Published in final edited form as:

Optom Vis Sci. 2011 January ; 88(1): 113–123. doi:10.1097/OPX.0b013e3181fc3625.

Reliability of a Computer-Aided Manual Procedure for Segmenting Optical Coherence Tomography Scans

Donald C. Hood, PhD, Jung Suk Cho, Ali S. Raza, BA, Elizabeth A. Dale, BA, and Min Wang, MD, PhD

Department of Psychology, Columbia University, New York, New York (DCH, JC, ASR, EAD), and Department of Ophthalmology and Visual Science, Eye and ENT Hospital, Fudan University, Shanghai, China (MW)

Abstract

Purpose—To assess the within- and between-operator agreement of a computer-aided manual segmentation procedure for frequency-domain optical coherence tomography scans.

Methods—Four individuals (segmenters) used a computer-aided manual procedure to mark the borders defining the layers analyzed in glaucoma studies. After training, they segmented two sets of scans, an Assessment Set and a Test Set. Each set had scans from 10 patients with glaucoma and 10 healthy controls. Based on an analysis of the Assessment Set, a set of guidelines was written. The Test Set was segmented twice with a ≥ 1 month separation. Various measures were used to compare test and retest (within-segmenter) variability and between-segmenter variability including concordance correlations between layer borders and the mean across scans ($n = 20$) of the mean of absolute differences between local border locations of individual scans, $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$.

Results—Within-segmenter reliability was good. The mean concordance correlations values for an individual segmenter and a particular border ranged from 0.999 ± 0.000 to 0.978 ± 0.084 . The $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values ranged from 1.6 to $4.7 \mu\text{m}$ depending on border and segmenter. Similarly, between-segmenter agreement was good. The mean concordance correlations values for an individual segmenter and a particular border ranged from 0.999 ± 0.001 to 0.992 ± 0.023 . The $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values ranged from 1.9 to $4.0 \mu\text{m}$ depending on border and segmenter. The signed and unsigned average positions were considerably smaller than the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values for both within- and between-segmenter comparisons. Measures of within-segmenter variability were only slightly larger than those of between-segmenter variability.

Conclusions—When human segmenters are trained, the within- and between-segmenter reliability of manual border segmentation is quite good. When expressed as a percentage of retinal layer thickness, the results suggest that manual segmentation provides a reliable measure of the thickness of layers typically measured in studies of glaucoma.

Keywords

optical coherence topography; OCT; segmentation; glaucoma

For more than 15 years, the thickness of the human retinal nerve fiber layer (RNFL) has been routinely measured with time-domain optical coherence tomography (OCT).^{1,2} With the newer frequency-domain (fd) OCT, other retinal layers can be easily discerned and

measured as well. Studies of glaucoma typically focus on layers of the inner retina—the RNFL, the retinal ganglion cell (RGC) layer, the combined thickness of the RNFL, RGC and inner plexiform layers (IPL)—and, in some cases, total retinal thickness as well. A variety of computer algorithms have been developed for segmenting two or more of these layers. Some of these algorithms are commercially available, included with particular fdOCT machines, whereas others are reserved for the use of individual research groups. However, different algorithms can produce different results. For example, we provided evidence that segmentation algorithms, rather than hardware, accounted for differences between the RNFL thickness measured with an fdOCT machine vs. that measured with time-domain OCT machine.³ Because there is no generally available and accepted algorithm for segmenting different retinal layers, computer-aided manual segmentation procedures have been used by a few groups, e.g., as shown in Refs. 4–9. For example, we have used a computer-aided manual procedure to measure the thickness of the RGC plus IPL in patients with glaucoma.⁸

In addition to the obvious need to assess the reproducibility of these procedures, there are three other reasons to be concerned with validating manual procedures. First, there is no “gold standard” for assessing the results of segmentation algorithms; thus, it is hard to compare the relative performance of different algorithms. Manual segmentation provides a possible vehicle for validation.^{5,6} Although, we do not mean to imply that the visually determined borders are the “gold standard,” it is true that many errors of automated algorithms can be detected visually. Second, a systematic attempt to visualize borders supplies information about the problems automated algorithms might confront, as will be illustrated below. Third, even if automated procedures become reliable enough for general use, it is very likely that they will include an option for the user to “correct” the segmented borders. In fact, both commercial and non-commercial programs presently have this capability. This also raises the question of the reliability of manual segmentations. That is, how consistent are different operators in how they “correct” the segmentation?

The purpose of this study was to assess the within- and between-segmenter agreement of a computer-aided manual procedure for segmenting fdOCT scans. The procedures here were designed to study the effects of glaucoma. Thus, we concentrate on the layers that were most vulnerable to this disease. We trained four individuals, whom we call “segmenters,” to mark the borders between these layers in fdOCT line scans of the horizontal meridian. After training, we asked How good is the within-segmenter reliability (i.e., repeatability)? How good is the agreement between segmenters? And, what have we learned about the problems confronting computerized algorithms?

METHODS

Subjects

Segmenters—Four individuals (the “segmenters”), two undergraduates, one medical student, and one ophthalmologist, segmented fdOCT scans. Segmenters C and D, who had segmented more than 90 scans before this experiment, served as advisers to segmenters A and B as part of a training period.

Patients/Controls—All four segmenters segmented the same 40 scans obtained from 40 individuals, 20 glaucoma patients and 20 controls. They were divided into two groups, Assessment Set and Test Set, with each set containing scans from 10 patients and 10 controls. Our purpose here was not to compare the results from controls vs. patients. A mixture of patients and controls were included to assure a range of ages and scanning conditions; the patients were scanned as part of a clinical visit. Inclusion criteria included a best-corrected visual acuity of 20/60 or better, spherical refraction within ± 6.00 diopters (D),

and cylinder correction within ± 2.00 D. Individuals with clinically significant cataracts, other media opacities, or other ocular diseases were excluded. In general, the patients had early to moderate field losses, average MD of -5.50 dB (range, 0.65 to -12.31) for the Assessment Set and average MD of -3.38 (range, 1 to -10.39) for the Test Set. Written informed consent was obtained from all participants. Procedures followed the tenets of the Declaration of Helsinki and the protocol was approved by the Committee of the Institutional Board of Research Associates of Columbia University.

The Scans

Forty eyes of 40 individuals had fdOCT scans of the horizontal meridian (3D-OCT 1000, Topcon Corporation, Tokyo, Japan), which consisted of an average of 16 overlapping B-scans (1024 A-scans) 6 mm in length. The A-scan depth was 1.68 mm and the resolution was $3.5 \mu\text{m}/\text{pixel}$.

Design

The Assessment Set was used to determine if more training/experience was needed. After the Assessment Set was completed by the four segmenters, the data were analyzed as described below (Data Analysis). Although the agreement among segmenters was good, the analysis of these data revealed consistent differences in decisions made during segmentation. On the basis of this information, a set of guidelines were written to help improve consistency across segmenters. In the interest of journal space, only the results for the Test Set are presented below.

Following the creation of the guidelines, the Test Set was completed by all four segmenters. Between-segmenter reliability was assessed by comparing the results on the Test Set. To assess within-segmenter reliability, this set was segmented again after an interval that was ≥ 1.0 month (range, 1.0 to 2.2 months). In the interim, all segmenters were segmenting other scans as part of other studies. No feedback was given about the Test Set segmentations until all segmenters had completed the retest.

The Computer-Aided Manual Segmentation Procedure

As previously described,^{7,8} the raw scan data were exported from the fdOCT machine and displayed by the segmentation program. Fig. 1A shows a sample 6-mm line scan of a control eye. The operator manually marked the boundary of a retinal layer by using the “mouse” and a “click” to place points along the boundary. The program, written in MATLAB (v7.4, Mathworks, Natick, MA), drew a curve through these points using a spline algorithm. Points could be added and/or changed until the operator was satisfied that the boundary was described by the smooth curve.

Four borders were marked (Fig. 1A): (1) the border between the vitreous and the RNFL (red); (2) the border between the RNFL and the RGC layer (orange); (3) the border between the IPL and the inner nuclear layer (INL) (green); and (4) the border between retinal pigment epithelium (RPE)/Bruch’s membrane (BM) and the choroid (violet). By using these borders, the thicknesses of four layers, which have been used in studies of glaucoma, were calculated. In particular, the RNFL thickness (red vertical line) was calculated as the difference between borders 1 (vitreous/RNFL border) and 2 (RNFL/RGC border). Similarly, the RGC layer plus IPL (RGC + IPL; orange line) was calculated as $(2 - 3)$ and is shown as the orange vertical line. We measured the combined thickness, RGC + IPL, because on most scans, the border between RGC and IPL could not be visualized. The thickness of the total retina (white line) was calculated as $(1 - 4)$. In addition, we included what has been called the ganglion cell complex (GCC) because it has been used in commercial machines and in

other studies. The GCC thickness (green line) is the sum of the RNFL and RGC + IPL and is calculated as $(1 - 3)$.

Data Analysis

For both the within- and between-segmenter analyses, the unsigned (absolute value) and signed differences were analyzed as described below. The results for the patients and controls were similar, presumably because the patients had relatively minor field defects, especially in the regions covered by the scans. Thus, the patient and control results were combined in the interest of brevity. Because we are dealing with signed and unsigned differences, four segmenters, 20 scans, four borders, and four thicknesses for each scan, and two testing sessions, the analysis can seem complicated. For clarity, the following abbreviations are used.

MBL: mean border location of a particular border on a single scan (of a particular subject) segmented by a single segmenter. By convention, the location of any point of a border is the distance from the top of the scan in millimeters. The absolute values here do not provide any useful information. Thus we use,

Δ MBL: the difference between the MBL on test and retest (within-segmenter analysis) or between a segmenter and the average of the other three segmenters (between-segmenter analysis).

Δ MBL: the absolute (unsigned) difference between MBL on test and retest or between a segmenter and the average of the other three segmenters.

MEAN(Δ MBL): mean of Δ MBL across all 20 scans/subjects.

MEAN(Δ MBL): mean of Δ MBL across all 20 scans/subjects.

The MBL underestimates local differences in a border location. To assess local perturbations in the segmentation of borders we also define.

LBL: local border position of a particular location on a particular scan segmented by a single segmenter.

Δ LBL: absolute difference between LBL on test and retest or between a segmenter and the average of the other three segmenters.

mean(Δ LBL): mean of Δ LBL across the length of a particular scan.

MEAN{mean(Δ LBL)}: mean of mean(Δ LBL) values across all 20 scans/subjects.

Note: to conserve space we did not report the signed Δ LBL, “mean(Δ LBL),” or MEAN{mean(Δ LBL)}.

Concordance correlations were also calculated. These correlations were nearly always very high and, in general, added little to our understanding. Thus, they are only briefly summarized below.

RESULTS

Within-Segmenter Variability

Fig. 1A shows a typical scan from one of the patients. The dashed and solid lines are the borders segmented by one segmenter, D, on two occasions separated in time by a little over

1 month. The lines in Fig. 1B are the locations of these borders plotted as a function of distance from the center of the fovea. (Note that the y axis is the distance from the top of the scan in millimeters.) The test and retest borders are shown as the dashed and solid lines, respectively. Although borders for the test and retest segmentations are similar, there are deviations. To put the deviations in perspective, the error bar is 20 μm . We assessed the agreement in a number of ways. (See Methods for a definition of the abbreviations used.)

First, consider the MBL for this particular scan and this segmenter. As an illustration, the MBL of the four borders in Fig. 1B were 250.6 (test) and 249.1 (retest) μm (Vitreous/RNFL), 265.4 and 265.7 μm (RNFL/RGC), 337.4 and 336.4 μm (IPL/INL), and 556.0 and 554.3 μm (RPE/BM). To obtain a measure of the within-segmenter variability of the MBL for segmenter D, the signed (ΔMBL) and unsigned (absolute value) (ΔMBL) differences between the MBL of the test and retest borders were computed for each of the four borders and each of the 20 scans. For segmenter D, the mean of the 20 signed, $\text{MEAN}(\Delta\text{MBL})$, and unsigned, $\text{MEAN}(\Delta\text{MBL})$, border locations are shown in the next to the last row in Tables 1 (signed) and 2 (unsigned). A negative number in Table 1 indicates that the retest border was below the test border on average for the 20 scans. All but one of the values in Table 1 were within 1.2 μm of 0. Although this indicates that on average, there is little or no consistent difference in border placements between test and retest, it does not provide a good measure of variability. For example, for any given segmenter, if MBLs were $-3 \mu\text{m}$ for one scan and $+3 \mu\text{m}$ for another, the overall mean would be 0 μm . Although the standard deviations in Table 1 provide a measure of variability, the unsigned difference analysis in Table 2 has been used as an easy-to-interpret measure of border accuracy.^{5,6} The mean unsigned differences were $\leq 2.5 \mu\text{m}$, with 12 of the 26 cells in Table 2 showing values $\leq 1.8 \mu\text{m}$. In general, segmenter A showed the best agreement between mean locations calculated on test and retest segmentations; for each border, both her signed and unsigned differences are within 1 μm of 0.

MBLs do not provide an indication of the degree of variability in local border locations. Because the local thickness of the individual layers is of interest to us, a measure of local border variation was also obtained. The black curves in Fig. 1C show the ΔLBL , i.e., the difference between the test and retest border locations in Fig. 1B. The other 19 curves in each panel show the difference profiles for the other 19 scans, i.e., the scans for the other 9 patients and 10 controls. The unsigned mean, $\text{mean}(\Delta\text{LBL})$, of the points in each of the 20 curves in each panel of Fig. 1C was obtained and the mean of these 20 values, $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$, is shown in Table 3. For example, for segmenter D and the vitreous/RNFL border, the $\text{mean}(\Delta\text{LBL})$ for each of the 20 curves in Fig. 1C (red curves) was obtained and then averaged to get the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ value of 1.9 μm in Table 3. For segmenter D, these $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values ranged from 1.9 to 3.3 μm depending on the border. Note this segmenter's values are larger than segmenter A's and within 0.6 μm of the mean of all four segmenters. Overall, the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values varied from 1.6 μm (segmenter A and vitreous/RNFL border) to 4.7 μm (segmenter C and RPE/BM border). On average, the least variability was seen in the vitreous/RNFL border, typically the most distinct border in most scans.

In addition, the agreement between border locations was assessed with concordance correlations. For the within-segmenter comparisons, the mean correlations for an individual segmenter and a particular border ranged from 0.999 ± 0.000 to 0.978 ± 0.084 .

Between-Segmenter Variability

Fig. 2A shows the same scan as in Fig. 1A with the segmented borders from all four segmenters; the dashed curves are the same as in Fig. 1A for segmenter D. The lines in Fig. 2B are the locations of these borders plotted as a function of distance from the center of the

fovea for all four segmenters. To obtain a measure of between-segmenter variability, three Tables analogous to Tables 1 to 3 were generated based on their first segmentation of the Test Set. As an illustration, consider the MBL of the vitreous/RNFL border represented in Fig. 2A, B. The MBL of the vitreous/RNFL border was 249.9, 249.3, 250.6, and 250.6 μm for segmenters A, B, C, and D, respectively. So for this single scan, the signed, ΔMBL , and unsigned, ΔMBL , difference between segmenter D and the mean of the other three segmenters was 0.7 and 0.7 μm , respectively. The averages across all 20 scans are shown as the signed, $\text{MEAN}(\Delta\text{MBL})$, and unsigned, $\text{MEAN}(\Delta\text{MBL})$, border locations in Tables 4 and 5, respectively. For the vitreous/RNFL border and segmenter D, the $\text{MEAN}(\Delta\text{MBL})$ and $\text{MEAN}(\Delta\text{MBL})$, were -0.2 and 0.7 μm , respectively. The $\text{MEAN}(\Delta\text{MBL})$ in Table 4 provides valuable information, as discussed below, about systematic differences between segmenters. A negative number in Table 4 indicates that segmenter D's boundary was below the average border location of the other three segmenters. Most of the $\text{MEAN}(\Delta\text{MBL})$ values in Table 4 were within 1 μm of 0, indicating that there was little or no systematic difference across subjects between border location of a particular segmenter and the average location of the other three segmenters. However, the three values shown as bold for segmenters B (RNFL/RGC) and C (RNFL/RGC and IPL/INL) suggest that there were systematic differences on the order of 2 μm in these cases.

Similar to Table 2, Table 5 shows the $\text{MEAN}(\Delta\text{MBL})$; these values were typically below 2 μm , except for the same three cells as in Table 3, as well as the RPE/BM values for segmenter C. The reason for these larger values will be discussed below. In general, segmenters A and D show the closest agreement with the mean of the other three segmenters.

To compare the local border locations (LBL) across segmenters, a similar analysis to that shown in Fig. 1C and Table 3 was performed. Fig. 2C shows the point-by-point difference between segmenter D's border location and the MBL for the other three segmenters for the scan in Fig. 2A (black), as well as for the other 19 scans (colored). The unsigned mean of the points, $\text{mean}(\Delta\text{LBL})$, in each of the 20 curves in each panel of Fig. 2C was obtained and the mean of these 20 values, $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$, is shown in Table 6. For example, for segmenter D and the vitreous/RNFL border, the $\text{mean}(\Delta\text{LBL})$ for each of the 20 curves in Fig. 2C (red curves) was obtained and then averaged to get the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ value of 2.0 μm in Table 6. For segmenter D, these means ranged from 2.0 to 2.7 μm depending on the layer and were larger than those of segmenter A, but, in general, smaller than those of segmenters B and C. Overall, the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values varied from 1.9 μm (segmenter A: vitreous/RNFL border) to about 4 μm (segmenter B: RNFL/RGC border; segmenter C: RNFL/RGC and IPL/INL border). Again, on average, the least variability was seen in the vitreous/RNFL border.

In addition, the agreement between border locations was assessed with concordance correlations. For the between-segmenter comparisons for a particular border, the mean correlation ranged from 0.999 ± 0.001 to 0.992 ± 0.023 .

A Comparison of Within- and Between-Segmenter Variability

In general, the between-segmenter variability was surprising close to the within (retest) segmenter variability. For example, compare the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ values in the bottom rows of Tables 3 and 6. Of course, for within-segmenter reliability, we compared the segmentation of a test scan to a single replication, the retest, whereas for the between-segmenter reliability, we compared the same test scan to the mean of three segmentations, effectively reducing measurement error. To make the comparison fairer, the between-segmenter analyses were repeated with pair-wise comparisons and the $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ results are in Table 7. The averages of these pair-wise comparisons,

bottom row in Table 7, are now consistently larger than the corresponding values for the within-segmenter comparisons in Table 3. However, the within- and between-segmenter values are still relatively close, within 1 μm , with the between-segmenter variability 9 to 30% larger depending on the border.

Layer Thickness

In general, borders are segmented so that the local or average thickness can be calculated. Fig. 1A, shows the typical retinal layer thicknesses of interest in glaucoma. The curves in Figs. 1D, 2D are the thickness profiles of these layers based on the borders segmented in panel A of these figures. The same tables generated above were generated for layer thickness. In the interest of space, only Tables 8 and 9, equivalent to Tables 1 and 4 are shown. As expected, the variation is larger here as two borders are involved in determining each thickness value. However, the differences are relatively small.

Lessons Learned During Training

During the training sessions, we discovered a number of points of confusion that led to between-segmenter differences in border locations. To resolve these confusions we developed guidelines. Recall that these guidelines were available to the segmenters before and during segmentation of the test and retest sessions reported here. As many of the suggestions can be generalized to other devices and segmentation methods, a few of the key points of the guidelines are worth mentioning.

We found that segmenters can differ in where they place a border even when the border is relatively distinct. As an example, take the edge of the vitreous/RNFL border shown in Fig. 3A, B. A segmenter may choose to follow the upper edge of the white pixels as in Fig. 3A, including some black regions. However, the segmenter could seek to avoid including any black regions below the border as in Fig. 3B. Because our segmenters showed consistent differences during training, our guidelines explicitly instructed them to avoid including black regions as in Fig. 3B. Note that automated algorithms will also differ in where they place borders depending on the contrast threshold and the amount of spatial averaging involved.

Second, initially segmenters were confused about how to deal with small local perturbations in a border. The guidelines tell the segmenter to ignore small local perturbations in a border, similar to an algorithm with some local averaging.

Third, there were also “artifacts” that segmenters learned to ignore. Horizontal scans often show a light reflex both in the fovea and just outside the fovea. Segmenters were instructed not include the apparent thickening that is associated with these artifacts (red arrow in Fig. 3C). Similarly, what we take to be a vitreal separations (Fig. 3D) should be ignored. Although computer algorithms typically exclude the later, the light reflexes can create a problem if the apparent thickening is not local.

Finally, the horizontal scans that include the fovea present two additional problems for both computer algorithms and human segmenters, problems that are not present to the same extent a few degrees above or below the midline. The first involves the RNFL on the temporal side of the fovea. The horizontal meridian temporal to the fovea is the locus of the Raphe. It should include few, if any, RGC axons. However, on the scans with the best resolution, it is often possible to see a very thin RNFL. This is probably largely because of glial structures and not to residual RGC axons, as we have seen this thin layer in patients with a total loss of vision because of glaucoma or ischemic optic neuropathy. In any case, the point here is that this layer is easy to discern in some scans and difficult or impossible to see in others. This created confusion for our segmenters and accounted for a significant part

of the variation seen in the RNFL/RGC border. The largest numbers in Tables 4 to 6 are associated with this border and segmenters B and C. Note in Table 4, on average, the border for segmenter B was 2.4 μm below the mean of the others, and for segmenter C, it was 2.1 μm above the mean. In the case of segmenter B, it can be traced to the tendency for this segmenter to be more likely than the other segmenters to identify an RNFL layer on the temporal side. Fig. 3E, F illustrates this point with one of the scans segmented by segmenter B (panel E) and C (panel F). To test this hypothesis, we recalculated the mean RNFL/RGC border omitting the points on the nasal side of the fovea. The values of RNFL/RGC border for segmenter B now fell in the range of values for the other segmenters. However, segmenter C had a tendency to place both the RNFL/GC and IPL/INL borders slightly higher than the other segmenters. Although automated algorithms will segment the same scan the same way every time, they too will differ across scans depending on how visible this residual RNFL is. (Note: algorithms that take 3D information into consideration may have an advantage in dealing with this problem, as well as with the problem with the light reflex mentioned above.)

There is a second problem unique to segmentation in and near the fovea. If the scan transverses the center of the fovea, then in most individuals, there should be no RNFL, RGC layer, IPL, or INL in the very center. Although this may create a problem for computer algorithms, human segmenters were simply instructed by the guidelines to extrapolate the layers seen outside the fovea to the same point in the center of the fovea. That is, segmenters were instructed to mark borders in the center of the scan if they saw them. However, if, as was typically the case, the layer was missing in the center, then they were to extrapolate the border in question (e.g., RNFL/RGC) to the same point on the vitreous/RNFL border in the center to the fovea. This was relatively easy to do, as Fig. 3H illustrates for the scan in Fig. 3G.

DISCUSSION

Manual segmentation will be increasingly important because studies seek to quantify changes in retinal layers seen on fdOCT scans. First, until we have automated algorithms that are generally available, some laboratories will measure the thickness of retinal layers with human segmenters and manual procedures. Second, many of the automated algorithms that currently exist allow for manual segmentation to “correct” the borders segmented by the algorithms; it is likely that future algorithms will as well. In addition, the results from manual segmentations can help inform the development and validation of automated algorithms. With these considerations in mind, we compared the within- and between-segmenter agreement of four individuals who manually segmented fdOCT scans.

In particular, four individuals were trained to segment fdOCT scans of the horizontal meridian using a computer-aided manual procedure. These individuals segmented the same 20 scans twice with the segmentations separated by at least 1 month. First, the retest reliability was good with mean concordance correlations for the each of the four segmenters exceeding 0.978. The $\text{MEAN}(\Delta\text{MBL})$ (Table 1) was typically within $\pm 1 \mu\text{m}$ (range, -0.7 to $2.2 \mu\text{m}$) and the $\text{MEAN}(\Delta\text{MBL})$ of the test and retest borders only slightly higher (range, 0.5 to $2.5 \mu\text{m}$). We also calculated $\text{mean}(\Delta\text{LBL})$, the mean of the absolute value of the point-by-point difference between the border on the test and retest segmentations for each scan. As would be expected, the mean of these values across the 20 scans, $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$, was larger, ranging from 1.6 to $4.7 \mu\text{m}$ depending on the segmenter and border. It also appears that not all segmenters are equally reliable. For example, segmenter A had the smallest retest values for all borders. However, this study was not designed to determine the degree to which variations in time between repeat tests and degree of experience may have

influenced these differences because segmenter A had the shortest time between test and retest.

Similarly, the agreement among segmenters was also good. The between-segmenter values for $\text{MEAN}(\Delta\text{MBL})$, $\text{MEAN}(\Delta\text{MBL})$, and $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$ based on pair-wise comparisons, were, on average, slightly higher than the within-segmenter values.

As a caveat, it should be mentioned that we used reasonably high-quality averaged scans. Scans of poorer quality and/or scans from patients with aberrant retinal anatomy because of extreme pathology present more serious challenges for the human segmenter. The reliability measures for such scans will undoubtedly be worse. Of course, these same scans will be more difficult for automated algorithms as well.

How Good Are These Results?

As we have no gold standard for border segmentation, there is no simple way to answer this question. Therefore, in assessing these results, we must consider the question, “How good compared to what?” Obviously, computer algorithms will show perfect retest reliability when segmenting the same scan twice. So in that sense, the human retest reliability will always be poorer than that of any particular algorithm. Comparing the results of different computer algorithms to our intersegmenter comparisons is more problematic. Because of the proprietary nature of current algorithms, it is not yet feasible to compare the performance of different algorithms on the same scan. If it were possible, we predict that the disagreement among algorithms would be greater than the disagreement among our segmenters. We base this prediction on our observation of the performance of current commercial algorithms. In any case, although it is not our purpose here to compare manual segmentation to those of computer algorithms, all computer algorithms we have seen make “mistakes” that a human segmenter would not make. Of course, the reverse can also be true, although in our experience considerably less frequently.

A second approach is to compare our results to the literature. Unfortunately, we know of only two studies of a quantitative evaluation of manual segmentation. Garvin et al.^{5,6} compared the results of manual segmenters to each other, and to those of an automated segmentation algorithm. They calculated $\text{MEAN}(\Delta\text{LBL})$ for a number of borders including what we call vitreous/RNFL, RNFL/RGC, and RGC/INL. The mean pair-wise values for our four segmenters taken two at a time were 2.8, 3.9 and 3.7 μm for these three borders (Table 7). These values were considerably smaller than the comparable values of 4.9, 6.7, and 8.3 μm for the three segmenters in Garvin et al.⁵ or the values of 3.5, 5.5, and 6.7 μm for the three segmenters in Garvin et al.⁶ Two factors probably contribute our lower values. On one hand, our scans probably had better resolution than those in Garvin et al.,⁵ because we used fdOCT scans and they used time-domain OCT scans. However, our training procedure probably also contributed to the better agreement seen in this study because Garvin et al.⁶ also used fdOCT scans.

A third approach is to compare the results to the resolution of the scan. The scans had a vertical resolution of 3.5 $\mu\text{m}/\text{pixel}$, whereas the data here were typically better. However, the segmenters marked the borders on enlarged images on a monitor. Although the scan resolution is still a factor, the greater resolution of the monitor display (about 1.5 μm) combined with averaging information across adjacent pixels allows the segmenters to achieve a higher effective precision than the limit imposed by the OCT scan resolution. At their best, the results here are closer to the monitor’s resolution and are probably within a factor of two of the 1.5 μm limit.

Finally, we can consider our results in the context of what we seek to measure. Ultimately, we are interested in the thickness of various layers of the inner retina. Fig. 1A shows four possible layers of interest to those studying glaucomatous damage. For the 20 scans in this study, the mean thickness of these four layers was as follows: 13.4 μm (RNFL), 67.7 μm (RGC+IPL), 81.1 μm (GCC), and 296.2 μm (total retina). To get an estimate of the relative size of the intersegmenter variability, the mean SD in Table 9 (bottom row) was divided by these mean thickness values. These SDs were 12.9%, 3.0%, 2.0%, and <1% of the RNFL, RGC + IPL, GCC, and total retina mean thickness, respectively. Given the SD values in Table 9, these values would be smaller for some individuals and larger for others. They will be relatively similar for within-segmenter/retest reliability (Table 8). With the possible exception of the RNFL thickness, the variability is relatively small relative to the mean thickness. However, the value of 12.9% is an overestimate because the variability of the RNFL thickness includes the fovea and temporal retina where there is no RNFL. This percentage would be considerably smaller in regions of higher RNFL thickness. In any case, if one is interested in RNFL thickness, the peripapillary region is probably a better place to make these measurements.

Implications for Automated Segmentation

Those developing automated algorithms undoubtedly confront some of the same problems confronting our segmenters. First, artifacts, such as the light reflex, vitreous separation or epiretinal membranes, can lead to an apparent local thickening if not identified and discounted. Second, the thin RNFL in and around the horizontal Raphe, as well as the inner retinal layers in the fovea, will create problems for automated algorithms. Use of 3-D information (i.e., scans above and below the horizontal meridian) should help here. Finally, a local perturbation in a border and/or a loss of border clarity will also create a problem for computerized algorithms. The degree of local spatial smoothing/filtering will influence how the algorithm treats these discontinuities.³

Caveat

To simplify the presentation, we combined the results for the controls and the patients because we found little difference between the two groups in within- or between-segmenter reliability. For example, consider the MEAN(Δ MBL) values in Table 2. The equivalent values for the controls and patients for each cell differed between -0.6 (patients values more discrepant) and $+0.5$ (control values more discrepant) with a mean difference of $0.11 \mu\text{m}$. However, our patients had very mild glaucomatous damage. Thus, we cannot be sure that the reliability would not be poorer with severely affected patients, where there is almost no RNFL in large regions of the retina. However, we find that although mild to moderate glaucomatous damage causes thinning of the RNFL and RGC layer, it does not dramatically change the appearance of the structure.

Improvements in Manual Segmentations

We have learned that individuals segmenting fdOCT scans start with different criteria regarding how to define a border and make different decisions about ambiguous situations. These differences can be decreased by training. In addition, we believe that written guidelines are important in training and maintaining consistent criteria among, and within, individual segmenters. In our experience, it is hard to obtain consistent agreement among segmenters without written guidelines. In addition, we found that two groups working on different projects came to a different consensus about their informal guidelines, e.g., how to segment the RNFL along the horizontal raphe. Thus, the use of written guidelines cannot be overemphasized.

However, we should be clear that we cannot quantify the effect of the guidelines in this study. The guidelines were developed during the training period because it became clear that they needed both to codify rules among segmenters and to remind individual segmenters about these rules. However, the comparison with the Garvin et al.⁵ results mentioned above, supplies quantitative support for our anecdotal evidence for the value of these guidelines.

However, even with training and with guidelines, differences will exist among segmenters and within segmenters over time. These differences should be taken into consideration when designing studies. For example, the same segmenters should be used when comparisons are made across conditions. Further, to minimize variability, one could “re-calibrate” segmenters by repeating segmentations performed at an earlier time.

Finally, given the importance of manual segmentation, there is a need for further work. Among the questions to be answered: To what extent is variability among segmenters reduced by training? Is training needed for very experienced OCT viewers, for example, those who were developing the algorithms?

SUMMARY

When human segmenters are trained, the intra- and intersegmenter reliability of manual border segmentation is quite good. When expressed as a percentage of retinal layer thickness, the results suggest that manual segmentation provides a reliable measure of the layers typically measured in studies of glaucoma.

Acknowledgments

We thank Drs. De Moraes, Liebmann, and Ritch for referring the patients for fdOCT scanning; Dr. Xian Zhang for technical assistance with the program, and Margot Lazow, Dr. Moura, and Jean Kim for serving as segmenters and helpful colleagues.

This work was supported by National Eye Institute, National Institute of Health grant EY02115. The fdOCT machine was on loan from Topcon, Inc.

References

1. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA, Fujimoto JG. Optical coherence tomography. *Science*. 1991; 254:1178–81. [PubMed: 1957169]
2. Schuman JS, Hee MR, Puliafito CA, Wong C, Pedut-Kloizman T, Lin CP, Hertzmark E, Izatt JA, Swanson EA, Fujimoto JG. Quantification of nerve fiber layer thickness in normal and glaucomatous eyes using optical coherence tomography. *Arch Ophthalmol*. 1995; 113:586–96. [PubMed: 7748128]
3. Hood DC, Raza AS, Kay KY, Sandler SF, Xin D, Ritch R, Liebmann JM. A comparison of retinal nerve fiber layer (RNFL) thickness obtained with frequency and time domain optical coherence tomography (OCT). *Opt Express*. 2009; 17:3997–4003. [PubMed: 19259241]
4. Lim JJ, Tan O, Fawzi AA, Hopkins JJ, Gil-Flamer JH, Huang D. A pilot study of Fourier-domain optical coherence tomography of retinal dystrophy patients. *Am J Ophthalmol*. 2008; 146:417–26. [PubMed: 18635153]
5. Garvin MK, Abramoff MD, Kardon R, Russell SR, Wu X, Sonka M. Intraretinal layer segmentation of macular optical coherence tomography images using optimal 3-D graph search. *IEEE Trans Med Imaging*. 2008; 27:1495–505. [PubMed: 18815101]
6. Garvin MK, Abramoff MD, Wu X, Russell SR, Burns TL, Sonka M. Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *IEEE Trans Med Imaging*. 2009; 28:1436–47. [PubMed: 19278927]

7. Hood DC, Lin CE, Lazow MA, Locke KG, Zhang X, Birch DG. Thickness of receptor and post-receptor retinal layers in patients with retinitis pigmentosa measured with frequency-domain optical coherence tomography. *Invest Ophthalmol Vis Sci.* 2009; 50:2328–36. [PubMed: 19011017]
8. Wang M, Hood DC, Cho JS, Ghadiali Q, De Moraes GV, Zhang X, Ritch R, Liebmann JM. Measurement of local retinal ganglion cell layer thickness in patients with glaucoma using frequency-domain optical coherence tomography. *Arch Ophthalmol.* 2009; 127:875–81. [PubMed: 19597108]
9. Horn FK, Mardin CY, Laemmer R, Baleanu D, Juenemann AM, Kruse FE, Tornow RP. Correlation between local glaucomatous visual field defects and loss of nerve fiber layer thickness measured with polarimetry and spectral domain OCT. *Invest Ophthalmol Vis Sci.* 2009; 50:1971–7. [PubMed: 19151389]

FIGURE 1.

(A) Horizontal midline scan of a control subject showing the borders segmented and the layers identified for thickness measurements. The dashed and solid borders are for segmenter D's test and retest segmentations, respectively. (B) The border location as a function of the distance from the center of the fovea for test (dashed) and retest (solid) segmentations. (C) The black curves are the difference between the dashed and solid curves in panel B. The other colored curves are for a similar analysis of the other 19 scans. (D) The layer thickness as a function of the distance from the center of the fovea for test (dashed) and retest (solid) segmentations.

FIGURE 2.

(A) Horizontal midline scan of a control subject showing the borders segmented by segmenter D (dashed) and by the other three segmenters (solid). (B) The border location as a function of the distance from the center of the fovea for segmenter D (dashed) and the other three segmenters (solid). (C) The black curves are the difference between the dashed and the average of the three solid curves in panel B. The other colored curves are for a similar analysis of the other 19 scans. (D) The layer thickness as a function of the distance from the center of the fovea for segmenter D (dashed) and the other three segmenters (solid).

FIGURE 3.

(A) Segmentation of the vitreous/RNFL (red) border in which the upper edge of the white pixels are approximately followed. (B) The recommended segmentation of the scan from panel A showing the recommended procedure in which the inclusion of black regions is avoided. (C) An illustration of a segmentation of the vitreous/RNFL (red) border that ignores the white region due to a light reflex. (D) An illustration of a segmentation of the vitreous/RNFL (red) border that ignores a vitreal separation. (E, F) The placement of the RNFL/RGC border (orange) border on the temporal side of the fovea differed among segmenters depending on the criterion used to define the RNFL. (G, H) If a border were missing in the fovea (G), segmenters were instructed to extrapolate the border seen outside the fovea to a point on the vitreous/RNFL border in the center of the fovea (H).

TABLE 1

Within-segmenter MEAN(Δ MBL): mean (SD) (in μ m) across 20 scans (10 controls and 10 patients) of the difference between the mean test and retest border locations

Segmenter	Vitreous/RNFL	RNFL/GC	IPL/INL	RPE/BM
A	−0.3 (0.7)	−0.1 (0.6)	−0.1 (1.2)	0.5 (0.8)
B	1.2 (1.8)	0.9 (2.7)	1.1 (1.4)	0.7 (2.6)
C	0.7 (1.2)	−0.7 (1.9)	−0.1 (2.9)	2.2 (2.7)
D	0.0 (0.6)	0.6 (0.8)	1.2 (1.5)	0.7 (2.3)
Mean	0.4 (1.1)	0.2 (1.5)	0.5 (1.8)	1 (2.2)

TABLE 2

Within-segmenter MEAN(Δ MBL): mean (SD) (in μ m) across 20 scans (10 controls and 10 patients) of the unsigned (absolute) difference between the mean test and retest border locations

Segmenter	Vitreous/RNFL	RNFL/RGC	IPL/INL	RPE/BM
A	0.7 (0.4)	0.5 (0.4)	0.9 (0.8)	0.8 (0.4)
B	1.8 (1.2)	2.3 (1.5)	1.6 (0.8)	2 (1.7)
C	1.1 (0.7)	1.6 (1.3)	2.2 (1.8)	2.5 (2.4)
D	0.5 (0.3)	0.8 (0.5)	1.3 (1.4)	1.8 (1.6)
Mean	1	1.3	1.5	1.8

TABLE 3

Within-segmenter $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$: mean (in μm) across 20 scans (10 controls and 10 patients) of the mean of the unsigned (absolute) point-by-point difference between the test and retest border locations for each scan

Segmenter	Vitreous/RNFL	RNFL/RGC	IPL/INL	RPE/BM
A	1.6 (0.6)	1.7 (0.5)	2.2 (0.6)	2 (0.8)
B	2.7 (1.2)	3.8 (1.6)	3 (0.8)	3.1 (1.5)
C	3.2 (1.2)	4.1 (1.4)	4.1 (1.6)	4.7 (1.8)
D	1.9 (0.4)	2.4 (0.5)	3.3 (0.9)	3 (1.2)
Mean	2.4	3	3.1	3.2

TABLE 4

Between-segmenter MEAN(Δ MBL): mean (SD) (in μ m) across 20 scans (10 controls and 10 patients) of the signed difference between the mean locations for segmenter X and the mean locations of the other three segmenters

Segmenter	Vitreous/RNFL	RNFL/GC	IPL/INL	RPE/BM
A	0 (1.1)	-0.2 (1.3)	0.8 (1.5)	0.2 (1.4)
B	-0.6 (1.3)	2.4 (2.7)	0.1 (1.6)	-0.8 (1.8)
C	0.8 (1.1)	-2.1 (2.1)	-1.8 (2.1)	0.8 (2.5)
D	-0.2 (0.9)	-0.1 (1)	0.8 (1.2)	-0.2 (1.8)
Mean	0 (1.1)	0 (1.8)	0 (1.6)	0 (1.9)

The values given in bold for segmenters B (RNFL/RGC) and C (RNFL/RGC and IPL/INL) suggest that there were systematic differences on the order of 2 μ m in these cases.

TABLE 5

Between-segmenter MEAN(Δ MBL): mean (SD) (in μ m) across 20 scans (10 controls and 10 patients) of the unsigned (absolute) difference between the mean border locations for segmenter X and the mean locations of the other three segmenters

Segmenter	Vitreous/RNFL	RNFL/RGC	IPL/INL	RPE/BM
A	0.8 (0.7)	1 (1)	1.4 (1.1)	1.1 (0.8)
B	1.2 (0.8)	3.1 (1.8)	1.2 (1)	1.7 (1)
C	1.1 (0.9)	2.5 (1.5)	2.3 (1.6)	2.1 (1.5)
D	0.7 (0.6)	0.8 (0.6)	1.2 (0.8)	1.5 (1.1)
Mean	0.9	1.9	1.5	1.6

TABLE 6

Between-segmenter $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$: mean (in μm) across 20 scans (10 controls and 10 patients) of the mean of the absolute (unsigned) point-by-point difference between border locations for each scan for segmenter X and the mean of the other three segmenters

Segmenter	Vitreous/RNFL	RNFL/RGC	IPL/INL	RPE/BM
A	1.9 (1.1)	2.2 (1.2)	2.8 (0.9)	2.4 (0.9)
B	2.3 (0.6)	4 (1.5)	2.6 (0.8)	2.6 (0.8)
C	2.8 (1.6)	3.9 (1.4)	3.9 (1)	3.6 (1.7)
D	2 (0.8)	2.5 (0.5)	2.7 (0.7)	2.7 (0.7)
Mean	2.3	3.2	3	2.8

TABLE 7

Pair-wise between-segmenter $\text{MEAN}\{\text{mean}(\Delta\text{LBL})\}$: mean (in μm) across 20 scans (10 controls and 10 patients) of the mean of the absolute (unsigned) point-by-point difference between border locations for pair-wise comparisons of segmenters

Segmenter	Vitreous/RNFL	RNFL/RGC	IPL/INL	RPE/BM
A vs. B	2.6 (1.1)	3.9 (1.6)	3.3 (1.1)	2.9 (1.2)
A vs. C	3 (2)	3.7 (1.9)	4.4 (1.4)	3.9 (1.6)
A vs. D	2.2 (0.7)	2.4 (0.6)	3 (0.8)	3 (0.8)
B vs. C	3.3 (1.4)	5.3 (2)	4.1 (1.1)	3.9 (1.7)
B vs. D	2.4 (0.8)	4.1 (1.4)	3.1 (0.8)	3 (0.8)
C vs. D	3.3 (1.6)	3.9 (1.3)	4.4 (1)	4.2 (1.6)
Mean	2.8	3.9	3.7	3.5

TABLE 8

Layer thickness (within-segmenter): mean (SD) (in μm) across 20 scans (10 controls and 10 patients) of the difference between the mean test and retest retinal layer thicknesses

Segmenter	RNFL	RGC + IPL	GCC	Total retina
A	0.3 (0.7)	0 (1.1)	0.3 (1.2)	0.8 (0.9)
B	-0.4 (2.5)	0.2 (3.1)	-0.1 (1.9)	-0.4 (3.3)
C	-1.6 (1.5)	0.6 (2.2)	-0.9 (2.6)	1.5 (2.7)
D	0.6 (1.1)	0.6 (1.6)	1.2 (1.7)	0.7 (2.4)
Mean	-0.3 (1.4)	0.4 (2)	0.1 (1.8)	0.7 (2.3)

TABLE 9

Layer thickness (between-segmenter): mean (SD) (in μm) across 20 scans (10 controls and 10 patients) of the signed difference between the mean retinal layer thickness for segmenter X and the mean retinal layer thickness of the other three segmenters

Segmenter	RNFL	RGC + IPL	GCC	Total retina
A	−0.3 (1.3)	1.2 (1.5)	0.9 (1.4)	0.2 (1.9)
B	3 (2.8)	−2.4 (3.2)	0.6 (1.8)	−0.2 (2.3)
C	−2.8 (1.5)	0.2 (1.8)	−2.6 (2.2)	0.2 (2.5)
D	0.1 (1.2)	1 (1.6)	1.1 (1.3)	−0.1 (2)
Mean	0 (1.7)	0 (2.0)	0 (1.7)	0 (2.2)