithms for segmentation of retinal image data using optical coherence tomography (2011).
[28] J. D. Oakley, I. Gabilondo, C. Songster, D. Russakoff, A. Green, and P. Villoslada, Invest. Ophthalmol. Vis. Sci. **55**, 4790 (2014).
[29] J. Tian, B. Varga, G. M. Somfai, W.-H. Lee, W. E. Smiddy, and Delia Cabrera DeBuc, PlosOne **8**, 0133908 (2015).
[30] K. Lee, M. D. Abramoff, M. Garvin, and M. Sonka, The Iowa Reference Algorithms (Retinal Image Analysis Lab, Iowa Institute for Biomedical Imaging, Iowa City, IA). [Online]. Available: https://www.iibi.uiowa.edu/content/iowa-reference-algorithms-human-and-murine-oct-retinal-layer-analysis-and-display.
[31] A. Lang, NITRC: AURA tools : AUtomated Retinal Analysis tools: Tool/Resource Info, 2015. [Online]. Available: http://www.nitrc.org/projects/aura_tools/ [Accessed: 22-Jun-2015].
[32] P. A. Dufour, OCT Segmentation Application. [Online]. Available: http://pascaldufour.net/Research/software_data.html [Accessed: 22-Jun-2015].
[33] G. Staurenghi, S. Sadda, U. Chakravarthy, and R. F. Spaide, Ophthalmology **121**, 1572–1578 (2014).
[34] E. W. Dijkstra, Numer. Math. **1**, 269–271 (1959).
[35] K. Li, X. Wu, D. Z. Chen, and M. Sonka, IEEE Trans. Pattern Anal. Mach. Intell. **28**, 119–134 (2006).
[36] M. Mayer, Read heidelberg engineering (he) oct raw files., 2011. [Online]. Available: http://www5.informatik.unierlangen.de/fileadmin/Persons/MayerMarkus/openVol.m URL.
[37] MathWorks, MATLAB Runtime – MATLAB Compiler. [Online]. Available: http://www.mathworks.com/products/compiler/mcr/?refresh=true [Accessed: 22-Jun-2015].
[38] D. DeBuc, E. Tatrai, L. Laurik, B. E. Varga, V. Olvedy, A. Somogyi, W. E. Smiddy, and G. M. Somfai, J. Clin. Exp. Ophthalmol. **4**, 289 (2013).
[39] Z. Hu, M. Nittala, and S. Sadda, Invest. Ophthalmol. Vis. Sci. **54**, 5492 (2013).
[40] H. Chen, X. Chen, Z. Qiu, D. Xiang, W. Chen, F. Shi, J. Zheng, W. Zhu, and M. Sonka, Sci. Rep. **5**, 9269 (2015).
[41] K. A. Vermeer, J. van der Schoot, H. G. Lemij, and J. F. de Boer, Invest. Ophthalmol. Vis. Sci. **53**, 6102–6108 (2012).
[42] T. E. de Carlo, A. Romano, N. K. Waheed, and J. S. Duker, Int. J. Retin. Vitr. **1**, 5 (2015).
[43] M. Tiemann, History of the OSI | Open Source Initiative, Sciencecommons.org., 2010. [Online]. Available: http://opensource.org/history [Accessed: 07-Sep-2015].