Automated Glaucoma Detection with Optical Coherence Tomography

Automatische Glaukomerkennung mit optischer Kohärenztomographie

Der Technischen Fakultät der
Friedrich-Alexander-Universität Erlangen-Nürnberg

zur Erlangung des Grades

Doktor-Ingenieur (Dr.-Ing.)

vorgelegt von

Markus Anton Mayer

aus

Ingolstadt

Als Dissertation genehmigt von der
Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	13.07.2018
Vorsitzender des Promotionsorgans:	Prof. Dr.-Ing. Reinhard Lerch
Gutachter:	Prof. Dr.-Ing. Joachim Hornegger
	Prof. Sina Farsiu, Ph.D.

Abstract

The number of patients suffering from the glaucoma disease will increase in the future. A further automation of parts of the diagnostic routine is inevitable to use limited examination times more efficiently. Optical coherence tomography (OCT) technology has become a widespread tool for glaucoma diagnosis, and data collections in the clinics have been built up in recent years that now allow for data mining and pattern recognition approaches to be applied to the diagnostic challenge. A complete pattern recognition pipeline to automatically discriminate glaucomatous from normal eyes with OCT data is proposed, implemented and evaluated. A data collection of 1024 Spectralis HRA+OCT circular scans around the optic nerve head from 565 subjects build the basis for this work. The data collection is labeled with 4 diagnoses: 453 healthy (H), 179 ocular hypertension (OHT), 168 preperimetric glaucoma (PPG), and 224 perimetric glaucoma (PG) eyes.

In a first step, 6 retinal layer boundaries are automatically segmented by edge detection and the minimization of a custom energy functional, which was established in preceeding work by the author. The segmentation algorithm is evaluated on a subset consisting of 120 scans. The automatically segmented layer boundaries are compared to a gold standard (GS) created from manual corrections to the automated results by 5 observers. The mean absolute difference of the automated segmentation to the GS for the outer nerve fiber layer boundary is 2.84μm. The other layers have less or almost no segmentation error. No significant correlation between the segmentation error and scans of bad quality or glaucomatous eyes could be found for any layer boundary. The difference of the automated segmentation to the GS is not much worse than the single observer's manual correction difference to the GS.

In a second step, the thickness profiles generated by the segmentation are used in a classification system: In total, 762 features are generated, including novel ratio and principal component analysis features. "Forward selection and backward elimination" selects the best performing features with respect to the classwise averaged classification rate (CR) on the training data. The segmentations of the complete dataset were manually corrected so that the classification experiments could either be run on manually corrected or purely automated segmentations. Three classifiers were compared. The support vector machine classifier (SVM) performed best in a 10-fold cross-validation and differentiated non-glaucomatous (H and OHT) from glaucomatous (PPG and PG) eyes with a CR of 0.859 on manually corrected data. The classification system adapts to the less reliable purely automated segmentations by choosing features of a more global scale. Training with manually corrected and testing with purely automated data and vice versa shows that it is of advance to use manually corrected data for training, no matter what the type of test data is. The distance of the feature vectors to the SVM decision boundary is used as a basis for a novel glaucoma probability score based on OCT data, the OCT-GPS.

Zusammenfassung

Eine steigende Anzahl von Glaukompatienten wird es unabdingbar machen Teile der diagnostischen Routine weiter zu automatisieren. Die optische Kohärenztomographie (OCT) ist inzwischen ein fester Bestandteil der Glaukomdiagnose geworden und die Kliniken haben Datensammlungen aufgebaut, die Datenbankauswertungen und Mustererkennungsansätze für die diagnostischen Herausforderungen erlauben: Augen mit Glaukom sollen anhand von OCT automatisch von gesunden Augen unterschieden werden. Hierfür wird eine vollständige Mustererkennungskette vorgeschlagen, implementiert und evaluiert. Die Arbeit basiert auf einem Datensatz aus 1024 kreisförmigen Scans um den optischen Nervenkopf von 565 Personen, die mit einem Spectralis HRA+OCT aufgenommen wurden. Die Daten stammen von Augen mit 4 verschiedenen Diagnosen: 453 gesunde Augen (H), 179 Augen mit erhöhtem Augeninnendruck (OHT), 168 Augen mit präperimetrischem Glaukom (PPG) und 224 Augen mit perimetrischem Glaukom (PG).

In einem ersten Schritt werden sechs Retinaschichtgrenzen mit Kantendetektion und der Minimierung eines Energiefunktionals, das in vorrangegangenen Arbeiten eingeführt wurde, segmentiert. Der Segmentierungsalgorithmus wird mit Hilfe eines Goldstandards evaluiert, der aus manuellen Korrekturen am automatischen Ergebnis von fünf unabhängigen Beobachtern abgeleitet wurde. Die automatische Segmentierung der äußeren Begrenzung der Nervenfaserschicht weicht im Mittel 2.84μm vom Goldstandard ab. Die Segmentierungsfehler bei den anderen Schichtgrenzen sind geringer oder kaum vorhanden. Es wurde keine signifikante Korrelation zwischen den Segmentierungsfehlern und Scans von schlechter Qualität oder der Glaukomdiagnose festgestellt. Das Ergebnis der automatischen Segmentierung unterscheidet sich vom Goldstandard nicht deutlich mehr als die manuellen Korrekturen der einzelnen Beobachter von Goldstandard.

In einem zweiten Schritt werden die aus der Segmentierung gewonnenen Retinaschichtdickenprofile als Eingabe eines Klassifikationssystems verwendet: Es werden 762 Merkmale generiert, u.a. neuartige Verhältnis- und Hauptachsenmerkmale. "Alternierende Merkmalshinzufügung und Ausschluss" wählt die besten Merkmale automatisch aus. Die automatischen Segmentierungen des ganzen Datensatzes wurden manuell korrigiert, um Klassifikationsexperimente sowohl auf manuell korrigierten, als auch auf komplett automatisch erzeugten Segmentierungen durchführen zu können. Drei Klassifikatoren werden verglichen, wobei die Support Vektor Maschine (SVM) das beste Ergebnis in einer 10-fachen Kreuzvalidierung liefert. Es werden Nicht-Glaukom (H und OHT) von Glaukomaugen (PPG und PG) mit einer klassenweise gemittelten Klassifikationsrate von 0.859 auf manuell korrigierten Daten unterschieden. Das Klassifikationssystem adaptiert sich an die weniger zuverlässigen, komplett automatischen Segmentierungen, indem aus größeren Regionen berechnete Merkmale ausgewählt werden. Wenn das Training auf manuell korrigierten Daten und der Test mit komplett automatisch generierten Daten und umgekehrt durchgeführt werden, zeigt sich, dass es von Vorteil ist, immer manuell korrigierte Daten zum Training zu verwenden, unabhängig vom Datentyp der Testdaten. Die Distanz eines Merkmals zur SVM Entscheidungsgrenze wird abschließend benutzt, um einen neuartigen Glaukomwahrscheinlichkeitsindex für OCT zu konstruieren, den OCT-GPS.

Acknowledgment

The years I spent as a PhD student at the pattern recognition lab at the FAU were among the best in my life. The working atmosphere was great and enabled ideas to be pushed into reality. To allow such an atmosphere to grow is the effort of the Chair of the lab and my supervisor,

Prof. Dr.-Ing. Joachim Hornegger. First of all, thanks go to him for giving me the opportunity to start my PhD in the field of ophthalmic imaging. He never denied the hard work that his PhD students have to put into research and teaching - but always gave support when necessary and set me back on track. Especially during the time my first child was born he helped me greatly by ensuring my employment at the lab and managing my paternity leave. His way to motivate people, his honesty and absolute fairness are character traits to look up to. This PhD thesis would not have been possible without guidance from the ophthalmic clinic, which was provided by

Dr. Ralf Tornow. He is among the most humble persons I have ever met. Working together with him was a joy from the beginning of the PhD time on until the final writing stage, were he was my main assistance in getting the medical and physical details right. Working as a researcher would not have had such an impact on my life without the

Other colleagues and the staff at the pattern recognition lab. You could always ask anyone for suggestions, support for scientific and for technical problems. They were the best company for leisure activities, too. Some became and always will be close friends - Anja, Jörg and Rüdiger: Thank you. The opportunities given to me to visit other universities and conference visits have lead to contacts around the globe. I am grateful to my

International collaborators. Radim Kolar from Brno University, Axel Petzold and Lisane Balk from VUmc Amsterdam, Shahab Chitchian from the University of Texas, Ben Potsaid and Jim Fujimoto from the MIT. For the fruitful scientific conversations that pushed me further and for their examples as exceptional researchers I thank Sina Farsiu, Duke University, and Ivan Selesnick, NYU Tandon School of Engineering.

I gratefully acknowledge the funding by the German Academic Exchange Service (DAAD) and the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German National Science Foundation (DFG) in the framework of the excellence initiative.

Without the understanding and help of my supervisor at my current employer, ARRI, this thesis would not have been finished. Thank you, Dietmar Püttmann. Finally, there were always those around me who accepted what I am doing and gave me support:

My family. My deepest thanks go to my parents, my brother Christian and sister Veronika, my girlfriend Katharina and to the mother of my two children, Xaver and Valentin, Lioba.

Markus Mayer

Contents

Chapter 1

Introduction

1.1 Motivation

An estimation of the number of people suffering from the glaucoma disease from the year 2006 yielded that there were approximately 60.5 million open angle glaucoma (OAG) and angle closure glaucoma (ACG) patients worldwide in 2010. Of those, 8.4 million were bilaterally blind [Quig 06]. The blindness caused by glaucoma and the structural damage done is irreversible. However, it is possible to slow down the progression of the disease [Heij 02, Lesk 03, Lee 05]. Therefore, it is essential to diagnose glaucoma at an early stage, before severe vision loss has occurred.

There are various forms of glaucoma. It is a chronic disease that cannot be diagnosed depending on a single measurement or incidence. Thus, the ophthalmologist utilizes a variety of modalities together with the anamnesis of the person to identify the disease. The diagnosing process is time-consuming, due to the variety of modalities that may be involved, like visual field (VF) test, fundus photography, Heidelberg retina tomograph (HRT), and optical coherence tomography (OCT). On the one hand, a multitude of modalities and complex images, e.g. 3D volume scans of the retina, make a diagnosis more precise. On the other hand, each modality requires examination time and time to study its result. The number of glaucoma patients will increase in the future. The estimation of OAG and ACG patients is 79.6 million in 2020 [Quig 06]. Both challenges, the time demand of modalities involved in a precise diagnosis as well as the increasing number of patients, may be approached by automating parts of a diagnostic routine.

Before a person enters the eye clinic, dedicated screening centers can differentiate between patient suspects and healthy people in an efficient manner. For such a differentiation, only a limited number of diagnostic modalities is necessary and examinations and diagnoses can be carried out automatically or by trained personnel to a large extent. When a detailed examination of a patient suspect is carried out in the clinics, reports may be automatically generated for the ophthalmologist to break down huge amounts of image data into a few meaningful parameters. Instead of performing a time demanding manual inspection of the data, the ophthalmologist only needs to check the automated results.

Automated computerized methods are already in widespread use in eye clinics today, not only for research purposes, but also in commercial products. OCT sys-

tems like the Zeiss Cirrus (Carl Zeiss Meditec AG, Jena, Germany) or the Heidelberg Engineering Spectralis (Heidelberg Engineering, Heidelberg, Germany) provide segmentations of the retina and the retinal nerve fiber layer with the possibility to calculate mean thickness values or compare the results with a normative database. One step further, the HRT (Heidelberg Engineering, Heidelberg, Germany) uses machine learning methods to calculate a glaucoma probability score (GPS) based on the imaged topography of the retina [Swin 00].

In this work, an automated glaucoma score similar to the GPS is proposed for circular scan OCT data. The discrimination between glaucoma patients and normal subjects is not performed directly on the OCT images, but on thickness profiles of retinal layers. Therefore, the boundaries of these layers have to be segmented beforehand. The first part of this thesis presents and evaluates an approach for segmenting retinal layers on circular scan OCT data that is applicable on scans of both normal and glaucomatous eyes. In the second part of the thesis, the thickness profiles of multiple retinal layers form the data on which a classification system for the glaucoma disease is built upon. Feature selection, a typical data mining method [Fayy 96], is used to automatically detect the relevant information within of the data. Classification experiments are constructed, and both the results and the selected features are presented. Besides using manually corrected segmentations, the possibilities for a completely automated screening system are investigated by performing a classification on the untouched automated segmentation results that include possible segmentation errors. Finally, a method for transforming classification results into a glaucoma score is presented.

1.2 OCT in ophthalmology

OCT was invented by Huang et al. in 1991 [Huan 91]. It is based on the principles of a Michelson interferometer and is the optical counterpart to ultrasound B mode. Figure 1.1 shows a schemativ view of a time domain OCT (TD-OCT) system. Short coherent light is split at a half-translucent mirror into a measurement and a reference arm. A moveable mirror reflects the light in the reference arm. In the measurement arm, the light is reflected and backscattered inside the object. The beams recombine at the half-translucent mirror. Due to the interference of the combined beams, the signal measured at the detector oscillates when the reference mirror is moved within the coherence length of the light. The intensity of the light coming from the object can be calculated from these oscillations. A single depth profile of the object is acquired by moving the reference mirror over the desired depth range. The beam may be scanned over the object in transverse direction for 2D or 3D imaging. In OCT, the transverse resolution is independent from the axial resolution. While the axial resolution depends on the wavelength and spectrum of the light source, the transverse resolution is determined by the focusing properties of the light beam [Ferc 03]. A more detailed explanation of OCT technology can be found in [Ferc 03, Fuji 03, Wojt 10].

Since the invention of the OCT technology, ophthalmology has been its main application area. In the original OCT paper, image examples of a human retina ex vitro are shown [Huan 91]. Soon in-vivo imaging was possible [Ferc 93]. Early research and commercial TD-OCT systems image up to a few hundred depth profiles, called A-

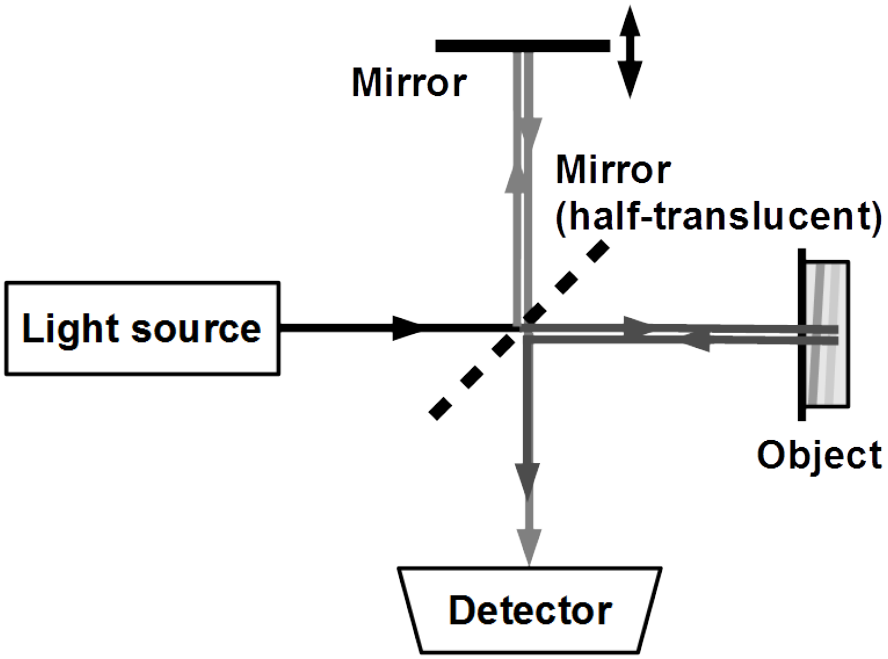


Figure 1.1: Schematic structure of a time domain OCT system. The short coherent light emitted at the light source is split at a half-translucent mirror into a reference and a measurement arm. The light reflected in both arms is recombined, and the intensity of the light reflected and backscattered in the object is measured at the detector indirectly through interference. Light paths in this drawing are schematic. The light reflected may actually overlap the incident light.

Scans, per second. A 2D OCT image is composed of multiple A-Scans. Depending on the system, this may be several hundred to well above a thousand A-Scans per image. With the slow scanning rates of the early systems, the acquisition of a 2D OCT image took up to a few seconds. As the eye constantly moves due to heartbeat, respiration, slow drifts, and fast saccades distortions due to motion on TD-OCT images were a common unwanted imaging artifact. The clinical acceptance of OCT-systems was boosted by the commercial availability of frequency domain OCT (FD-OCT) systems. Contrary to TD-OCT, the mirror that moved in TD-OCT systems is now in a fixed position, and depth information is acquired by analyzing the spectrum of the back reflected light from the object [Ferc 95, Haus 98]. Today, about 20000 A-Scans/s are common in commercial systems, while research OCT systems reach up to 300000 A-Scans/s [Pots 08]. Imaging in 2D became the matter of a fraction of a second and 3D OCT volumes may be acquired. Motion artifacts are negligible for 2D images with fast FD-OCT acquisitions.

The part of the eye that is commonly imaged with OCT is the retina. As mentioned before, in the early days of the OCT technology imaging was limited to a 2D space and up to now, the two most common scan patterns used in the clinics are 2D scans, namely a linear scan through the macula or a circular scan around the optic nerve head (ONH). An example of a circular scan around the optic nerve head is shown in Figure 1.2 a) and c). The scan path is drawn on the respective scanning

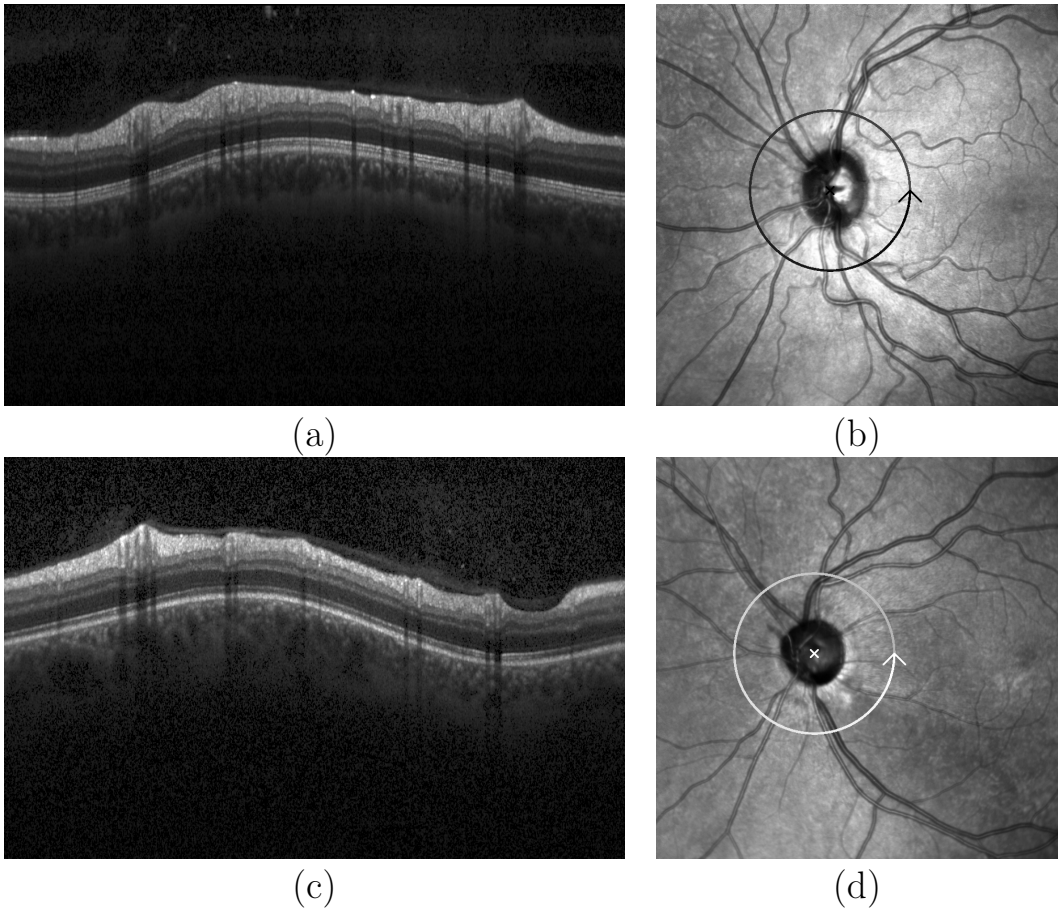


Figure 1.2: Example circular B-Scans imaged with a Heidelberg Engineering Spectralis HRA+OCT (a) OCT B-Scan of the right eye of a normal subject. (b) Scanning Laser Ophthalmoscope (SLO) image captured during the same scanning process. The circular scan path position is marked. The gray value of the scan position path does not have any relevance. The scan begins and ends at the arrow tip. The positions along the scan path on the SLO image correspond to the columns of the OCT image from left to right. (c) OCT B-Scan of the left eye of a glaucoma patient. (d) Corresponding SLO image for scan (c).

laser ophthalmoscope (SLO) image in Figure 1.2 b) and d). The SLO technique delivers images similar to fundus photography. The SLO images are acquired during the same scanning process as the OCT image. The layered structure of the retina can be clearly observed on the OCT scans in Figure 1.2 a) and c). The vitreous humor (VH) lays on top of the retina as a black, non-scattering region. The uppermost layer visible is the retinal nerve fiber layer (RNFL), which is separated from the VH by the inner limiting membrane (ILM). Then, from the inner to the outer retina, follows the ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), external limiting membrane (ELM), inner photoreceptors (IPR), outer photoreceptors (OPR), and finally the retinal pigment epithelium (RPE). Blood vessels cast black shadows on to the imaged tissue below, as the flowing blood hinders a deeper penetration and backscattering of

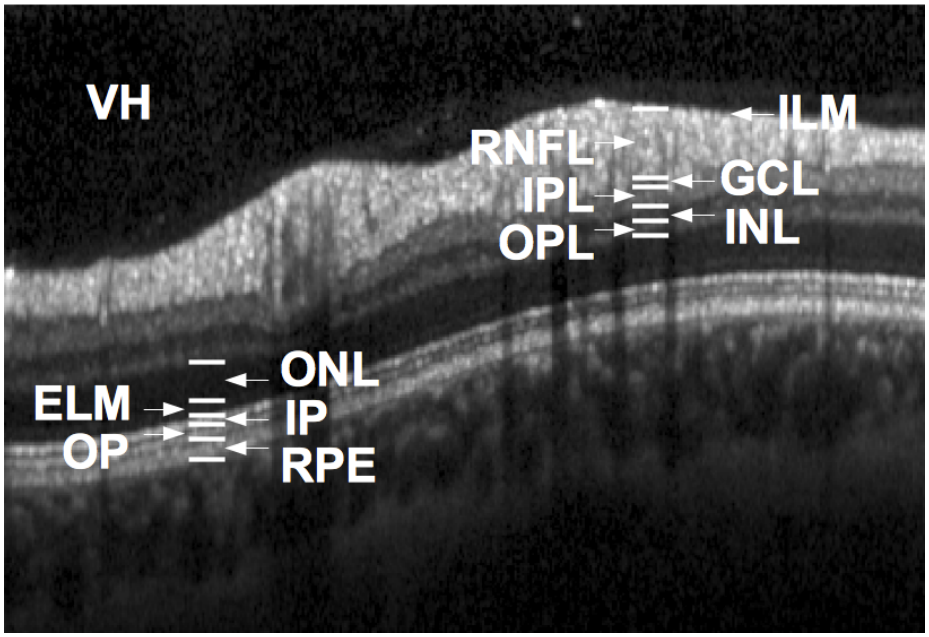
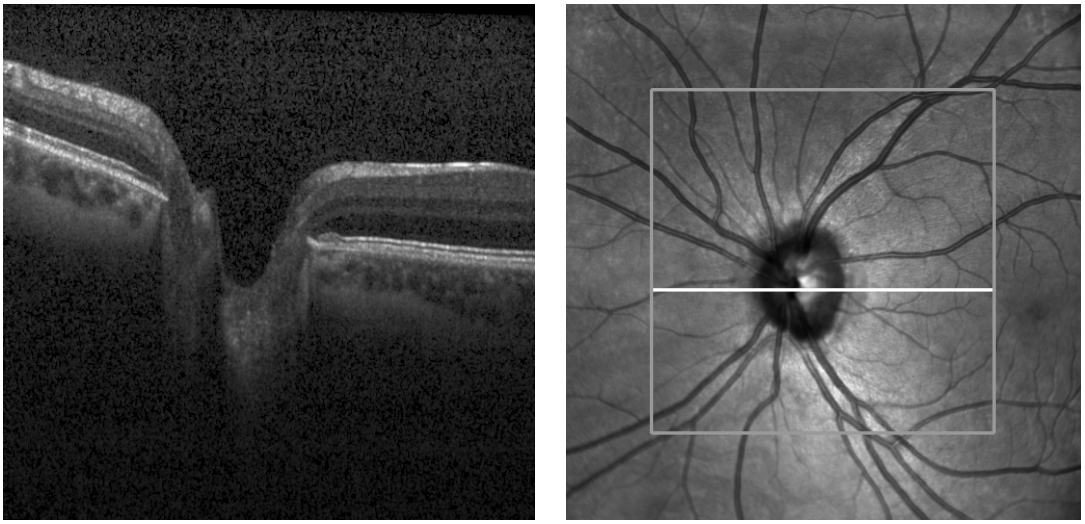


Figure 1.3: Denominations of the retinal layers in a circular OCT scan. Abbreviations: Vitreous humor (VH), inner limiting membrane (ILM), retinal nerve fiber layer (RNFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), external limiting membrane (ELM), inner photoreceptors (IPR), outer photoreceptors (OPR), retinal pigment epithelium (RPE). The thickness of the ILM is below the resolution capabilities of the Spectralis OCT system that was used to acquire this image. The contrast of the OCT image was adjusted for better layer visibility.

light. In Figure 1.3, a section of the OCT scan of Figure 1.2 a) is annotated with the layer names. The influence of measurements on these layers for glaucoma diagnosis is briefly touched in the next section.

Volume imaging of the retina is not as standardized as the circular scan pattern. Possibilities include scanning the macula region and the ONH region. Research systems allow wide angle scans of a large field of the retina, including both the macula and ONH. Figure 1.4 a) shows an example slice out of a volume scan of the ONH. The complete scan area and the position of the slice is again marked on the corresponding SLO image in Figure 1.4 b). Not only the retina, but also the anterior parts of the eye may be imaged with OCT [Izat 94, Leun 05, Kale 07], e.g. the cornea, iris, anterior and posterior chamber, and lens. Besides its application in ophthalmology, OCT is now used in other medical fields like gastroenterology and dermatology. An overview of the usage of OCT in medicine is given by [Ferc 10]. Applications of OCT outside the biomedical field are presented in [Stif 07].

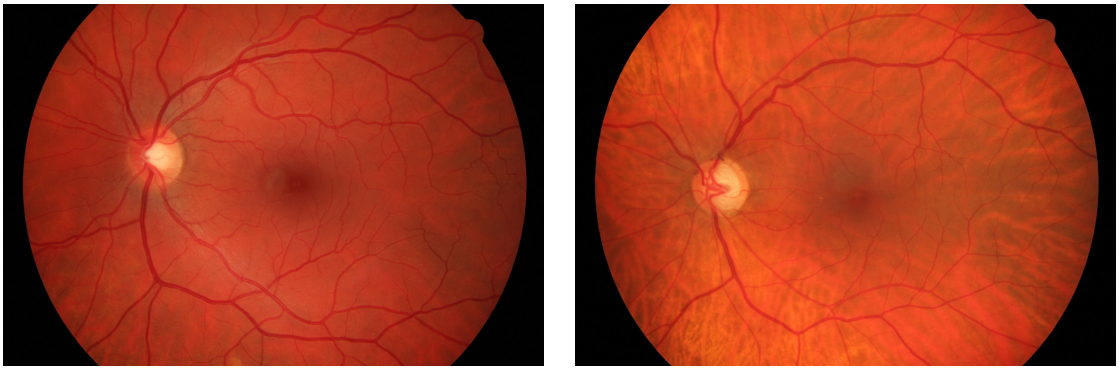


(a) (b)

Figure 1.4: Example B-Scan from an OCT volume imaged with a Heidelberg Engineering Spectralis HRA+OCT (a) OCT B-Scan through the optic nerve head of the left eye of a normal subject. (b) Scanning Laser Ophthalmoscope (SLO) image captured during the same scanning process. The scan path of the B-Scan in (a) marked in white. The positions along the scan path on the SLO image correspond to the columns of the OCT image. The complete scan area of the OCT volume, which consists of 97 B-Scans parallel to the one shown, is marked in grey.

1.3 Glaucoma diagnosis with OCT

The chronic glaucoma disease appears in various forms, like OAG and ACG. But they all have common characteristics that allow for a diagnosis. Most importantly, defects in the visual field of the patients appear. But by the time these defects are noticed by the patient or even by visual field measurements in a clinic, structural damage has already occured [Somm 91, Tuul 93, Woll 05]. However, recent studies propose a linear model of the relationship between VF loss and RNFL thinning [Hood 07, Horn 09]. Which one is detected earlier depends on the standard deviations of the tests and its conditions. The structural damage is a result of the death of retinal ganglion cells (RGC) and their axons. This leads to an excavation of the optic disc, an optic nerve degeneration and a thinning of the RNFL. These effects can be observed on (stereo) fundus photographs. A thinning of the optical rim, and thus a change in the cup-to-disc ratio, can be seen in Figure 1.5 b) compared to the fundus photograph of a normal subject in Figure 1.5 a). Fundus photographs provide a detailed 2D view on top of the retina, with additional topography information available when stereo fundus photographs are acquired. The HRT acquires the topography information in a single acquisition. Topography information allows for a better diagnosis based on ONH information, as the steepening of the ONH rim, the volume of the ONH cup, and the cup-to-disc ratio can be measured more precisely compared to a pure 2D image.



(a) (b)

Figure 1.5: Example fundus photographs from a publicly available database of high resolution fundus photographs [Odst 13]. (a) Normal subject. (b) Glaucoma patient.

With OCT, the effect of the dying RGC and their axons, the nerve fibers, can be directly observed by the visualization of the thickness of the RNFL and macula, the point on the retina which allows for the sharpest sight and is not covered by nerve fibers. These two measurements, RNFL and macula thickness, have been two of the most important glaucoma indicators from their first reports [Gued 03, Woll 05] until today. It was shown that the mean RNFL thickness correlated better with glaucoma than macula thickness [Gued 03]. This was verified in [Na 11]. The mean RNFL thickness was measured on the standard 3.46mm diameter circular scan around the ONH. The circular scan pattern has the advantage that all nerve fiber bundles going from the retina trough the ONH to the brain pass trough the scanning circle exactly once. In addition, the mean RNFL measure on the standard circular scan is not or only minimally affected by the size of the optic disc [Oddo 11, Huan 12]. Figure 1.2 shows examples of a circular scan of a normal subject and a glaucoma patient. The RNFL of the glaucoma patient is thin and has also disappeared in some areas.

The parameters computed out of RNFL thickness measurements were refined over time, e.g. by computing the mean in the quadrants around the ONH [Polo 08, Leun 09]. In [Vizz 09], it is shown that OCT may detect very localized RNFL defects. Complete 2D RNFL maps generated out of OCT volumes around the ONH show a promising diagnostic performance that may be superior to the traditional measurements on circular scans [Leun 10a, Leun 10b, Kana 13, Maya 13, Shin 15]. It is also noted that the progression of the RNFL defects does not affect all quadrants in the same way, but rather starts locally, before expanding. A detailed overview of OCT and other modalities as tools for glaucoma diagnosis is given in [Shar 08] and [Koto 14].

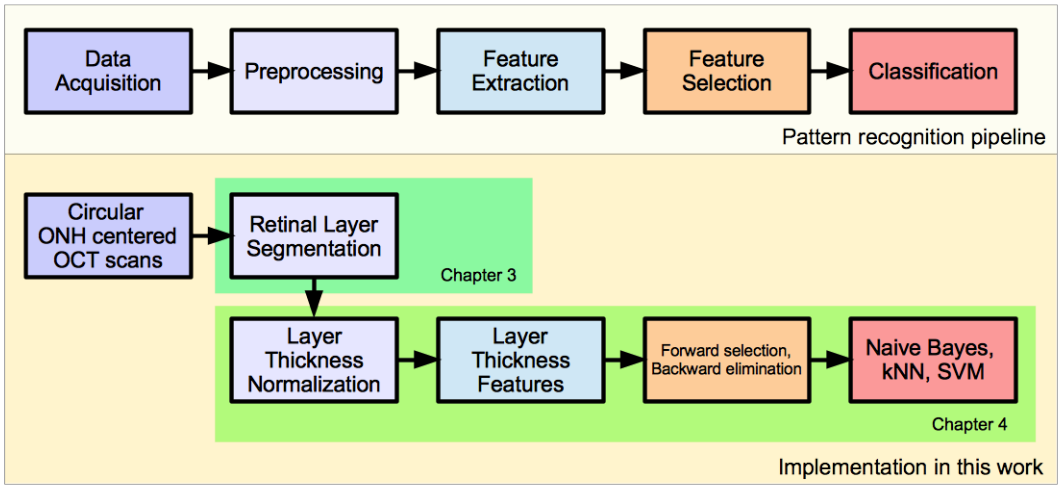


Figure 1.6: The standard pattern recognition pipeline and its implementation in this work. The data used are optic nerve head (ONH) centered circular OCT scans. The preprocessing step is a retinal layer segmentation and the normalization of the derived thickness measurements. After preprocessing, established methods are used to classify the data for the presence of glaucoma: Features are computed out of the thickness profiles, "Forward selection and backward elimination" is used as a feature selection method and three exemplary classifiers, namely naïve Bayes, k-nearest neighbor (kNN) and support vector machine (SVM) are tested for their discriminative ability on the features. The retinal layer segmentation is detailed in Chapter 3, and the rest of the pattern recognition pipeline in Chapter 4.

1.4 Contribution of this work

A complete pattern recognition pipeline is proposed, implemented and evaluated. The standard pattern recognition pipeline and its mapping to the contents of this work is shown in Figure 1.6. The classification task is to discriminate glaucoma patients from normal subjects. The data used are ONH centered circular OCT scans. The data preprocessing involves segmenting the retinal layers from the OCT scans to obtain thickness profiles of multiple layers, which then are optionally normalized. From the retinal layer thickness profiles numerous features are generated. Widely accepted algorithms are used for the feature selection and classification, i.e. "Forward selection and backward elimination" for feature selection and three basic classifiers: Naïve Bayes, k-nearest neighbor (kNN) and support vector machines (SVM). Each of the pipeline steps may be further improved at a later stage, e.g. by utilizing the newest trends in retinal layer segmentation or by testing more elaborated classifiers. The localization of the presented methods in the field of current research and future improvements are given in the respective State of the art and Outlook sections. The utilization of algorithms at the cutting edge of research is not the focus in this work.

The main scientific contribution of the work is to shed a light on how OCT data from daily clinical practice influences automated methods. OCT scans in daily clinical practice sometimes lack the strict quality criteria that are commonly applied

to data in segmentation evaluations or clinical trials. It is common sense that all automated segmentation methods will eventually fail to provide perfect results, be it due to low quality or an unexpected content of the images, e.g. the presence of a disease that was not included in the evaluation data set. The failures of automated segmentation methods might be corrected manually, which takes less effort than complete manual segmentations. Most of the published work on glaucoma detection from OCT data is based on the parameters the proprietary manufacturer software delivers. These parameters are either not manually corrected or the machines even lack the possibility for a manual correction, i.e. clinical trials do most often only use purely automated segmentation results. To the best of out knowledge, we present the first work that evaluates how a manual correction of layer segmentations influences the feature selection and the glaucoma classification scores.

In addition to this main contribution, there are other novelties:

- As mentioned before we base this work on a data collection that should represent daily clinical practice. No OCT scan was excluded in this work due to quality issues or other diseases present, as it is common in other works on retinal layer segmentation [Lang 13, Cara 14] and especially glaucoma classification [Huan 05, Bask 12, Garc 12, Mwan 13, Belg 15]. It should be noted that the scan quality issue has come into focus in recent years. A retinal layer segmentation was developed with the robustness against scan quality by Dufour et al. [Dufo 13]. However, they did not include glaucomatous cases in their evaluation.

- To the author's best knowledge, no work published in the field of retinal layer segmentation has so far evaluated a multi-layer segmentation algorithm not only for global measurements of segmentation errors, but also detailed where the errors occurred most likely. This is an important factor for the influence of segmentation errors on glaucoma detection. Such a local evaluation for the RNFL only was presented in a preceding work by the author in [Maye 10] and was, for example, taken up in [Kaba 15]. [Ehne 14] did at least present measures in 9 local fields. All other works only evaluate for global measurements like a summed absolute difference to a gold standard.

- The segmentation process is designed such that it takes into account complete losses of the retinal nerve fiber layer which is essential for severe glaucomatous cases. [Srin 14a] incorporated a global and local missing layer detection, while this was only done for mice OCT data. A level set based segmentation approach with sub-pixel accuracy was developed by [Cara 14] that should theoretically be able to deal with missing layers. However, the authors note themselves that the method was only evaluated on normal appearing eyes and the application to pathology is yet to study.

- A data mining approach from the retinal layer segmentations is proposed. For each of 6 segmented layer groups and the blood vessel positions, multiple features are computed and an automated feature selection chooses the most discriminative ones. Up to now, only the complete retina, RNFL, GCL+IPL and ONH features were utilized from OCT data to automatically classify for

glaucoma as authors rely on parameters from the manufacturer built-in segmentation algorithms [Burg 05, Bask 12, Garc 12, Mwan 13, Yiu 14].

- Principal component analysis (PCA) features are proposed to be computed from the thickness profiles in addition to the traditional minimum, maximum and mean features. The PCA features were first presented in an abstract and poster by the author on a conference [Maye 09], but missed a detailed description and a database large enough for a resilient evaluation was still to follow.

- An age normalization for the thickness profiles inspired by Bendschneider et al. [Bend 10] is proposed. The normalization in this preceding work is only calculated for the RNFL and not within a classification framework.

- A novel glaucoma probability score based on circular scan OCT data, a retinal layer segmentation, and a classification system is proposed.

1.5 Structure of this work

The properties of the OCT data we utilize are outlined in Chapter 2. The diagnose process on the patients' eyes is briefly summarized. The steps in creating two subsets out of the full OCT scan data collection used in the segmentation evaluation and glaucoma classification are detailed with respect to inclusion and exclusion criteria, as well as scan quality. The two subsets are used for the automated segmentation evaluation and the glaucoma classification evaluation respectively.

The main body of this work is split in two parts, as sketched in Figure 1.6. Chapter 3 describes the automated retinal layer segmentation algorithm. This algorithm is an extension of former work by the author [Maye 10] and is first put into the context of the current state-of-the-art research in Section 3.1. The segmentation method itself is described in Section 3.2, and the evaluation process in Section 3.3. We decided to let the automated segmentation results be corrected by multiple observers and constructed a gold standard from these manual corrections. The measures and results for the inter-observer evaluation are presented and discussed in Section 3.4. For the classification evaluation later in this work, only a manual correction by the author could be utilized due to practical reasons. The relation of this manual correction to the other observers and the gold standard is described. Section 3.5 presents and discusses the measures and results for the automated segmentation method. The chapter on the automated segmentation method is closed by an outlook in Section 3.6 on how the presented algorithm can be extended to volume scans and how the ideas of the method can be combined with recently published approaches to layer segmentation.

The following steps of the pattern recognition pipeline, namely the layer thickness profile normalization, the feature extraction, feature selection and classification (see Figure 1.6) are the content of Chapter 4. The state-of-the-art review in Section 4.1 does not limit itself to glaucoma classification from OCT images, but also takes a look at similar fields, e.g. classifying glaucoma from other measures and imaging techniques related to the eye, as well as the usage of OCT data for an automatic identification of diseases besides glaucoma. In Section 4.2 two approaches to layer

thickness normalization, i.e. age and magnification normalization, are presented. The subsequent Section 4.3, describes the features computed from the layer thicknesses. As the feature selection and the classifiers used are common and widely accepted methods, they are only briefly summarized in Section 4.4. The results of the classification experiments are presented in Section 4.5. Not all possible combinations of classification challenges, thickness normalization methods, classifiers and automated or manually corrected data have informational value. Therefore we first outline how we broke down the parameter matrix by keeping always all parameters except one fixed, and how we ordered the experiments such that a reliable statement can be made. First, the possible classification challenges derived from the diagnoses in our dataset are defined and judged for relevance in Section 4.5.2. Then the influence of the thickness normalization features is evaluated in Section 4.5.3. The best performing classifier is searched for in Section 4.5.4. Finally, the results and changes in the feature selection when switching from manual corrected results to purely automated segmentations are investigated in Section 4.5.5. With the results of the classification experiments in mind, we propose a glaucoma score for circular scan OCT data in Section 4.6. Similar to Chapter 3 on segmentation we conclude the chapter on classification with an outlook to the adaptation of our method to volume data and possible future enhancements.

The work is concluded with a summary of the automated layer segmentation and the glaucoma classification tasks presented and their implications on the field of research in Chapter 5.

Chapter 2

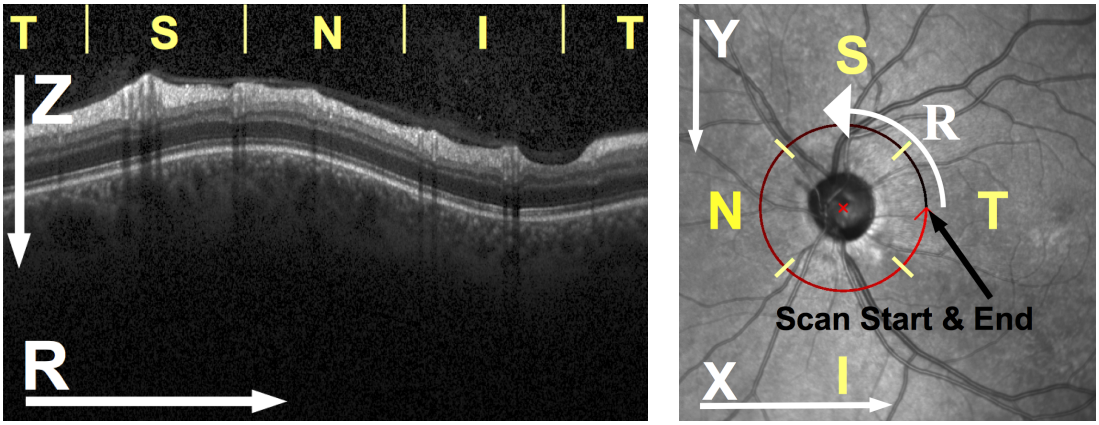
Optical coherence tomography data

This chapter describes the data that is used for the main body of this work. First, in Section 2.1 properties and naming conventions regarding image dimensions, coordinates and anatomical structures are clarified. Second, the construction of the datasets used for evaluating the methods proposed is presented in Section 2.2. The properties of the datasets, especially in differentiation to former works published in the field of OCT retinal layer segmentation and glaucoma classification, are emphasized.

2.1 Properties, names and conventions

The OCT system used for scan acquisition is a Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany). This OCT device is referred to as Spectralis for the remainder of this work. As mentioned in Section 1.2, A-Scan and depth profile are used interchangeably. OCT image and 2D frame are also used as synonyms, referring to scans with a circular scan pattern. A B-Scan denotes either a circular scan or a line scan of an OCT volume. OCT volumes are also be referred to as 3D data.

The circular scans are centered at the optic disk and have a diameter of 3.46mm. All consist of 768 A-Scans. The volume examples shown are centered at the optic nerve head and have a varying number of A-Scans per B-Scan (384 to 512) and a varying number of B-Scans (49 to 97). Each A-Scan consists of 496 pixels. The axial resolution of the Spectralis is 7μm in tissue, although the pixel spacing is 3.87μm. The images are thus oversampled in the axial direction. The raw data was exported using the VOL file format of Heidelberg Engineering. The pixel intensity value range in the VOL files is $[0; 1]$ saved as 32-bit floating point values. All computations are performed in this data format or with 64-bit floating point values. Unless stated otherwise, the intensity values of the VOL file are double square-rooted for display as proposed by Heidelberg Engineering.



(a) (b)

Figure 2.1: Example circular B-Scan of a left eye with coordinate system denominations. Right eye denominations and scan pattern are equivalent and follow the common rules for the mapping between left and right eye. (a) OCT B-Scan. The axial direction is denominated by Z. The transversal direction is denominated by R. (b) SLO image captured by the Spectralis HRA+OCT during the same scanning process. The circular scan pattern position and its direction corresponding to the R-direction in the images is marked. The quadrant borders on the SLO image scan position and on the OCT scan are shown with yellow lines. The quadrants are: Temporal (T), Superior (S), Nasal (N), Inferior (I).

The axial direction of B-Scans is Z. To simplify formulas, the transversal direction in a circular scan, i.e. the position of an A-Scan in the resulting image, is denominated by R. The transversal directions in a volume are X and Y. The Z-direction, as well as the Y and R-directions have their origins in the upper left corner of the corresponding images. Figure 2.1 illustrates these notations on a circular scan. The rough location of landmarks on OCT scans is commonly given in quadrants: Temporal (T), Superior (S), Nasal (N), and Inferior (I) quadrant.

All abbreviations and symbols are summarized in Table A.1 and Table A.2.

2.2 Datasets

The data utilized in this work is derived from the "Erlangen Glaucoma Registry" and reflects data from daily clinical practice. The subjects included visit the Erlangen glaucoma service once a year. The inclusion/exclusion criteria and the type of examinations are defined in a protocol that was approved by the local ethics committee. The study is registered at www.clinicaltrials.gov (NCT00494923) and it followed the tenets of the declaration of Helsinki for research involving human subjects. Informed consent was obtained from all participants in the study.

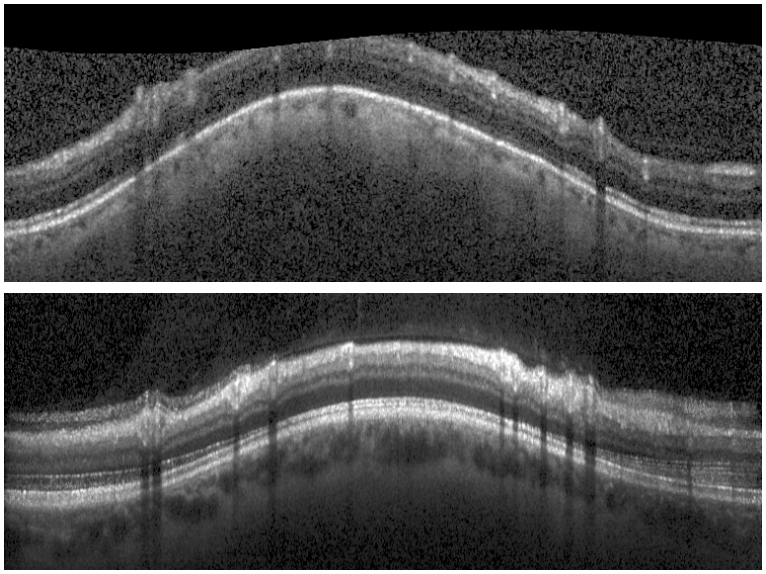


Figure 2.2: Examples of B-Scans *excluded* from the dataset. (a) Retina not completely visible in scan region. (b) Heavy averaging errors that do not allow differentiation of layers by a human observer.

The original dataset consists of 1065 scans from 577 patients. Each eye is included only once at most. After visual inspection 18 images where excluded from the dataset for the following reasons:

- Retina cut off. See example Figure 2.2 a).
- Severe averaging artifacts (even a human observer cannot judge layer boundaries from experience). See example Figure 2.2 b).

Images are explicitly not excluded for following reasons:

- Mild averaging errors that may appear at the left and right borders of a circular OCT scan. See example Figure 2.3 a).
- Low image quality. Even images with complete sections that do not show any structure are included in the dataset. See example Figure 2.3 b).
- Scans of eyes with an obvious disease besides glaucoma. See example Figure 2.3 c).
- Algorithm failures. If the segmentation algorithm fails completely (e.g. due to the reasons above), the image was not excluded from the dataset (as it is practiced in other works [Ishi 05, Tan 09]).

After the exclusion of images with extremely severe scan errors 1046 images from 575 patients remained.

The subjects were diagnosed by medical experts based on an ophthalmic examination using slit lamp inspection, applanation tonometry, funduscopy, gonioscopy,

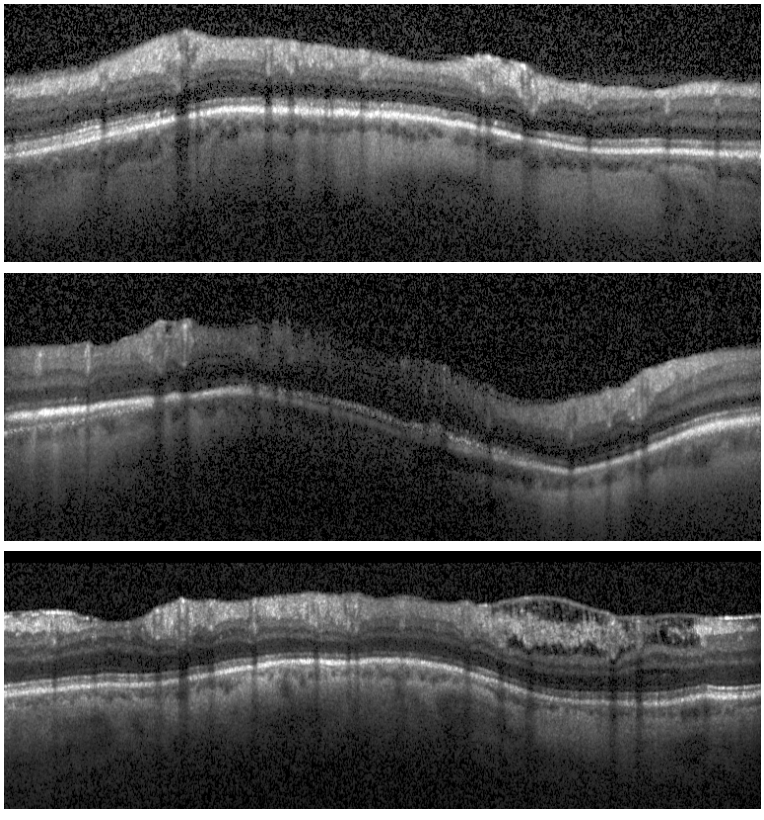


Figure 2.3: Examples of B-Scans *included* in the dataset. The scans exemplify the broad quality and variance in the dataset. (a) Mild averaging errors from a ocular hypertension patient that still allow differentiation of layers by a human observer (HE quality: 19.82dB, Zero quality: 0.67). The averaging errors are mainly located in the temporal quadrant. (b) Very low quality scan from a preperimetric glaucoma patient (HE quality: 19.25dB , Zero quality: 0.62). (c) Besides perimetric glaucoma, the scan also shows signs of an other disease (HE quality: 27.71dB, Zero quality: 0.69).

perimetry and papillometry. A 24-hours intraocular pressure profile with 6 determinations was also obtained. A detailed review of the used diagnostic routine can be found in Baleanu et al. [Bale 09] and Horn et al. [Horn 09] and is not within the scope of this work. The dataset contains the following four diagnoses defined in [Horn 11]:

- **Healthy subjects** (H): Findings in slit lamp inspection, tonometry without medication, and funduscopy were in the normal range. White-on-white perimetry was classified as normal. Optic discs were inspected and classified as normal.
- **Ocular hypertension** (OHT): Patients of this group had intraocular pressures above 22 mm Hg upon repeated applanation tonometry measurements. All OHT patients had normal white-on-white perimetry and normal appearing optic discs.
- **Preperimetric glaucoma patients** (PPG): Patients showed glaucomatous abnormalities of the optic discs (diffuse or localized loss of neuroretinal rim).

OS/OD	N	OHT	PPG	PG	U
H	192	1	4	0	21
OHT	0	68	9	8	4
PPG	4	9	43	14	12
PG	0	8	20	79	11
U	39	4	10	5	10

Table 2.1: Diagnosis left-eye and right-eye distribution among the complete dataset with scan failures already excluded. Rows: Left-eye diagnosis. Columns: Right-eye diagnosis. The diagnosis groups are Healthy (H), ocular hypertension (OHT), preperimetric glaucoma (PPG), perimetric glaucoma (PG) and unknown diagnosis or scan missing (U).

Computerized visual field examinations with white-on-white perimetry were normal.

- **Perimetric glaucoma patients** (PG): The patients of this group had glaucomatous optic disc damage and non-normal white-on-white perimetry.

Unfortunately, for some scans in the dataset diagnoses or age information is unclear or missing. No sex information was available for the subjects. Some are only included with one eye, some with both. The diagnosis in between the eyes may differ, as is reflected in Table 2.1. The table shows the left eye (OS) against the right eye (OD) diagnosis.

Scans without age information or diagnosis were excluded. The dataset of valid sets, e.g. with diagnosis and age, therefore consists of 1024 scans from 565 patients. This is the dataset used for the classification experiments (see Chapter 4) and is named classification dataset within this work. The number of scans for each of the four diagnoses and the age statistics for this dataset are shown in Table 2.2. The healthy group consists of 453 scans, the ocular hypertension group of 179 scans, the preperimetric glaucoma group of 168 scans and the perimetric glaucoma group of 224 scans. The mean age of the subjects increases in the order of the groups mentioned. The overall mean age is 54.98 years.

As mentioned before, no scan was excluded for low quality if no severe artifacts were present that would prevent the examination from even a human expert. The dataset thus consists of scans of varying quality. We intend to refer to the scan quality in our segmentation evaluation (see Section 3.4 and 3.5), therefore a quantification of the quality of a scan is desired. The author proposed such a quality measure in a former work [Maye 10] (Zero quality). During the period of time the subject scans were taken Heidelberg Engineering introduced a quality measure for its Spectralis data. For 820 scans of the classification dataset, this Heidelberg Engineering quality measure (HE quality) is available. The relationship between the two quality measurements is as follows: In Figure 2.4 the distribution of the two measures among the 820 scans of which the HE quality was available is shown. The value range differs and the distribution is not completely linearly related. The correlation coefficient in

Group	#Scans	Age
All	1024	54.98 ± 14.38
H	453	49.48 ± 15.03
OHT	179	55.22 ± 12.65
PPG	168	59.29 ± 10.87
PG	224	62.24 ± 11.24

Table 2.2: Classification dataset statistics. The number of scans for the diagnose and the mean and standard deviation of the respective subjects are shown. The calculation of the mean and standard deviation of the age was performed on a scan basis. If a subject has differing diagnosis for the left and right eye it was included in both of the respective groups.

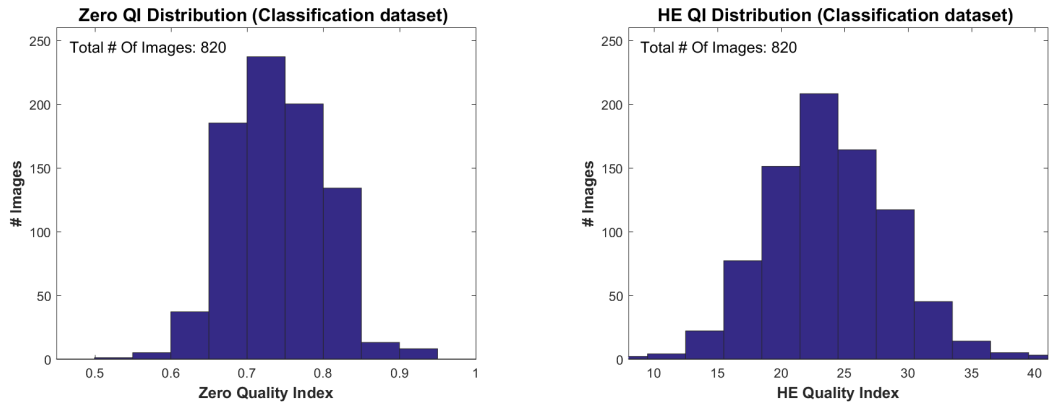
between the two measures on the 820 scans is 0.52 ($p < 0.001$). By visual inspection both measures relate fairly well to the perceived scan quality.

From the classification dataset a subset was created for the evaluation of the automated segmentation algorithm presented in this work. The reason for creating a subset is simple: Manual inspection and a detailed correction of automated segmentation results is time-demanding. More than one hour is necessary for correcting 20 B-Scans. As multiple experienced observers should perform the task, a manual correction of the complete classification dataset was not possible. Therefore, a subset called segmentation evaluation dataset was created following these rules:

- 30 scans were selected from each diagnosis group, i.e. 120 scans in total. This tradeoff between evaluation database size and observer work time was reasonable. 120 scans is a database size comparable to other works published in the field (see Tables B.1 and following).
- 15 left eyes and 15 right eyes are contained in each diagnosis group.
- From one subject only one eye was included.
- HE quality measure is available for each included scan.

This set was created from a random permutation of the full classification dataset. Scans were added to the segmentation evaluation dataset if one of the above rules was not violated, and until all groups had 30 images. After a final check for the conditions, this dataset was stored and not modified further. The selection was completely random and no image was included or excluded for quality or algorithm failure reasons. The statistics of the segmentation evaluation dataset are summarized in Table 2.3. In addition to the mean and standard deviation of the age, the mean and standard deviations of the HE and Zero quality measures are shown.

The mean of the quality measures for each group always lays within the standard deviation around the mean of all the other groups. To define a separation boundary between low and high quality scans, the median of the measures is taken. This median is 0.734 for the Zero quality and 22.79dB for the HE quality. It is of interest whether the diagnosis of the subject correlates to the quality measure, e.g. whether the age



(a) (b)

Figure 2.4: Comparison of the distribution of two quality measures, the Zero quality measure as proposed in [Maye 10] and the built-in quality measure of the Heidelberg Engineering Spectralis (HE quality, in [dB]). The histograms were computed from the 820 scans from the classification dataset were the HE quality was available. For both measures, the relationship "higher value is higher quality" holds. The correlation between the two measures on the 820 scans is 0.52 ($p < 0.001$).

and pathology of glaucoma patients influences the scanning process such that a lower quality scan can be expected. The PPG and PG groups together form a glaucoma (G) diagnosis group. Both the low quality (quality measure is below or equal to the median quality) and glaucoma group have 60 subjects. The correlation of a subject belonging to the glaucoma group and the subject having a low quality scan gives a hint to the independence of the two properties of a scan. The correlation of the HE low quality scans to the glaucoma group is -0.03 and the correlation of the Zero low quality scans to the glaucoma group is -0.07. Both measures are thus nearly linearly independent from the glaucoma diagnosis. This reflects the result in the previous work [Maye 10]. For the remainder of the work we focus on one quality measure. As the quantitative statistics and subjective visual inspection do not favor one measure,

Group	**#Scans**	**Age**	**HE Quality**	**Zero Quality**
All	120	56.75 ± 13.15	23.67 ± 5.08	0.741 ± 0.063
H	30	47.76 ± 14.81	25.12 ± 5.88	0.745 ± 0.066
OHT	30	57.83 ± 11.76	23.69 ± 4.66	0.731 ± 0.052
PPG	30	59.67 ± 10.00	22.49 ± 5.56	0.737 ± 0.072
PG	30	61.74 ± 11.52	23.38 ± 3.85	0.750 ± 0.061

Table 2.3: Segmentation evaluation dataset statistics. The number of scans for the diagnoses and the mean and standard deviation of the respective subjects are shown. In addition, the mean and standard deviation of two quality measures are given: The quality measurement built into the Heidelberg Engineering Spectralis (HE Quality) and the quality measurement defined in [Maye 10] (Zero Quality).

the HE quality is selected. It does not require additional computations and is already included in the scan data, and therefore more easy to access in daily clinical practice.

In addition to the circular scan database, 10 volume scans were available that were centered on the ONH but have differing scan areas and B-Scan numbers. 3 scans of healthy subjects and 7 scans glaucoma patients were available. For each subject only one eye was scanned. An example scan of a healthy eye is shown in Figure 1.4.

Finally, let us answer the question how the properties of the two datasets compare to other works published: While the majority of recently published works on retinal layer segmentation is aimed at segmenting volume data [Anto 13, Cara 14, Chiu 15] the research on finding suitable methods for circular scan segmentation is still ongoing [Golz 11, Kaba 15]. Almost every work published does not consider the quality of the data or explicitly excludes bad quality scans from the evaluation dataset [Ishi 05, Tan 09]. Varying image quality is considered in [Somf 07, Dufo 13, Chiu 15]. The only other method known to the author that does not exclude bad quality images and has glaucomatous scans in its evaluation is [Rath 14]. The authors suggest that the increased error rate of their segmentation in progressed glaucomatous eyes relates to their use of a shape model and less scan quality, but do not evaluate the influence of the quality explicitly. The segmentation evaluation database we utilize allows a differentiation of the influence of bad quality and the glaucoma disease on the segmentation result.

The first works published on glaucoma classification from OCT data could only use datasets with samples in the number of hundreds due to the novelty of the OCT machine in eye clinics and the resulting the lack of data [Burg 05, Huan 05]. The same holds for the preliminary work by the author regarding the glaucoma classification topic from OCT data published at a conference [Maye 09]. Modalities that are of common use in eye clinics for a longer period of time, e.g. visual field test, allow classification databases in the number of thousands [Wrob 09, Gold 05]. The only OCT image database for automated glaucoma detection comparable in size to ours was used in [Bask 12]. They excluded bad quality scans. Garcia-Martin et. al [Garc 12] also reject bad quality scans with a HE quality index below 25db in their work on multiple sclerosis (MS) detection, but explicitly mention in their discussion that daily clinical practice does not always allow good quality scans. Our classification database reflects data from daily clinical practice. In particular, scans with a HE quality index below 20dB are included, which is the manufacturer's suggestion for scan acceptance [Belg 15], but may not be reached for all, especially elderly and glaucomatous, patients.

Chapter 3

Retinal layer segmentation

The glaucoma classification approach of this work is based on the retinal layer thickness measurements obtained from the positions of the boundaries in between retinal layers. These boundaries are detected by an automated retinal layer segmentation algorithm and optionally corrected for segmentation errors afterwards. This chapter presents the retinal layer segmentation algorithm and quantifies possible segmentation errors in an evaluation.

The segmentation algorithm is an extension of the work published in [Maye 10]. In this preceding work, only the RPE and the RNFL boundaries were detected. The present form of the algorithm detects 6 retinal layer boundaries. This change is not only an addition to the algorithm, but also affects the detection of the 3 layer boundaries from [Maye 10]. Parts of the algorithm description in Section 3.2 are taken from the former publication where no change has been made, others are supplemented or altered. The inner layer detection is new and has only been presented in the form of an abstract and poster at a conference [Maye 11]. The conference presentation aimed at volume, not circular scan segmentation. We will refer to that work once again in the outlook of this chapter in Section 3.6. The State of the art Section 3.1 is completely restructured compared to [Maye 10] and updated for the year 2016. The evaluation in Sections 3.3,3.4 and 3.5 is new in terms of the data utilized, the observer corrections and the evaluation measures.

3.1 State of the art

Since the first appearance of commercially available OCT systems, automated retinal layer segmentation algorithms were presented to objectively quantify the thickness of the retina or retinal layers. While segmentation algorithms are built into commercial systems and are refined over time, their designs remain undisclosed. In the following, we give a short review of the published research on retinal layer segmentation and mention only those works that are the most important. A complete overview is given in the Tables B.1 and following in the Appendix. They contain the retinal layer segmentation overview from [Maye 10] and extend it to all retinal layer segmentation methods known to the author up to the beginning of the year 2016. In the following, we do not mention the number of retinal layers segmented, which for recently published algorithms usually are between 5 and 10. Almost all algorithms can be

extended to segment more layers when the quality of the data is sufficient. Specific numbers of segmented layers are included in the table overview in the Appendix.

There are multiple possibilities to group the various algorithms in the field. One could for example focus on the data the algorithm is applied to, i.e. TD- or FD-OCT data, linear scans through the macula or circular scans around the ONH, and volume scans in the macula or ONH region. In our days, the FD-OCT systems have almost replaced TD-OCT in the clinics. Most 2D segmentation methods can be either applied to 3D data directly by a slice-by-slice application or reformulated for a real 3D approach. A more appropriate grouping of the research in the automated retinal layer segmentation field is by the mathematical method utilized. The majority of works published can be divided into the following groups: **Edge detection or intensity-based methods** that build on various pre-processing and regularization or post-processing steps. The presented algorithm falls into this category, too. The usage of **active contours or active appearance models** enables the inclusion of shape priors. The most dominant group of algorithms in recent years are graph-based algorithms, either using **dynamic programming** or **graph cuts** to find the segmentation result. Some algorithms use ideas from multiple algorithm groups. The methods developed over time:

Edge detection and intensity-based methods: Koozekanani et al. [Kooz 01] published the first automated retina segmentation algorithm for OCT scans. An edge detection approach was introduced that also takes the leading sign of the derivative into account to differ between rising and falling intensity along a depth profile. One major challenge of OCT data segmentation was already mentioned: The speckle noise that corrupts OCT images is non-Gaussian, multiplicative and neighborhood correlated (for more information on OCT speckle noise see [Schm 99, Grzy 10, Lee 11, Kiri 14]). It cannot be easily suppressed by standard software denoising methods. Ishikawa et al. [Ishi 02, Ishi 05] use gradient-based edge detection followed by an integrity check. The median filter used by Ishikawa et al. for pre-processing of the OCT image was replaced by anisotropic or complex diffusion in later works [Fern 05, Muja 05]. The post-processing of the initial segmentation was refined. Baroni et al. [Baro 07] detect outliers in the segmentation by the distance to an average constant line and assign the boundaries by a dynamic programming approach on detected edges. Tan et al. [Tan 08] and Fabritius et al. [Fabr 09] use a multiscale approach for iterative segmentation refinement. The inclusion of 3D neighborhood information was also proposed by Tan et al. [Tan 09]. Blood vessel detection and removal further enhances segmentation results [Chiu 10, Lu 10, Golz 11]. Up to now, methods based on pre-processing and edge detection are proposed [Niu 14].

Active contour or active appearance model: As mentioned, active contour methods have the advantage that model information may be included. Yazdanpanah et al. [Yazd 09, Yazd 11] proposed this by including a shape prior, the deviation of the segmentation from a circle, into the energy functional to minimize. Also included in the energy term are a measure to prefer regions of homogeneous intensities and a gradient measure. The segmentation was tested on scans of rat eyes. The drawback of the active contour methods is that a proper initialization for the contour optimization has to be given. Yazdanpanah et al. used manual initializations with few points on each boundary, which makes the method semi-automatic. Kajic et al. [Kaji 10] solve

the problem by an initial segmentation with adaptive thresholding. They include shape and texture information learned from training data into their active appearance model. Due to the use of the model shape information, the method is very robust to noise. A dynamic programming approach yields an initialization for the method of Mishra et al. [Mish 09]. K-means clustering delivered the segmentation initialization for Ghorbel et al. [Ghor 11]. Rathke et al. [Rath 14] were the first to include not only local shape information learned from training data, e.g. differences to a model made of layer positions, but global shape information. Their model is based on a mixture of Gaussians. This approach has several advantages: The shape term of the regularizer can be directly used as a discriminative feature for glaucoma detection and a quality index for the segmentation result can be derived from the model. But another drawback of the approaches utilizing model information also becomes clear: If the model is learned form healthy eyes only, the segmentation of glaucomatous eyes is more likely prone to error. Rathke et al. clearly point this out and propose future work to overcome this problem. Other works avoid the issue by evaluating only on data from normal subjects [Kaji 10, Ghor 11] in the case scans of humans are to be segmented.

Dynamic programming: Chiu et al. proposed a dynamic-programming-based approach [Chiu 10]. Graph weights are computed out of gradients and pixel distances, and the minimal path is found by Dijkstra's algorithm. A similar approach with slightly different graph weights, consisting of gradient information from two different scales, was proposed by Yang et al. [Yang 10] at the same time. While still multiple heuristical refinements and assumptions are used in the case of [Chiu 10], both publications show the strength of the method, i.e. a compact algorithm that delivers good results within small computation time and without model assumptions. The approach has been taken up by others [Ehne 14] and the groups around Chiu and Yang have further published works that enhance the original algorithms and show their applications on scans of diseased eyes [Yang 11, Srin 14a, Chiu 15]. However, the dynamic programming methods are B-scan based, as an extension of Dijkstra's algorithm to be utilized for a real 3D segmentation is not straightforward.

Graph cuts: The strength of minimum cut graph-based methods is the easy extension to 3D, therefore they are the most prominent methods used to segment retinal layers in volume scans. The first graph-cut segmentation algorithm for retinal layers was proposed by Haecker et al. [Haek 06] and further extended by Garvin et al. [Garv 08]. Weights in a graph are constructed by gradient, intensity, integrated intensity measures and known position boundaries. The segmentation is the surface or cut in the graph with lowest cost [Boyk 00, Boyk 01, Boyk 06]. Various enhancements to the original algorithm have been proposed, especially for the weights of the graph. Quellec et al. [Quel 10] improve the algorithm by incorporating texture features as graph weights and automatically detect abnormalities from the segmentation results. Dufour et al. [Dufo 13] include model information and shape constraints into the graph weights, which leads to good segmentation results on noisy images. To overcome heuristics, Antony et al. [Anto 13] use a machine-learning approach to design the entire cost function and graph weights for the graph-cut method. The method can be easily adapted to retinas from different species (mice, canines) but does rely on good training examples. A similar approach has been proposed by

Lang et al. [Lang 13], where the graph-cut segmentation utilized probabilities from a preceding pixel classification approach that classifies pixel for belonging to certain regions and boundaries. This method was further extended by Carass et al. [Cara 14], were the graph-cut segmentation holds as an initialization for a level set refinement. The level set method allows subpixel accuracy and therefore layers below 1 pixel in strength, i.e. missing layers may be detected. Furthermore Carass et al. introduce the concept of "flat space" as a new computational domain for OCT images that is learned from training data.

An approach that does not fall into the 4 categories above is segmentation by classification. Szulmowski et al. [Szul 07] used manually drawn regions inside a volume to train a classifier and segment the rest of the volume, thus it is a semi-automated method. Vermeer et al. [Verm 10] performed a pixel-wise classification to layers with support vector machines trained on separate data, followed by a level set regularization. This also avoids the use of heuristics at the cost of extremly high computation time.

Today, the goals of segmentation on OCT data expand beyond the retinal layers, e.g. the sklera/choroid boundary becomes another focus of work [Kaji 12, Alon 13, Tian 13, Chen 15].

While the evaluation of algorithms in early works was sometimes limited to visual inspection [Fern 05, Szul 07, Mish 09] or error marking [Ishi 02, Ishi 05, Haek 06], every recently published work shows the validity of the segmentation approach by a quantitative evaluation in comparison to manual segmentations or a gold standard derived from manual segmentations. However, most of the times only overall numbers over the complete or a subgroup dataset and no local error information, i.e. were the errors most likely take place in an image, are given, with the exception of [Ehne 14, Kaba 15]. Furthermore, the works that included datasets of glaucoma patients in their evaluation are limited [Ishi 02, Ishi 05, Tan 08, Tan 09, Verm 10, Golz 11, Kafi 13]. Glaucoma patient data offers challenges that scans of healthy subjects do not have: The thickness and shape of the RNFL may be altered, which invalidates models generated from healthy subjects. In some severe cases the RNFL may be even missing. Srinivasan et al. [Srin 14a] targeted missing layers, but only on mice data. Theoretically, the level set refinement with sub-pixel accuracy of Carass et al. [Cara 14] is able to detect missing layers. However, this was not evaluated for glaucoma patients. Also scan quality is most often not considered in evaluations, as low quality images are excluded in the evaluation by design [Lang 13, Cara 14], or the quality of the scans is not even mentioned. Only Somfai et al. [Somf 07] modeled operator errors and their influence on segmentations and Dufour et al. [Dufo 13] specifically target their approach on low quality images.

In the next Section 3.2, an automated segmentation of retinal layers mostly based on pre-processing and edge detection is presented, with the exception of the ONFL segmentation, where we use a discrete optimization of an energy functional, comparable with active contour approaches. The segmentation is a 2D approach working on circular FD-OCT scans, but it can be applied on 3D volume data, as it is shown in Section 3.6. The goal during the development of the algorithm was to make as few assumptions on the layer borders as possible. These few assumptions should be very general. Reliable application to pathological cases with the glaucoma disease

should be possible without changing any parameter, even in the case of a RNFL hole. Bad image quality should not lead to segmentation errors. The evaluation presented in Sections 3.3 to 3.5 includes data from glaucoma patients of different disease stages, good and bad quality images, and points out areas of higher segmentation error probability.

3.2 Automated segmentation method

The goal of the automated retinal layer segmentation algorithm presented is to segment 5 retinal layers or layer groups corresponding to 6 retinal layer boundaries. The denominations of all retinal layers are given in 1.3. The RNFL, GCL+IPL, INL+OPL, ONL+ELM and IP+OP+RPE are the layer groups to segment, of which the RNFL and GCL+IPL are of most importance to glaucoma diagnosis. At an early stage of glaucoma diagnosis with OCT, the RNFL was the main retinal layer for obtaining glaucoma parameters [Gued 03, Woll 05, Polo 08, Horn 09], with the GLC+IPL coming into focus recently [Taka 12, Begu 14, Mwan 14] as commercial retinal layer segmentation software is now able to segment this layer. The retinal layer boundaries corresponding to the 5 layers or layer groups are the ILM, outer retinal nerve fiber layer (ONFL), IPL/INL, OPL/ONL, ELM/IP and RPE boundaries. They are marked with consistent colors throughout this work, with exceptions clearly stated. The ILM is red, ONFL green, IPL/INL orange, OPL/ONL rose, ELM/IP yellow and RPE blue, as can be seen in Figure 3.3 g) or Figure 3.10.

The algorithm is built around a few general assumptions that hold for both normal and glaucomatous eyes:

- The intensity distribution along the A-Scans is such that the most reflecting layers are the RNFL on the inner side, and the RPE on the outer side of the retina.
- The RPE boudary is not disrupted and has a simple shape, i.e. the shape can be modeled by a polynomial of small degree.
- An adequate pre-processing followed by edge detection with regularization is sufficient for segmenting most layer boundaries except the ONFL.
- The inner layer boundaries IPL/INL, OPL/ONL and ELM/IP are to a large extent parallel to the RPE boundary.

As the focus of this work is the glaucoma disease and the data was collected for glaucoma research purposes, we expect most of the scans in the database to conform to the assumptions, but these may be violated on scans of subjects with other diseases or scans with operator errors.

All algorithm parameters were manually adapted by visually inspecting random sets of images from the database. The processing steps of the algorithm are shown in Figure 3.1 with visual examples of the steps provided in Figures 3.2 and 3.3. A scan from a glaucoma patient was chosen as an example. This scan is not included in the database. The full scan with the corresponding SLO image was already shown in Figure 1.2 c) and d). It shows an almost complete loss of the RNFL in the

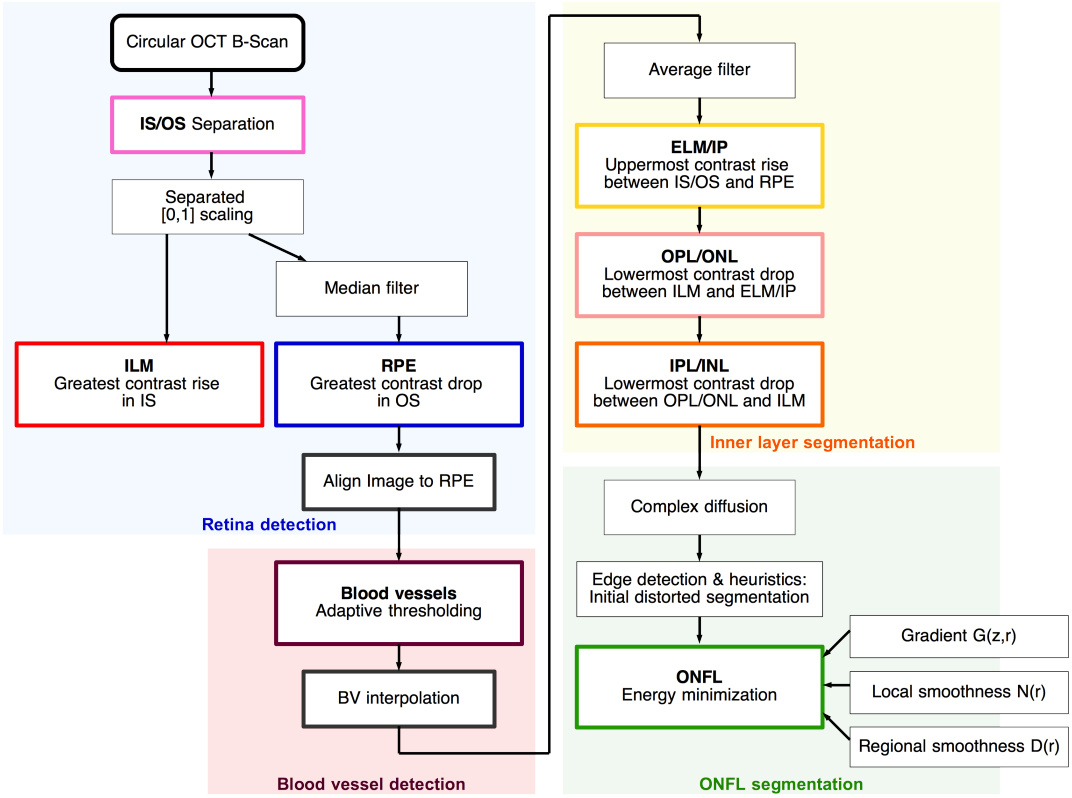


Figure 3.1: Algorithm overview. The input data is an circular OCT B-Scan around the optic nerve head. The retina detection is colored in blue, the blood vessel detection in dark red, the inner layer detection in orange and the outer nerve fiber layer detection in green. Important steps are marked with bold rectangles in the color used for the respective layer throughout the work. Layer smoothing steps are omitted in the overview for a clearer view.

transition between the inferior and temporal quadrant, while the other regions still have relative high RNFL thickness. The processing steps of the algorithm can be roughly divided into 4 groups: **Retina detection**, **blood vessel segmentation**, **inner layer segmentation** and **ONFL segmentation**. The steps are detailed in the following:

Retina detection: The input to the algorithm is a circular OCT B-Scan around the ONH with a diameter of 3.46mm captured with the HE Spectralis and with 768 A-Scans. To limit the search space for the retinal boundaries, first a separating line located inside the outer nuclear layer is identified. It splits the image content into the inner segment (IS) and the outer segment (OS) of the retina. The original image is blurred in the linear domain with a wide Gaussian filter (standard deviation $\sigma = 22$ pixels). The separating line is the minimum with the lowest intensity value between the two maximums with the highest intensity value in the Z-direction (see Figure 3.2 a) and Figure 3.4 a)). Note that the position of the separating line in Figure 3.2 a) has changed compared to the preceding work [Maye 10] as larger kernel sizes are used. The intensities of each A-Scan are scaled to $[0; 1]$ in the IS and OS

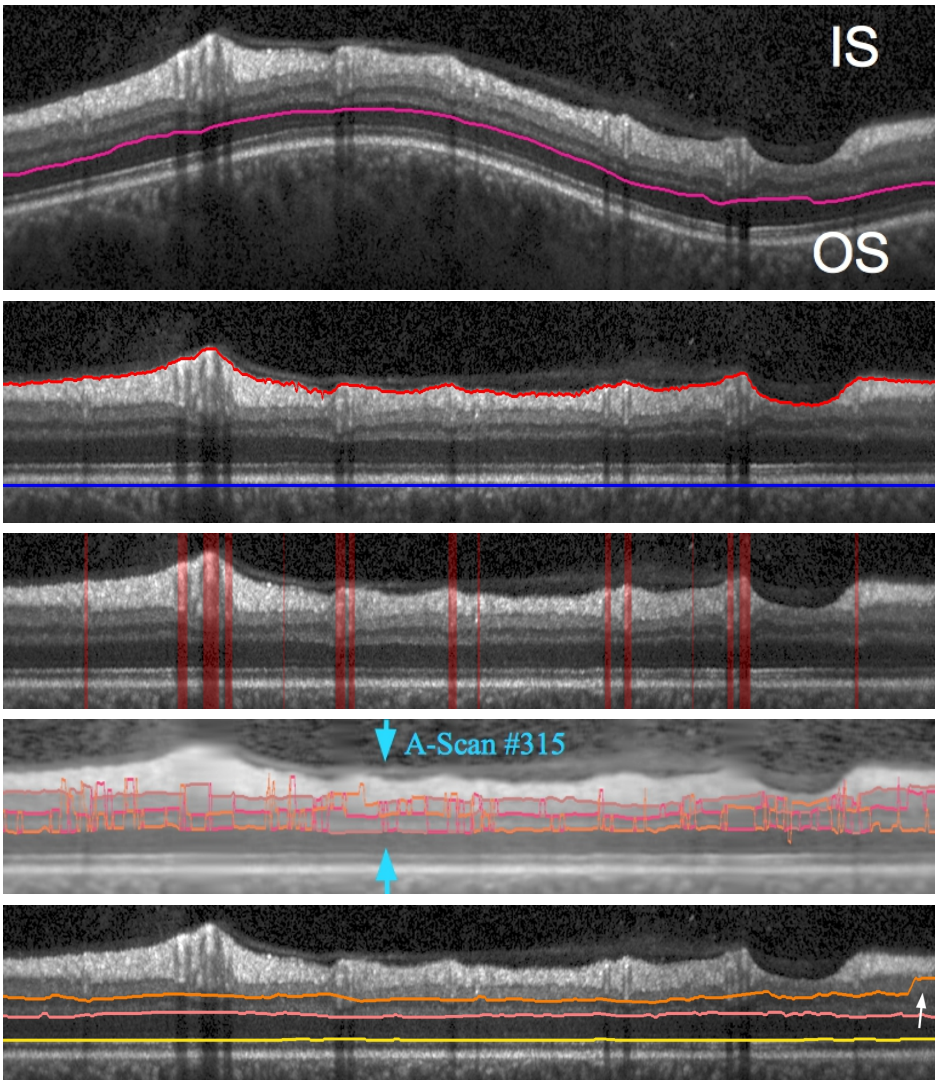
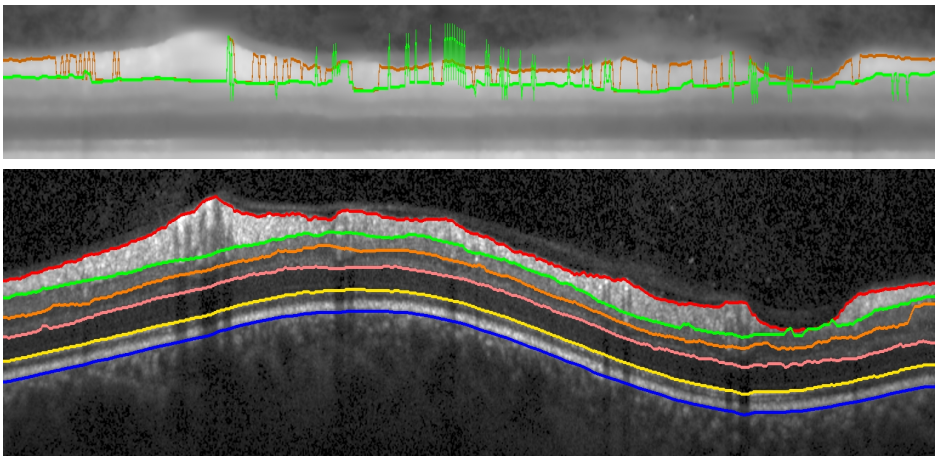


Figure 3.2: Processing steps of the retinal layer segmentation (first part). The example scan is of a glaucomatous eye with local nerve fiber layer loss. The original complete scan is shown in Figure 1.2 c). The algorithm step images are cut in vertical direction to show only the retina. (a) Separating line in the outer nuclear layer detected. IS: Inner segment of the retina. OS: Outer segment of the retina. (b) Inner limiting membrane (ILM) and retinal pigment epithelium (RPE) detected. A-Scans aligned such that the retinal pigment epithelium forms a constant even line. (c) Result of the blood vessel detection. Blood vessel positions are marked in transparent red. (d) Image with interpolated blood vessels and denoised with an average filter. 3 intensity drops in the inner segment of the retina are detected. Color assignment is by the order of the strength of the contrast drop in the A-Scan. The segmented OPL/ONL boundary before smoothing is the outermost contrast drop. (e) Segmentation result for the inner layers. At the right side of the image, there is a segmentation error in the IPL/INL boundary, marked with an arrow.



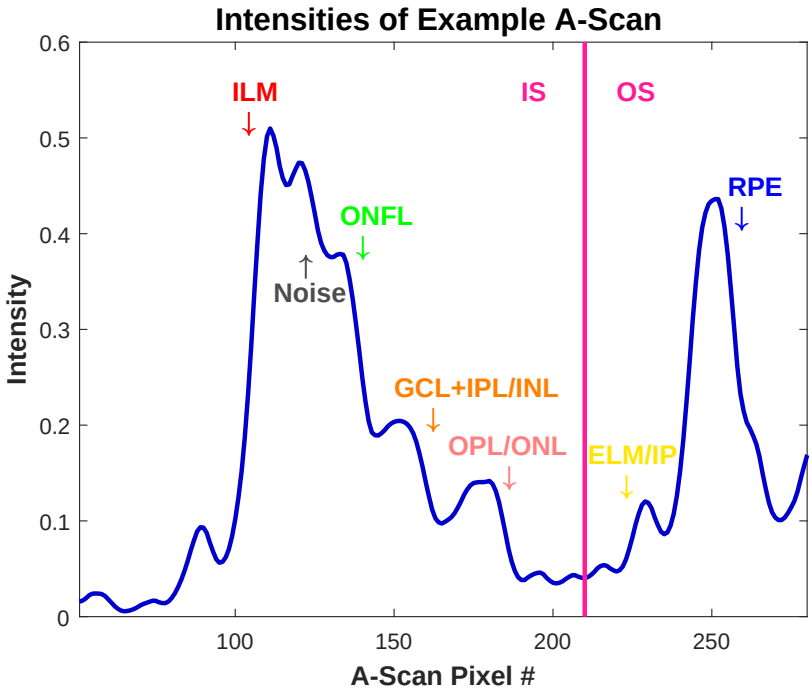
(f)

(g)

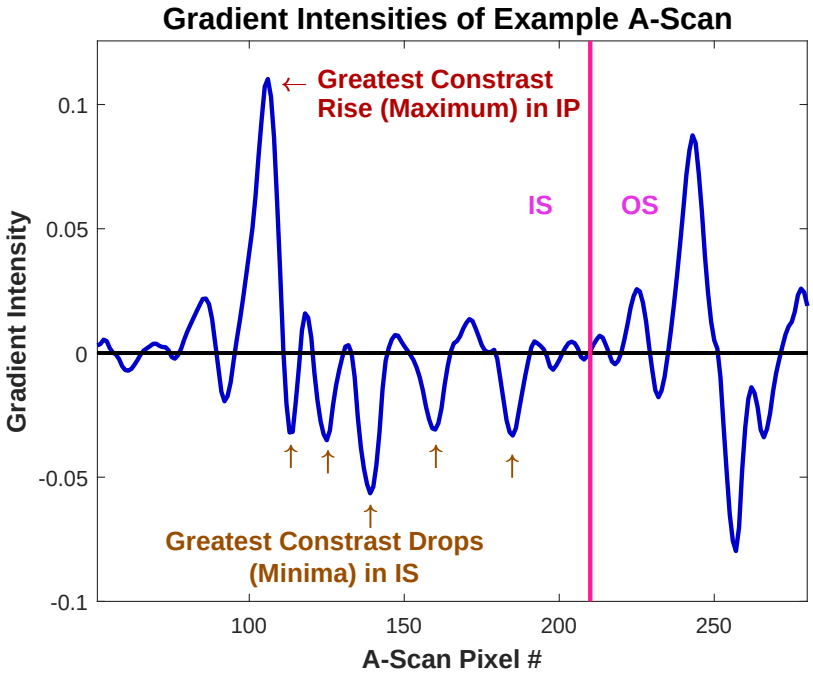
Figure 3.3: Processing steps of the retinal layer segmentation (second part). Further information on the example image see Figure 3.2. (f) Image with interpolated blood vessels and denoised with complex diffusion. It shows the initial distorted segmentation of the outer nerve fiber layer boundary formed by heuristic decisions (dark orange) and result after the energy minimization (green) before smoothing. (g) Final result with all segmented boundaries painted on the original (not flattened) image.

separately. The ILM is then set to the greatest intensity rise in the IS, i.e. the second derivative along the A-Scan is 0 and the gradient is highest. A contrast rise is an increase in intensity seen from the inner Z-direction. For the RPE segmentation, a rough speckle noise removal is used. A 2D median filter (size 5 in Z and 9 in R-direction) is applied twice, as proposed by Ishikawa et al. [Ishi 05]. The RPE is the greatest intensity drop in the OS.

The ILM and the RPE segmentation are smoothed with a procedure that is common to all the layer boundary results, also the layer boundaries following later in the segmentation process. The single line smoothing steps are displayed in Figure 3.5. First a median filter is applied. Then distant line segments, i.e. line segments that do not have any other line segment in a defined neighborhood around their left or right end, are removed. Additional outliers are detected by fitting a polynomial to the line and removing distant line segments afterwards. Then short segments are removed. The holes in the line resulting from the removal of points are filled with linear interpolation and finally a second median filter and Gaussian smoothing are applied. The parameters for the smoothing process differ slightly from layer boundary to layer boundary and some smoothing steps may be omitted. In the case of the RPE, the median filters have the width 9, no distant line segments are removed, the polynomial has the degree 6, the line segments must have at least the length 6 and the Gaussian filter has a standard deviation of $\sigma = 3$ pixels. In the next Section 3.3, it will be explained where the source code for the algorithm and example algorithm parameters are publicly available. Therefore, and as the smoothing parameters for each layer boundary differ only slightly, the enumeration of all smoothing parameter values has been omitted for the rest of the algorithm description to favor readability.



(a)



(b)

Figure 3.4: Intensity plot along an A-Scan (a) and corresponding derivative (b). The A-scan # 315 of the example image with an average filter applied (Figure 3.2 (d)) is shown. It is cropped to the retina region. The intensity rise at the ILM and ELM/IP boundary, as well as the intensity drops at the ONFL, IPL/INL, OPL/ONL and the RPE boundary are marked. The separation line between the inner and outer segment of the retina is also drawn. In b) the 5 greatest contrast drops in the inner segment (IS) are marked. The first 2 in the Z-direction correspond to remaining speckle noise in the retinal nerve fiber layer, the others to the layer boundaries to segment.

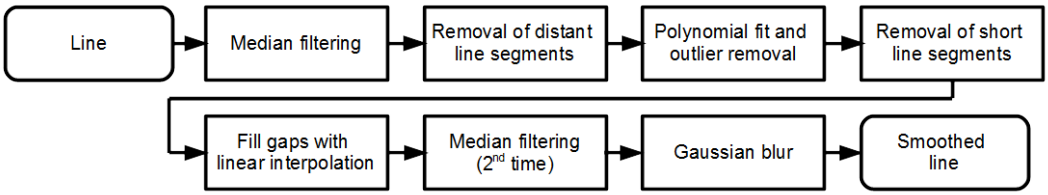


Figure 3.5: Line smoothing algorithm steps. All segmented layer boundaries are smoothed by using these processing steps in the displayed order. Parameters may change and certain steps may be excluded for particular layer boundaries.

After the smoothing, the A-Scans of the original unprocessed image are aligned such that the RPE forms a constant even line (see Figure 3.2 b)).

Blood vessel detection: For the blood vessel (BV) detection, we use the fact that the motion inside BVs casts a shadow along the remaining A-Scan depth. The image intensities are single square-rooted. This intensity domain proved to provide the best results for BV detection. The BV positions are determined by adaptive thresholding along the RPE. A layer of 8 pixels above the RPE is summed along the Z-direction to form a RPE intensity profile. The average of this profile is computed in a 61-pixel wide window. If the value of the profile in the middle of the window is lower than 0.7 times the average, it is marked as a blood vessel. An example result is shown in Figure 3.2 c). As the size parameter of the average window and the threshold are fixed, some large vessels above 12 pixels in width are not detected. However the results are sufficient for the next segmentation steps. After the BV detection, the RPE on BV positions is invalidated and linearly interpolated. The linear interpolation over BV regions is also incorporated in the smoothing of all following segmented layer boundaries.

For the remaining processing steps, the image intensities are double square-rooted from the original linear domain. The A-Scans of the image of BV positions are interpolated in R-direction from their direct neighbors. BVs with a diameter of more than 4 pixels are enlarged by a factor of 2 in width for this purpose, as the BV regions for large BV do span more A-Scans than the casted BV shadow. Slightly different but similar approaches to segment the layers over BV regions have been proposed by [Chiu 10] and [Golz 11].

Inner layer segmentation: The image is denoised with a mean filter (size 3 in the Z- and 7 in the R-direction). The averaging is a sufficient denoiser for the inner layer segmentation. The 2 highest intensity rises, i.e. with the largest gradient, between the IS/OS separation and RPE are detected, which should correspond to the ELM/IP and OP/RPE boundary. The ELM/IP is set to the innermost of these 2 intensity rises.

Afterwards, we find the 3 highest intensity drops between the ILM and the IS/OS separation. In Figure 3.2 d) these 3 highest intensity drops are drawn. The colors are assigned by gradient strength in the A-Scan. Note that the gradient strength does not correspond to a valid boundary assignment. The OPL/ONL is the outermost of the intensity drops and the IPL/INL the middle one. The inner boundaries are regularized by the distance to the RPE. If a boundary position to the RPE deviates a

more than the average boundary distance to the RPE, i.e. outside a certain threshold (3 pixels for the ELM/IP and OPL/ONL, 8 pixels for the IPL/INL), it is invalidated. There is one exception to this rule: This thresholding is not applied to the IPL/INL in the temporal quadrant, as the IPL+ONL layer group in general increases in thickness in this quadrant. The inner layer boundaries are smoothed with the steps shown in Figure 3.5.

ONFL segmentation: Average filtering as denoising followed by a simple edge detection will not give promising results for the ONFL, even when the ILM and IPL/INL positions are known. This holds especially in the cases of general low image quality, glaucoma patients with a complete local loss of the RNFL, or normal subjects with a very thick RNFL. For the last two challenges, a state-of-the-art pre-processing with sophisticated denoising as proposed by Fernandez et al. [Fern 05] and Mujat et al [Muja 05] is also insufficient. A neighborhood integrity check as mentioned in [Ishi 05] might not be able to cope with a jump of the segmented border in a whole region to a more contrasty outer layer border. Assumptions on the layers, as made by Garvin et al. [Garv 08], may be violated in pathological cases, or parameters have to be adapted for either normal subjects or glaucoma patients. Our approach is as follows:

The original image is denoised with complex diffusion (see Gilboa et al. [Gilb 04]) as proposed by Fernandez et al. [Fern 05]. Our implementation is not based on the traditional time-marching implementation, but uses lagged diffusivity [Voge 96, Chan 99]. The code of the algorithm can be downloaded from the homepage of the pattern recognition lab of the FAU (`http://www5.cs.fau.de`) from the author's personal page. The timestep parameter was set to 13, while the σ_{CD} parameter, that controls which gradients are detected as edges, is directly estimated from the image:

$$\sigma_{CD} = \frac{1}{3} \underset{r,z}{std}(|I(r,z) - I_{medfilt}(r,z)|); \tag{3.1}$$

$I(r,z)$ denotes the original image matrix, $I_{medfilt}(r,z)$ is the original image on which every A-Scan is filtered with a median filter of width 7. The $\frac{1}{3}$ is a heuristic weighting factor. The computation of the standard deviation of all pixels is abbreviated by $\underset{r,z}{std}(...)$. The noise estimate does not correspond to a physically meaningful noise measurement on the OCT data, but it has proven to adapt to the different noise levels and qualities of the OCT B-Scans by visual inspection.

If they are present, the two largest contrast drops are detected between the ILM and IPL/INL. Actually, only one layer boundary with falling contrast should lie between the ILM ans IPL/INL boundary, namely the ONFL. To derive an initialization for the following energy minimization, the following heuristic is used: We choose the lowermost contrast drop and blur the line by applying a median (width 5 pixels) and Gaussian (sigma 15 pixels) filter. If only one contrast drop is detected, we trust this initialization. In all other regions, the initialization is set to the ILM, as we do not assume a correct segmentation there. Either no contrast drop is detected, in which case a complete RNFL loss is most likely, or remaining speckle noise within the RNFL produced a second contrast drop. This method delivers a very distorted initializa-

tion for the segmentation shown in Figure 3.3 f) in dark orange. To improve it, we formulate an energy-minimization-based approach:

$$\begin{aligned} E(r, ONFL(r)) = & G(r, ONFL(r)) \\ & + \alpha N(ONFL(r-1), ONFL(r), ONFL(r+1)) \\ & + \beta D(r, ONFL(r)); \end{aligned} \tag{3.2}$$

$$\underset{ONFL(r)}{minimize} \sum_r E(r); \tag{3.3}$$

$ONFL(r)$ gives the Z-position of the boundary at A-Scan position r. $E(r)$ is the energy at A-Scan r that is minimized. It consists of three terms. Two factors, α and β weight these terms. They are set to $\frac{1}{100}$ and $\frac{1}{1000}$ respectively. The first term, $G(r, ONFL(r))$ is the gradient of A-Scan r at depth $ONFL(r)$. As the ONFL is a contrast falloff in Z-direction, this scalar gradient should have a negative sign with an absolute value as high as possible. $N(ONFL(r-1), ONFL(r), ONFL(r+1))$ is a first smoothing term involved in the formulation that ensures that there are no high jumps in the border, while allowing for some edges. It is the sum of the absolute differences in depth $ONFL(r)$ of A-Scan r and its neighbors:

$$\begin{aligned} N(ONFL(r-1), & ONFL(r), ONFL(r+1)) = \\ & |ONFL(r-1) - ONFL(r)| \\ & + |ONFL(r+1) - ONFL(r)|. \end{aligned} \tag{3.4}$$

The second smoothness term $D(r, ONFL(r))$ works on a more global scale. It is motivated by the observation that, when the A-Scans are aligned for an even RPE, the ONFL is part-wise almost even, too. In Baroni et al. [Baro 07] the distance to a constant line along the whole B-Scan was taken as a smoothness term. We take this idea up and extend it. The RNFL should not be as constant as possible over the whole B-Scan, but within certain regions. To avoid using arbitrary positions on the image, the regions between blood vessels are used. $D(r, ONFL(r))$ is therefore the distance to the average height of the segmented boundary between two blood vessels:

$$\overline{BVR}_k = \frac{1}{\#BVR_k} \sum_{r \in BVR_k} ONFL(r); \tag{3.5}$$

$$D(r, ONFL(r)) = ONFL(r) - \overline{BVR}_k; \tag{3.6}$$

BVR_k is a region between two blood vessel centers with identifier k. Depending on the blood vessel segmentation, there can be an arbitrary number of such regions. To keep the number of regions reasonable, very small blood vessels with a diameter below 4 pixels are ignored for the blood vessel center computation. $\#BVR_k$ is the number of A-Scans in the blood vessel region k. $\overline{BVR}_k$ is the mean depth of the ONFL within in the blood vessel region k.

The energy formulation in Equation 3.2 is solved iteratively. For each A-Scan r, the energy above and below the current position is computed and the ONFL position is moved by one pixel in the direction with decreased energy, until an iteration limit is reached or no change happens anymore. The algorithm is summarized as pseudo

code in Figure 3.6. The result is painted in green in Figure 3.3 f). It can be observed that only a local minimum of the energy formulation is found. Some of the ONFL positions get stuck on wrong layer boundaries or speckle noise due to the gradient weight overpowering the smoothness weights for the direct neighboring positions. The simple optimization that moved the boundary only by one pixel in each step is suboptimal. It could be replaced by more elaborate discrete optimization techniques, e.g. simulated annealing or genetic optimization. However, after smoothing the result similar to the other layers with the steps given in Figure 3.5 the outliers vanish and the results have proven to be adequate by (1) the evaluation in the preceding work [Maye 10] and (2) by inspecting randomly selected images from the current database during the process of the manual algorithm parameter adjustment. At A-scan positions r were no contrast drops was detected between the ILM and IPL/INL at all, the ONFL is set to the ILM before the smoothing process, i.e. a complete RNFL loss is assumed. The result of the algorithm transformed back to the original (not flattened) image is shown in Figure 3.3 f).

The code was written in Matlab (Mathworks, Inc.) and is freely available (see upcoming Section 3.3). The average runtime was measured using a MacBook Pro, Intel Core 2 Duo, 2,66 GHz with 4GB main memory in the preceding work [Maye 10] and did not change to a large extent. The computation time of the additional steps is compensated by the possibility to run the complex diffusion filter on a more restricted area. Only one processor core was utilized. The average runtime for each B-Scan is around 20s. It did not differ substantially from normal subjects to glaucoma patients, or between images of good to bad quality. Included in the average runtime are the loading of the data from the hard disc and the storing of the results. The main part of the time spent is used for the complex diffusion filter. Diffusion filters can be massively optimized for speed by using multi-grid technologies. Other parts of the algorithm can also be sped up by implementing them in a more efficient language, for example C++. However, algorithm speed is not the focus of this work. Therefore, we did not optimize for computational efficiency. In fact, some of the computations are carried out multiple times for a more structured, easy-to-read code.

```
iteration = 1;
somethingMoved = true;
while iteration < MAXITER and somethingMoved do
  somethingMoved = false;
  for all r do
    Compute E(r, ONFL(r));
    Compute E(r, ONFL(r) + 1);
    Compute E(r, ONFL(r) − 1);
  end for
  for all r do
    if E(r, ONFL(r) + 1) < E(r, ONFL(r)) then
      if E(r, ONFL(r) − 1) < E(r, ONFL(r)) then
        if E(r, ONFL(r) + 1) > E(r, ONFL(r) − 1) then
          ONFL(r) = ONFL(r) - 1;
          somethingMoved = true;
        else
          ONFL(r) = ONFL(r) + 1;
          somethingMoved = true;
        end if
      else
        ONFL(r) = ONFL(r) + 1;
        somethingMoved = true;
      end if
    else
      if E(r, ONFL(r) − 1) < E(r, ONFL(r)) then
        ONFL(r) = ONFL(r) - 1;
        somethingMoved = true;
      else
        Do nothing;
      end if
    end if
  end for
  iteration = iteration + 1;
end while
```

Figure 3.6: A simple iterative scheme for minimizing the energy formulation in Equation 3.2. The position of the ONFL at A-Scan r $ONFL(r)$ is moved one pixel up or down if the energy at this A-Scan is decreased thereby. First all the energy above, at and below the current ONFL position is computed, afterwards the decisions to move the position are made. The iteration limit MAXITER is set to 200. Variable and function names are chosen such that they describe their content or behaviour.

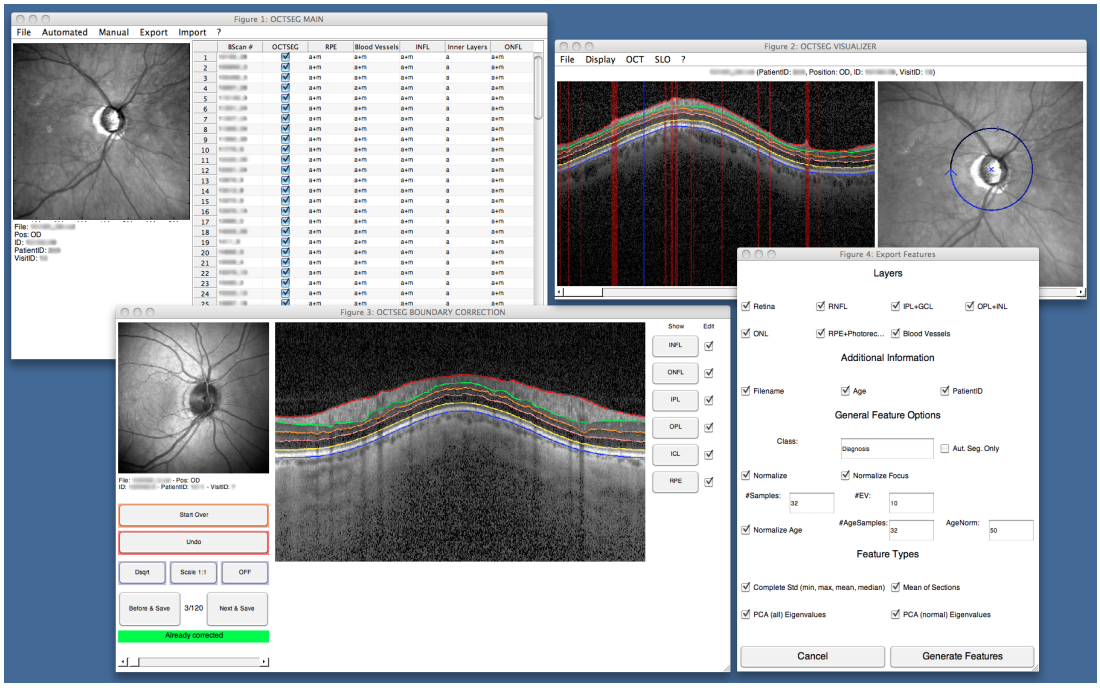


Figure 3.7: Screenshot of the OCTSEG tool. The main window on the left upper side controls the automated segmentations and opens the tools for manual correction, visualization and export. On the upper right side, the visualization window is displayed. It renders the segmented layer boundaries on to the OCT image and displays the corresponding SLO image. Various display modes are available, e.g. visualization of en-face views for OCT volumes. On the lower left side, the manual correction window is shown. Layers boundaries may be corrected by free hand drawing. Display modes for contrast change, scaling and zooming support the correction. On the lower right side, the feature export window is shown that enables feature export for the use in a classification system.

3.3 Evaluation construction

A graphical user interface (GUI) was written in Matlab (Mathworks, Inc., Natick, Massachusetts, USA) with the intention to display the segmentation results and to allow for a manual correction. The software is called OCTSEG (OCT segmentation and evaluation GUI). The abilities of OCTSEG expanded over time. The automated segmentation for circular scans and also volumes was included and export functions written, e.g. to export the layer data in text tables, feature export and batch image processing. A screenshot with some further explanation on the GUI is displayed in Figure 3.7. The tool has proven its usefulness in publications from various groups [Laem 11, Torn 11, Feno 13, Balk 13, Kola 13, Balk 14, Odst 14]. The program was published in compiled form on the homepage of the pattern recognition lab of the FAU (`http://www5.cs.fau.de`) on the author's personal page and under "Free Software". Since the beginning of 2016, the source code is freely available from the same address.

In the manual correction mode of OCTSEG, various display modes are available, for example the image contrast can be adjusted and layer borders can be switched

on and off. The image can be scaled and zoomed. The simultaneously acquired SLO image of the Spectralis is displayed in addition. Corrections to the automated segmentation can be made by free-hand repainting of the segmentation borders. No region has to be selected. The method allows correcting even for smallest errors.

We decided to let observers not draw complete manual segmentations, but to correct errors of the automated segmentation. Both methods have their advantages and disadvantages. The main advantage of a complete manual segmentation is that the observer is not biased by a given segmentation, and is therefore forced to look at each part and boundary of the image. When manual correction is applied, errors may be overseen by the observer, may it be due to fatigue or time pressure. However, the unbiased manual segmentation leads to offsets in the segmentation lines, even within one image and especially between observers. This means that lines or line segments are shifted by a few pixels distances. The human observers do not pay attention to this constant shift. This offset has to be taken into account in the evaluation. To our knowledge, only the group of Chiu and Farsiu [Chiu 10, Srin 14a, Chiu 15] have taken the offset of manual segmentations into account in their evaluation. Manual correction avoids a global bias of the observers.

The second reason for the correction of errors is that a complete manual segmentation of the segmentation evaluation dataset would be too time-consuming. For the manual correction alone, about 5 to 6 hours of pure correction time is the minimum for the the 120 scans of the segmentation evaluation dataset, not taking necessary breaks into account. Pure manual segmentations are therefore often restricted to prominent layer boundaries and the number of observers is limited, in most cases to 2 (see the retinal layer segmentation literature overview Tables B.1 and following). In addition, to speed up the manual segmentation process, methods like spline drawing are used, which do not allow to follow a layer boundary precisely. We decided for a manual correction of segmentation results which allowed us to have the automated segmentations corrected by 5 independent observers.

The observers included 4 computer scientists and 1 physicist, all working in the field of ophthalmic imaging and therefore familiar with OCT data. All but one had at least two or more years of experience in the field. The remaining observer had one year of experience with ophthalmic imaging. The observers were provided with an introductory manual that explains the use of the OCTSEG software for this task and points out helpful tools and procedures. Furthermore, instructions on the positioning of the layer boundaries were given by the author, with an example image not included in the segmentation evaluation dataset for further clarification. Special instructions were given for the IPL/INL and ONFL layer boundary in blood vessel regions and for images with poor quality. These special instructions were as follows (literally copied from the instruction text):

- "*(...), the IPL/INL boundary needs some additional explanation: In regions outside blood vessels it is positioned above a thin dark layer. (...) I suggest that you use the following rule: If the thin dark layer is visible, use it as a reference and follow it above. If the contrast is very low, or the blood vessel is positioned very deep, follow it on both sides until you are near the blood vessel shadow and then interpolate over the blood vessel region.*

- *For the ONFL boundary in blood vessel regions: If you can see/feel a clear boundary, follow it (e.g. the large blood vessel region on the left of the example scan). If there is no clear distinction between what is blood vessel and what are nerve fibers, interpolate over the blood vessel region (e.g. the middle blood vessels in the example image).*

- *If the quality of an image is so bad that even you as human cannot see the boundaries: Use your intuition.*"

The observers fulfilled their tasks independently. The automated segmentations were carried out on the full classification dataset with the same parameters for every scan. No segmentation was reperformed with different parameters. In addition to the 5 observers correcting the data from the segmentation evaluation dataset, the author corrected the automated segmentations of the full classification dataset, including the correction of the blood vessel segmentation, which was not performed by the observers. While it would be desirable to have multiple observer corrections on the full classification dataset, this was not possible due to time issues: The correction of all 1024 scans took round about two weeks.

From the manually corrected observer segmentations, a gold standard (GS) to evaluate the automated segmentations is constructed. Ideally the observer behavior, i.e. where and how corrections are made, should be identical among the observers given their experience with OCT data and the detailed instructions. This is by far not true, as will be shown in Section 3.4. The following scheme is used for the construction of a gold standard for each A-Scan position and layer boundary:

- At least 2 observers must have corrected the boundary at the specific A-Scan position, otherwise the automated segmentation is taken over as the gold standard.

- If exactly 2 observers corrected the boundary, the layer position nearer to the automated segmentation is preferred.

- When 3 or more observers corrected the boundary, the median position among the observers defines the gold standard.

This scheme ignores outlying corrections and observer opinions, and the resulting gold standard is not an average, but lies exactly on a position where at least one observer set it. The properties of the observer and the author corrections and their relationship to the GS are detailed in the following Section 3.4. The automated segmentations are evaluated with the help of the GS in Section 3.5. Contrary to the preceding work [Maye 10], where only the RNFL thickness was of interest and was therefore evaluated, a direct comparison of the layer boundary positions is carried out for each of the layers. Overall values averaged over all boundaries are only given where they make sense, as the results differ much between the layer boundaries. The measures are computed from layer boundaries in a [μm] scale:

$$L(r) = L_{pixel}(r) \cdot Scale_Z \tag{3.7}$$

where $L(r)$ is a retinal layer boundary position in μm from the inner side of the OCT scan at A-Scan position r, $L_{pixel}(r)$ the boundary position from the inner side of the scan in pixels and $Scale_Z$ the pixel spacing in the Z-direction in μm/pixel. $Scale_Z$ is 3.86μm/pixel in the case of the Spectralis. The evaluation measures computed from the layer positions are proposed in the respective Section 3.4 or 3.5 in conjunction with the results and a discussion, if appropriate or necessary.

3.4 Observer evaluation and discussion

A first measure to quantify observer behavior is to have a look at how much A-Scan positions each of them corrected, i.e.:

$$CF_{o,l} = \frac{100\%}{\#R * \#DB} \sum_{i \in DB} \sum_{r \in R} \chi(L_{o,i,l}(r) \neq L_{autom,i,l}(r)) \tag{3.8}$$

were $CF_{o,l}$ is the corrected fraction (CF) of A-Scans for a specific observer o and layer boundary l. DB is the set of images in the segmentation evaluation database, i a specific image from this database, and $\#DB$ the number of images in the database, which is 120 for the full segmentation evaluation database. R are all the possible A-Scan positions $\{1, 2, 3, \ldots, 768\}$, a subset of $\mathbb{N}$. Herein r is a specific A-Scan position and $\#R$ the number of possible A-Scan positions, which is 768 for our database. The indicator function $\chi(...)$ is 1 if the expression inside the brackets is true and 0 otherwise. $L_{o,i,l}(r)$ is the layer boundary position of layer boundary l as seen by observer o on image i at A-Scan position r. $L_{autom,i,l}(r)$ is the layer boundary position of layer boundary l from the automated segmentation on image i at A-Scan position r. For the following equations, $L(r)$ with one or multiple subindices always denotes a layer boundary position at A-Scan position r, e.g. from a specific image i, a specific layer l, a specific observer o, the gold standard GS, the automated segmentation $autom$ or the author's correction $author$.

The $CF_{o,l}$ measure tells us how much of the automated segmentations of layer l the observer o has "touched", i.e. moved from the position of the automated segmentation. A similar measure can be computed for the gold standard and the author's correction by replacing $L_{o,l}(r)$ with $L_{GS,l}(r)$ or $L_{author,l}(r)$ respectively. If all the observers would have seen the same layer positions as erroneous, $CF_{o,l}$ would be the same for each observer o. However, the corrected fraction of A-Scans differs from observer to observer, as the table of the CF results Table 3.1 tells us. In general, the RPE and ELM/IP boundary had to be corrected only to a minor extent with the CF values of all observers, the GS and the author being equally low. The ONFL and IPL/INL had to be corrected the most. While most of the numbers, including the GS and the author's CF fall at least into the same range, there are noticeable exceptions: Observer 4 corrected the ONFL boundary clearly less than the others. It has to be noted that observer 4 was the one with the least experience in ophthalmic imaging. Observer 5 performed only minor corrections to the ILM, but the most on the ONFL. The author's CF always lies between those of the observers, except

Boundary	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	GS	Author
ILM	7.5	4.1	7.2	2.2	0.6	4.3	3.5
ONFL	10.8	8.9	14.3	5.6	16.3	13.3	13.8
IPL/INL	9.7	14.7	18.3	14.6	12.6	16.2	12.1
OPL/ONL	3.8	4.3	4.2	6.3	1.2	3.7	6.5
ELM/IP	1.1	2.2	1.4	1.7	1.4	1.7	1.1
RPE	1.1	1.3	1.1	1.9	1.1	1.3	1.2

Table 3.1: Fraction of touched automated segmentation results in the manual correction per layer boundary l for the observers (Obs.)($CF_{o,l}$), the gold standard (GS)($CF_{GS,l}$) and the author ($CF_{author,l}$) in percent [%]. "Touched" in this sense means that the layer boundary was moved from the automated segmentation for the respective A-Scan.

Boundary	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	GS	Author
ILM	1.24	1.06	1.27	0.92	0.80	1.07	1.02
ONFL	2.79	1.97	2.85	1.32	4.21	2.84	2.98
IPL/INL	2.43	2.39	2.76	2.19	2.61	2.56	2.74
OPL/ONL	0.72	0.66	0.68	0.78	0.37	0.61	1.01
ELM/IP	0.49	0.63	0.54	0.55	0.59	0.58	0.53
RPE	0.51	0.50	0.48	0.55	0.49	0.51	0.52

Table 3.2: Mean absolute difference of the manual correction of the observers (Obs.)($ODA_{o,l}$), the gold standard (GS)($ODA_{GS,l}$) and the author ($ODA_{author,l}$) to the automatically segmented layer boundaries in [μm]. These values hold as a quantitative measure of how much correction the observers, the gold standard and the author applied to the automated segmentation results.

the OPL/ONL boundary, where the CF is greatest among all, closely followed by Observer 4. Looking at the CF numbers, the GS construction rule that at least 2 observers had to correct a certain position is reasonable. The single large corrected fractions of Observer 5 for the ONFL, Observer 3 for the IPL/INL and Observer 4 for the OPL/ONL get less weight, but the CF of the GS segmentation is always more then the average of the observers' CF values.

Next we will quantify how much the observers corrected in terms of distance to the automated segmentation:

$$ODA_{o,l} = \frac{1}{\#R * \#DB} \sum_{i \in DB} \sum_{r \in R} |L_{o,i,l}(r) - L_{autom,i,l}(r)|; \tag{3.9}$$

The mean absolute difference of the observer corrections to the automated segmentation ODA is a measure on how much correction was applied. Not only the number of corrected positions count, but also the distance the layer boundary was moved. Again, this measure can also be computed for the GS and author. The results, shown in Table 3.2, are in general more similar to each other for the different observers than

for the CF values, which suggests that a large corrected fraction CF shows the correction of minor errors by the observer. The only layer boundary that almost keeps the CF relationships between the observers is the the ONFL layer boundary. Observer 5 has a 3 times larger ODA value for this layer than Observer 4, roughly the relationship of the CF values. The GS lies in the middle of the observers for the ODA values, which is valid. The author applied roughly the same amount of correction than the GS, with the exception of the OPL/ONL, were he applied the most compared to the observers.

After quantifying the correction applied to the automated segmentation by the CF and ODA measures, we will have a look at how the observers' corrections relate to each other. Two measures can be computed, the mean absolute inter-observer difference IOD and the standard deviation $STDO$ among the observers. The mean absolute difference of one observer to the others for one specific image i is defined by:

$$IOD_{o,l,i} = \frac{1}{\#R * (\#O - 1)} \sum_{\hat{o} \in O, \hat{o} \neq o} \sum_{r \in R} |L_{o,i,l}(r) - L_{\hat{o},i,l}(r)| \tag{3.10}$$

where O is the set of observers and $\#O$ the number of observers, in our case 5. The mean absolute difference of one observer to the others for the segmentation evaluation database is then:

$$IOD_{o,l} = \frac{1}{\#DB} \sum_{i \in DB} IOD_{o,l,i}; \tag{3.11}$$

The inter-observer difference IOD_l for a specific layer is the mean of the $IOD_{o,l}$ values over all observers:

$$IOD_l = \frac{1}{\#O} \sum_{o \in O} IOD_{o,l}; \tag{3.12}$$

Beside this mean, also the standard deviation of the $IOD_{o,l}$ can be computed.

The author's mean absolute difference to the observers is given by:

$$IOD_{author,l} = \frac{1}{\#R * \#DB * \#O} \sum_{o \in O} \sum_{i \in DB} \sum_{r \in R} |L_{o,i,l}(r) - L_{author,i,l}(r)|; \tag{3.13}$$

The results for the $IOD_{o,l}$,IOD_l and $IOD_{author,l}$ values are given in Table 3.3. The $IOD_{o,l}$ values tell us how far the opinion of a single observer differs from the others. Except the ONFL values from Observer 5, they are very similar. We can conclude that Observer 5 corrected the ONFL in a different way than the other observers, despite the instructions given. The relationship of the IOD_l with its standard deviation and the $IOD_{author,l}$ gives insight in how much the author's corrections resemble a standard observer. The $IOD_{author,l}$ lies within the $IOD_l \pm std(IOD_{o,l})$ range except for the OPL/ONL and ELM/IP. The fraction of positions corrected for the ELM/IP is very small, so the only noticeable difference of the author to a standard observer is the way the OPL/ONL correction took place. Compared to the corrections performed on the ONFL an IPL/INL, the number of corrections for the OPL/ONL is also small. Therefore we conclude that the author's corrections are a valid representation of a single standard observer opinion for the bulk of manual corrections in the evaluation. If in clinical practice the OCT system operator decides to correct the retinal layer boundary segmentation, this is a single observer opinion. Corrections in daily clinical

Boundary	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	Obs. Avg.	Author
ILM	0.63	0.50	0.63	0.46	0.43	0.53 ± 0.10	0.45
ONFL	2.55	2.21	2.54	2.42	3.53	2.65 ± 0.51	2.41
IPL/INL	1.90	1.87	2.00	1.95	2.18	1.98 ± 0.12	1.95
OPL/ONL	0.63	0.53	0.54	0.66	0.55	0.58 ± 0.06	0.82
ELM/IP	0.21	0.20	0.18	0.19	0.19	0.19 ± 0.01	0.16
RPE	0.13	0.14	0.13	0.16	0.13	0.14 ± 0.01	0.13

Table 3.3: Mean absolute inter-observer differences (IOD) for the manually corrected layer boundaries among the dataset. The observer (Obs.) columns show the mean absolute difference of the layer positions in $[\mu m]$ of one observer compared to the others ($IOD_{o,l}$). The mean and standard deviation of the single observer differences is given in the observer average (Obs. Avg.) column (IOD_l). The column "Author" depicts the mean absolute differences of the author's manual correction to the observers' corrections ($IOD_{author,l}$).

practice, if they are carried out at all, are rarely done by the agreement of two or more operators. We assume that the evaluation results for the author's manual correction properties on the segmentation evaluation subset are transferable to the full classification database. The manual corrections performed by the author on the full classification dataset can thus be seen as representative for the corrections that an OCT operator might carry out, if perfect segmentation results are desired in the clinic.

The IOD_l can not only be computed over the whole database, but also on subsets of the database, i.e. scans of low or high quality and scans of normal (with diagnosis H or OHT) or glaucoma patients (with diagnosis PPG or PG). The results are given in Table 3.4. No connection between scans of low quality or scans of glaucomatous eyes can be directly observed. To further confirm these results we compute $IOD_{i,l}$, the mean absolute inter-observer difference for a specific layer l on a specific image i from the database, defined as:

$$IOD_{i,l} = \frac{1}{\#O} \sum_{o \in O} IOD_{o,l,i}; \tag{3.14}$$

The $IOD_{i,l}$ can be correlated with the presence of a low quality scan and the presence of glaucoma. No significant correlation with $P < 0.001$ can be found for any layer. But one must take care: Due to the construction of the evaluation we can not directly conclude that our observers are not influenced by the glaucoma disease or low quality. In the evaluation, only manual correction was performed. Therefore, an inter-observer difference can only appear in places that were corrected, i.e. the inter-observer difference and the automated segmentation error are most likely correlated. Only in the case that the automated segmentation error is not correlated to scans of low quality or the glaucoma disease, the conclusion that observers are not influenced by those criteria is valid. The correlation of the automated segmentation error with low quality or glaucoma will be looked at in the next Section 3.5.

Boundary	Glaucoma	Normal	Low Qality	High Qality
ILM	0.68	0.38	0.60	0.46
ONFL	2.42	2.88	2.61	2.68
IPL/INL	2.08	1.88	2.00	1.96
OPL/ONL	0.68	0.48	0.52	0.65
ELM/IP	0.33	0.06	0.06	0.33
RPE	0.19	0.09	0.04	0.24

Table 3.4: Mean absolute inter-observer differences for the manually corrected layer boundaries among different subject groups and B-Scan qualities. The columns show the mean absolute inter-observer difference of the layer positions in $[\mu m]$ among all observers. "Glaucoma" are the scans of preperimetric and perimetric glaucoma patients, "Normal" the scans of healthy and ocular hypertension subjects. "Low quality" depicts the half of the scans with low quality, "High quality" the half of the scans with high quality.

Despite the high probability of a correlation between the inter-observer difference and the automated segmentation error, we take a closer look at how the difference between the observers is distributed along the A-Scan positions by the standard deviation of the observer segmentations $STDO_l(r)$:

$$STDO_l(r) = \frac{1}{\#DB} \sum_{i \in DB} \sqrt{\frac{1}{\#O - 1} \sum_{o \in O} (L_{o,i,l}(r) - \bar{L}_{i,l}(r))^2}; \tag{3.15}$$

where $\bar{L}_{i,l}(r)$ is the observers' mean position of layer boundary l on image i at A-Scan position r. In Figure 3.8 the $STDO_l(r)$ is plotted for each segmented layer boundary. The plots are split into two graphs, one for inner layer boundaries in Figure 3.8 a) and one for the outer layer boundaries in Figure 3.8 b). Inner and outer does not resemble the separation of inner segment and outer segment from the algorithm description of Section 3.2, but is just a half-half partitioning of the layer boundaries for overview reasons. To those experienced in OCT layer segmentations, the plots of the observer standard deviations of the ONFL and, to a smaller extent, the IPL/INL boundary resemble a familiar shape, namely that of the retinal or RNFL thickness or the blood vessel density (BVD) among a circular scan dataset. Layer thicknesses are defined by:

$$LT(r) = L_{outer}(r) - L_{inner}(r) \tag{3.16}$$

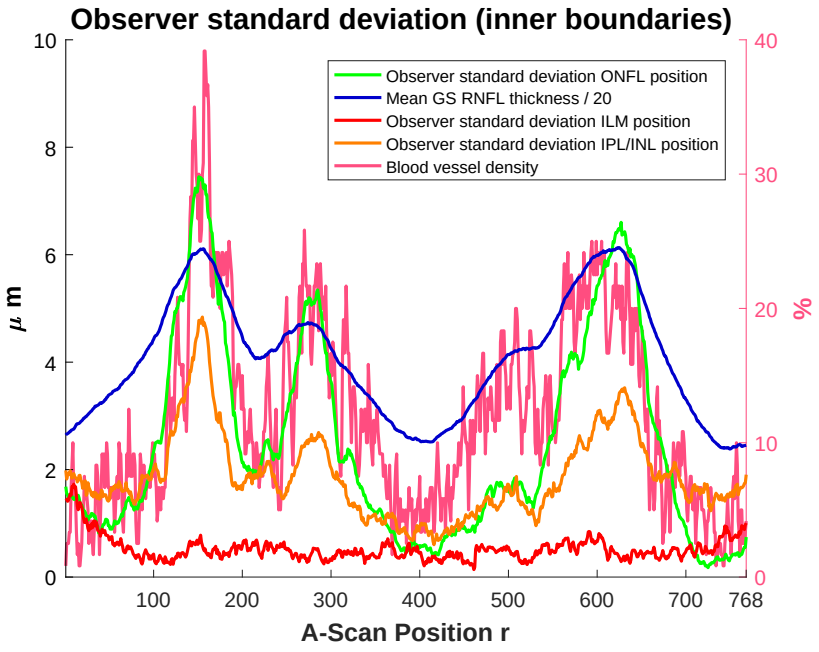
where $L_{outer}(r)$ is the position of the outer layer boundary and $L_{inner}(r)$ the position of the inner layer boundary of the respective layer at A-Scan position r. For the complete retina, the outer layer boundary is the RPE and the inner layer boundary the ILM. For the RNFL, the outer layer boundary is the ONFL and the inner layer boundary the ILM. The mean thickness among the segmentation evaluation dataset of the RNFL and retina are computed for the gold standard and included in Figure 3.8 a) and b) respectively. The BVD denotes the percentage of images among the segmentation evaluation dataset where a blood vessel is indicated on the author's correction of the automated blood vessel detection at a certain position. The BVD is drawn into

Boundary	BV Distrib.	Retina thickn.	RNFL thickn.
ILM	-0.18	0.05	-0.21
ONFL	0.84	0.80	0.93
IPL/INL	0.73	0.88	0.82
OPL/ONL	-0.12	0.46	0.02
ELM/IP	0.16	0.37	0.21
RPE	0.16	0.49	0.23

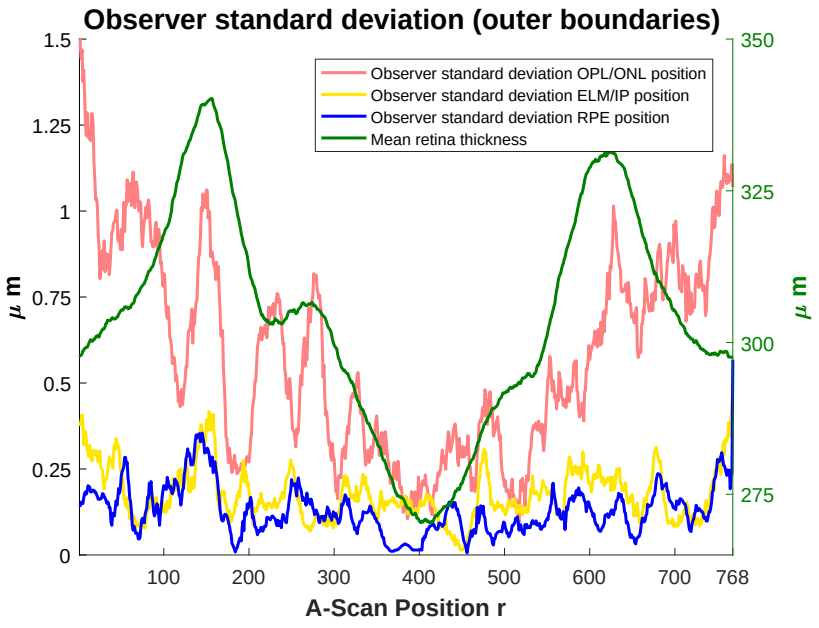
Table 3.5: Correlation of observer standard deviation along the A-Scan position r with the blood vessel distribution (BV Distrib.), retina thickness (thickn.) and retinal nerve fiber layer (RNFL) thickness. All correlations are significant with $P < 0.001$ except the correlation of observer standard deviation among the OPL/ONL boundary to the RNFL thickness and the correlation of observer standard deviation among the ILM boundary to the retina thickness. In both cases there is no correlation.

Figure 3.8 a). Indeed, the shapes of the observer standard deviations of the ONFL and the IPL/INL boundary are similar to the BVD, the retina or RNFL thickness, less so for the other layer boundaries. To quantify this similarity correlations can be computed. These correlations are shown in Table 3.5. Almost every observer standard deviation is correlated to the BVD, retina thickness and RNFL thickness with $P < 0.001$ with 2 exceptions only. The correlation is especially strong for the ONFL and IPL/INL boundary. As the three measures BVD, retina thickness and RNFL thickness themselve are correlated (correlation of BVD to RNFL thickness is 0.81, BVD to retina thickness 0.58, and retina to RNFL thickness 0.83, all with $P < 0.001$), one may conclude that the observers differ most in either regions of high BVD or large RNFL thickness. It is known from the literature that the blood vessel shadows are a severe challenge to automated segmentation algorithms [Chiu 10, Golz 11] and there is still no common rule to be found in the literature how to treat blood vessels even in manual segmentations. It is therefore most likely that the blood vessels are the reason for high inter-observer standard deviation, either directly due to a different opinion on the correction or indirectly due to more segmentation errors at this positions.

To conclude the evaluation on the observer corrections, their relation to the GS, which was constructed from them, is investigated. Figure 3.9 shows visual examples of the observers' corrections and the resulting GS. Obviously, a layer boundary with only few corrections leads to a high inter-observer agreement (see Figure 3.9 a)). Figure 3.9 b) is the scan with the second largest inter-observer difference for the ONFL boundary. It can be seen that the observers treated the BV areas very differently. The GS segmentation is a reasonable compromise. Figure 3.9 c) is a scan with very large inter-observer difference for the IPL/INL boundary. Some observer corrections are clearly mislead in BV regions, with the corrections been set to the ONFL or very near to the OPL/ONL boundary, despite the clear instruction for this layer: "*... follow it (the IPL/INL) on both sides until you are near the blood vessel shadow and then interpolate over the blood vessel region*". Nonetheless, the GS is reasonable.

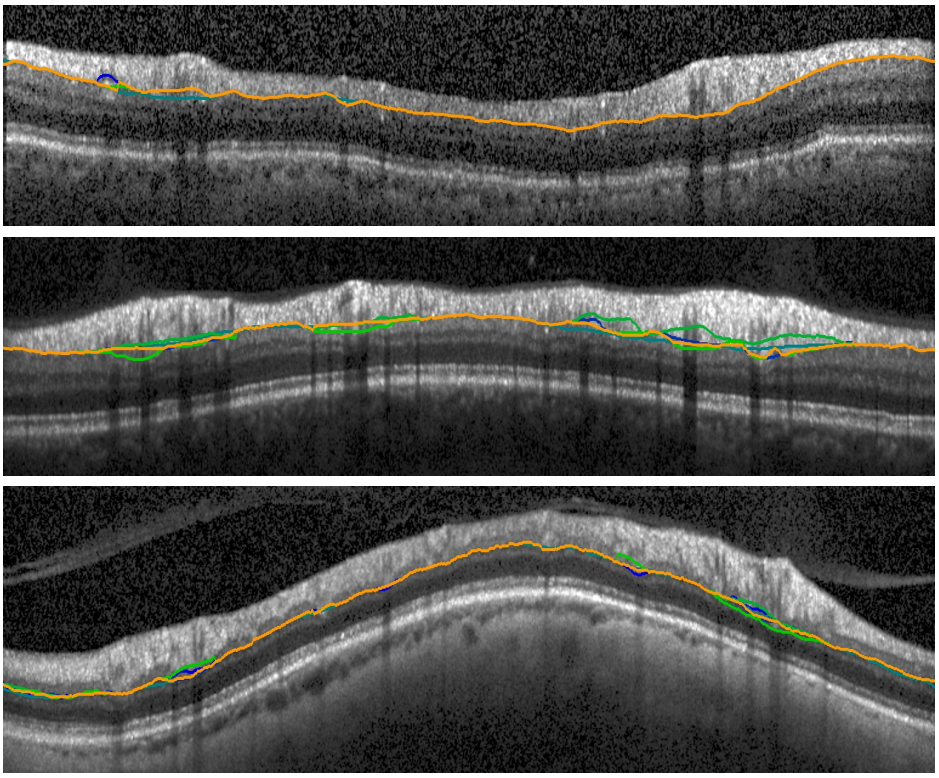


(a)



(b)

Figure 3.8: Observer standard deviation along the A-Scans. (a) The observer standard deviation for the inner layer boundaries plotted with the mean RNFL thickness and the blood vessel density. (b) The observer standard deviation for the outer boundaries plotted against the mean retina thickness. Note the different scales of the y-axis of the graphs a) and b).



(a)

(b)

(c)

Figure 3.9: Example images for inter-observer agreement. The layer boundaries are not shown in their default color. Only one layer boundary is drawn on each scan, with the 5 observer corrections in blue to green colors and the gold standard in orange. The images are cut in vertical direction to show only the retina. (a) High inter-observer agreement ($IOD_{Overall} = 0.56\mu$m), few corrections to the automated segmentation. PPG eye, quality 18.21dB. The ONFL segmentations are shown. (b) Low inter-observer agreement on the ONFL boundary ($IOD_{ONFL} = 10.23\mu$m). OHT eye, quality 29.00dB. (c) Low inter-observer agreement on the IPL/INL boundary ($IOD_{IPL/INL} = 4.19\mu$m). PPG eye, quality 21.47dB.

The visual examples can give only a sample expression on the validity of the GS. The absolute difference of an observer o to the GS for a layer l $DOG_{o,l}$ and the similarly computed signed difference $SDOG_{o,l}$ give a more global indication:

$$DOG_{o,l} = \frac{1}{\#R * \#DB} \sum_{i \in DB} \sum_{r \in R} |L_{o,i,l}(r) - L_{GS,i,l}(r)| \tag{3.17}$$

$$SDOG_{o,l} = \frac{1}{\#R * \#DB} \sum_{i \in DB} \sum_{r \in R} L_{o,i,l}(r) - L_{GS,i,l}(r); \tag{3.18}$$

The results of the $DOG_{o,l}$ computation are shown in Table 3.6 and the results of the $SDOG_{o,l}$ computation in Table 3.7. Both tables include the same measure computed for the author's correction, computed by replacing o by *author* in the equations above. The $DOG_{o,l}$ are similar between the observers. For most layers, the signs of the $SDOG_{o,l}$ values are distributed in a 2 to 3 fashion among the observers

Boundary	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	Author
ILM	0.43	0.30	0.36	0.35	0.31	0.35
ONFL	1.43	1.38	1.44	1.98	2.38	1.84
IPL/INL	1.30	1.06	1.15	1.30	1.59	1.71
OPL/ONL	0.44	0.25	0.25	0.46	0.40	0.71
ELM/IP	0.20	0.09	0.12	0.13	0.09	0.16
RPE	0.09	0.08	0.08	0.11	0.07	0.10

Table 3.6: Mean absolute difference of the manual correction by the observers (Obs.) and the author to the gold standard in [μm].

Boundary	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	Author
ILM	-0.25	0.00	-0.02	0.02	0.20	0.04
ONFL	-0.32	-0.80	-0.04	-1.40	1.64	-0.44
IPL/INL	0.03	-0.14	-0.03	-0.43	0.46	0.31
OPL/ONL	0.05	-0.02	-0.11	0.13	0.12	-0.12
ELM/IP	-0.15	0.04	-0.06	-0.05	0.05	-0.07
RPE	0.03	0.02	-0.01	-0.04	-0.01	-0.03

Table 3.7: Mean signed difference of the manual correction by the observers (Obs.) and the author to the gold standard (GS) in [μm]. A negative sign indicates that the mean position of the layer is more to the inner side of the scan compared to the GS.

except for the ONFL layer boundary. Observer 4 placed the ONFL boundary too much to the inner side of the scan and Observer 5 to the outer side compared to the GS. The mean signed differences for the other 3 observers are small for the ONFL layer boundary compared to the 2 extremes. This confirms the special opinion of these two observers on the segmentation of ONFL layer boundary but also shows that the GS is not influenced by the outliers and sticks to the majority of the observers, as does the similarity of the $DOG_{o,l}$ and the distribution of signs of the $SDOG_{o,l}$ of the other layers confirm the validity of the GS construction rule. The author's DOG and $SDOG$ values only add to the conclusion drawn above. They are, except the OPL/ONL boundary, never an extremum compared to the observers.

The summarized conclusions drawn from this section are the following: The observer differences are highest in areas of high blood vessel density. The gold standard construction rule formulated in Section 3.4 seems valid and reasonable. The author's corrections can be seen as representative for the corrections that a single OCT operator might carry out.

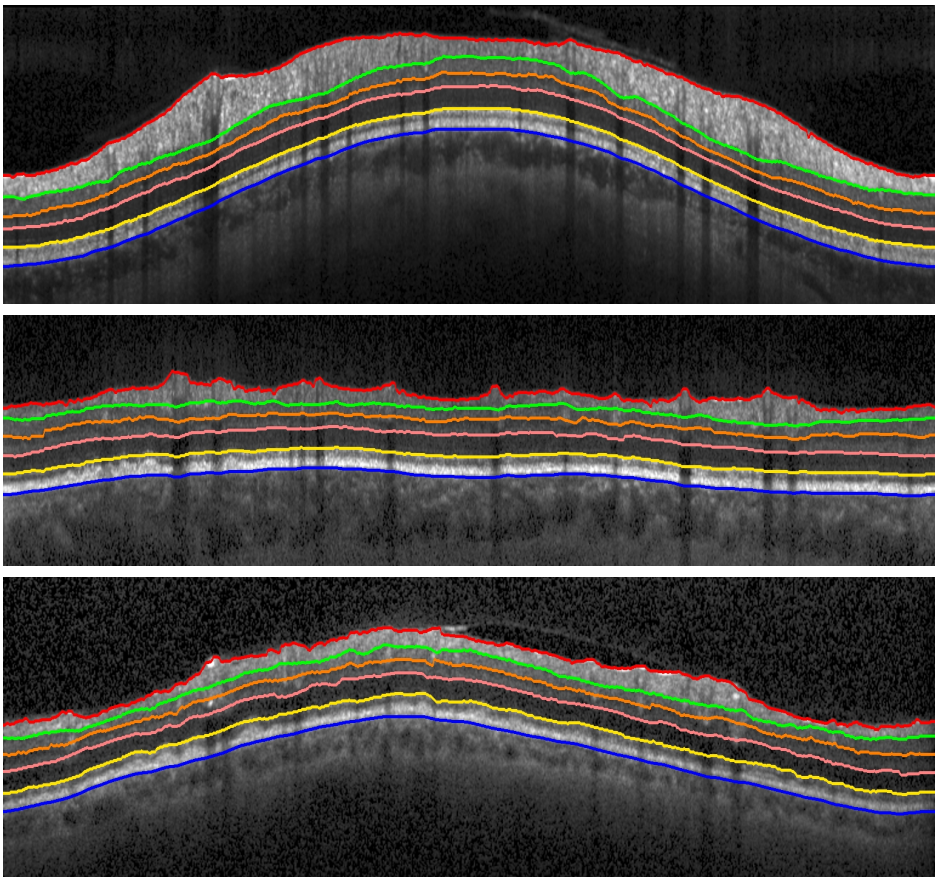
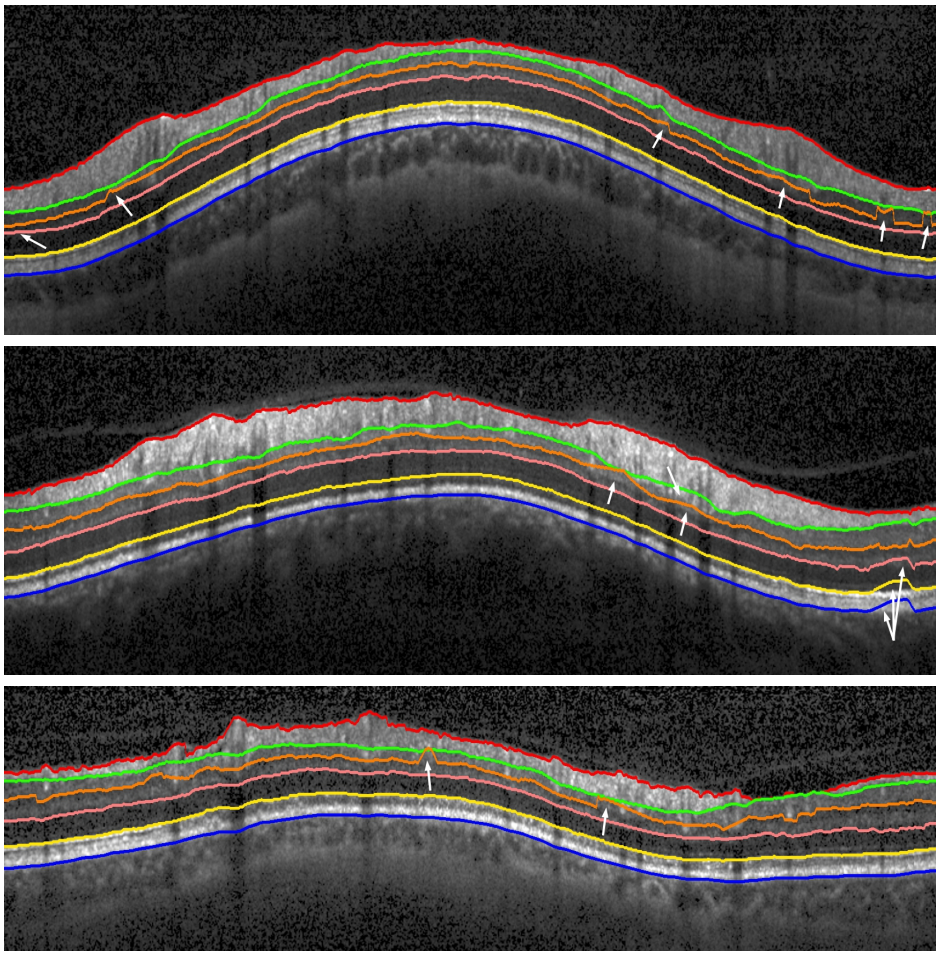


Figure 3.10: Examples of *successful* automated segmentations. The images are cut in vertical direction to show only the retina. (a) Scan of a healthy eye. Segmentation error $SE_i = 0\mu$m. Very high quality 40.31dB. (b) Scan of a preperimetric glaucoma eye. Segmentation error $SE_i = 0.13\mu$m. Very low quality 18.82dB. (c) Scan of perimetric glaucoma eye. Segmentation error $SE_i = 0.03\mu$m. High quality 24.19dB.

3.5 Automated segmentation evaluation and discussion

The evaluation of the automated segmentation results includes visual examples of the segmentation results, a brief look at the blood vessel segmentation, followed by a comparison of the automated layer segmentation results to the gold standard. This comparison is performed on the complete dataset, on scan groups and along the A-Scan positions.

At first, a qualitative impression on the segmentation results is given by example images in Figures 3.10, 3.11 and 3.12. Figure 3.10 shows successfull automated segmentations from a) a healthy eye, b) a PPG eye, and c) a PG eye. The measured scan quality is among the best in the dataset for the example 3.10 a) and among the worst for 3.10 b). The segmented layer boundaries follow the contrast changes inside the scan closely and the treatment of blood vessel regions is reasonable - a line connecting the segmentation on the left and right side of the blood vessel shadow.



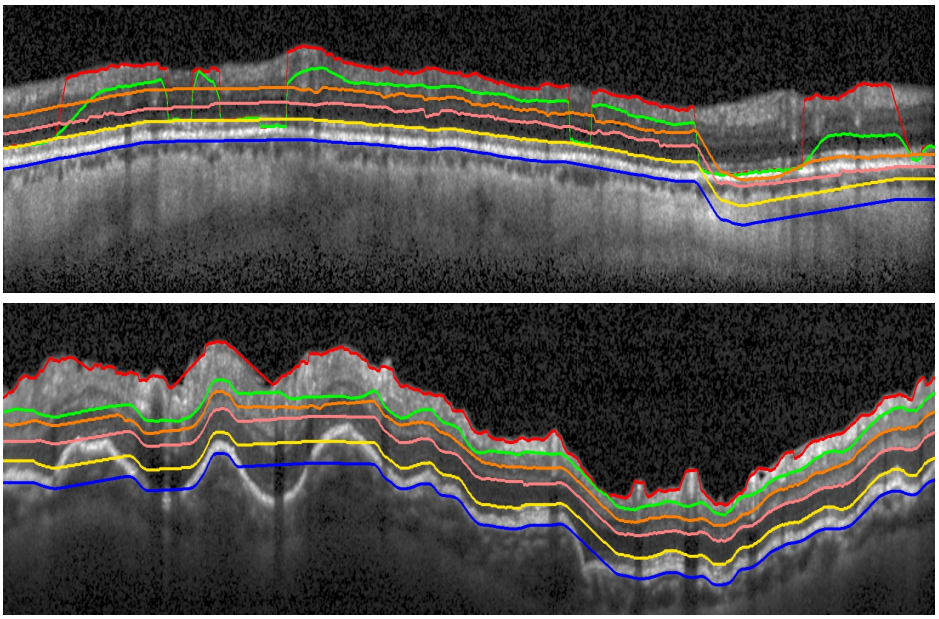
(a)

(b)

(c)

Figure 3.11: Examples of automated segmentations *with errors*. The images are cut in the vertical direction to show only the retina. Segmentation errors are marked with white arrows. (a) Scan of a healthy eye. Segmentation error $SE_i = 1.7\mu$m with segmentation failures on the IPL/INL boundary. Low quality 20.47dB. (b) Scan of an ocular hypertension eye. Segmentation error $SE_i = 1.77\mu$m with segmentation failures mainly on the ONFL and RPE boundaries. The failures on the RPE propagate to the inner layers. Low quality 21.24dB. (c) Scan of a perimetric glaucoma eye. Segmentation error $SE_i = 0.91\mu$m with segmentation failures on the IPL/INL boundary. Very low quality 17.64dB.

Moderate segmentation errors are visible in the examples in Figure 3.11 and marked with a white arrow. Most errors appear on the ONFL and IPL/INL boundary. The IPL/INL boundary has clear jumps in areas where it is barely visible to the human eye as it can be seen in the examples a) and c). An energy minimization approach similar to the ONFL segmentation could prevent these jumps and may be investigated in future work. The error on the RNFL in the example b) is located in the inferior quadrant of the scan with high blood vessel density and expected high RNFL thickness. The segmentation was mislead by the preceding wrong positioning of the IPL/INL and most likely by remaining speckle in the RNFL. A propagation



(a)

(b)

Figure 3.12: *Extreme failures* of the automated segmentation. The images are cut in the vertical direction to show only the retina. (a) Scan of a perimetric glaucoma eye. Segmentation error $SE_i = 40.49\mu$m. High quality 22.85dB. The intensities of the layers and sclera are very different from other scans due to the positioning of the retina at the outermost scan area (not visible in the cut of the image shown). (b) Scan of a perimetric glaucoma eye. Segmentation error $SE_i = 17.18\mu$m. High quality 26.70. The wavy retinal structure caused the algorithm failure.

of a segmentation error of the RPE to the inner layers is observed on the right side of the scan in the temporal quadrant. The IP layer is highly reflective at this position and this leads to the error. The assumption on the parallelism of the RPE and the inner layers are the reason for the error propagation. Segmentation errors on the RPE are, however, very uncommon, as we will see later in this section. A final remark has to be made on Figure 3.11 c), a scan of a PG eye with complete loss of the RNFL between the inferior and temporal quadrants. The algorithm detected the loss and remaining thin RNFL perfectly. The errors on the IPL/INL boundary did not influence the ONFL segmentation.

There were 2 extreme failures of the automated segmentation in the segmentation evaluation dataset, with one being a complete failure. This complete failue is shown in Figure 3.12 a). The failure can be reasoned: The retina was positioned at the very outer end of the scan area, contrary to the operating guidelines for the HE Spectralis. One can only guess that this was the only possibility to get a scan from the glaucomatous eye. This positioning of the retina lead to an intensity distribution along the A-Scans very different from the other scans. The outer layers and the sclera are much brighter than usual which caused the segmentation error, as even the IS/OS separation failed in parts. The error in Figure 3.12 b) is due to a violation of one of the algorithm's basic assumptions: The wavy structure of the retina caused the RPE detection to fail. As no further patient information beside the glaucoma diagnosis is

BV correction	**0-2**	**2-4**	**4-6**	**6-8**	**8-10**	**>10**
#Img	79	31	7	1	0	2

Table 3.8: Number of B-Scans from the segmentation evaluation dataset with the respective fraction of A-Scan blood vessel labels corrected. The values in the header are in [%], with the lower boundary included and the higher excluded.

available for this work, one might only guess the presence of a disease in addition to glaucoma. Nevertheless, we accepted scans like Figure 3.12 a) and b) on full purpose for this work and also for the segmentation evaluation. They represent scans from daily clinical practice that do not always favor perfect results in the evaluation of a segmentation algorithm. However, we will comment on how these two single scans influence the evaluation and where results have therefore to be judged carefully. Both extreme failure scans are from the same patient group (PG) and have the same scan quality (high), but the segmentation errors can not be reasoned with either of these two classifications.

While the correction of the automated blood vessel segmentation results was only performed by the author, some insight can be gained in how useful the results are for the following segmentation steps, as the blood vessel positions are used in each of them. The blood vessel segmentation only assigns a label to an A-Scan, if the A-Scan is at a blood vessel position or not. During the manual correction, 2.13% of these labels had to be corrected on the segmentation evaluation dataset. Beside this global number, the distribution of this error among the dataset is a more meaningful measure. It is shown in Table 3.8. The majority of scans did not need major corrections to the blood vessel positions with few exceptions. On two scans, more than 10% of the labels were wrong. One was the extreme failure shown in 3.12 b) with the wavy retina. On the other scan, severe speckle noise lead the thresholding to assign blood vessel labels to almost half of the A-Scans. This behavior could be prevented by a more speckle-sensitive algorithm without a fixed threshold parameter. However, the segmentation results were not that much affected, with segmentation errors being only minor and not due to the automated blood vessel labeling. Overall, by visual inspection of the results, it turned out that mostly large blood vessels, as stated before, are not found.

The layer segmentation error of an image i and layer boundary l can be quantified by the mean absolute distance of the automatically segmented layer boundary to the gold standard:

$$SE_{l,i} = \frac{1}{\#R} \sum_{r \in R} |L_{autom,i,l}(r) - L_{GS,i,l}(r)|; \tag{3.19}$$

The segmentation error SE_l for a specific layer l can be computed as the mean of the $SE_{l,i}$ among the dataset or a subgroup of the dataset:

$$SE_l = \frac{1}{\#DB} \sum_{i \in DB} SE_{l,i}; \tag{3.20}$$

In almost the same manner, an overall measure SE_i for a specific image i is created by the average of $SE_{l,i}$ among all layer boundaries, and an overall measure SE_{Group}

Group	ILM	ONFL	IPL/INL	O/O	ELM/IP	RPE	All
Overall	1.07	2.84	2.56	0.61	0.58	0.51	1.36
HQ	1.65	3.40	2.99	0.97	1.10	0.95	0.95
LQ	0.49	2.28	2.14	0.25	0.07	0.07	0.07
Normal	0.33	2.75	2.58	0.25	0.13	0.15	0.15
Glaucoma	1.81	2.93	2.55	0.97	1.04	0.87	0.87
Healthy	0.32	2.43	2.44	0.23	0.17	0.21	0.21
OHT	0.33	3.07	2.73	0.27	0.08	0.08	0.08
PPG	0.83	1.28	1.59	0.24	0.07	0.07	0.07
PG	2.80	4.58	3.50	1.70	2.01	1.67	1.67
N & HQ	0.08	2.56	2.72	0.26	0.18	0.22	0.22
N & LQ	0.57	2.94	2.44	0.24	0.07	0.07	0.07
G & HQ	3.21	4.24	3.25	1.68	2.01	1.67	1.67
G & LQ	0.42	1.62	1.85	0.26	0.07	0.06	0.06

Table 3.9: Mean absolute difference, i.e. the segmentation error SE of the automatically segmented layer boundary positions to the gold standard for all segmentation evaluation B-Scans and subgroups of the dataset. The abbreviations are: OPL/ONL boundary (O/O), High quality scans (HQ), low quality scans (LQ), healthy and ocular hypertension (OHT) subjects (Normal, N), preperimetric (PPG) and perimetric glaucoma (PG) patients (Glaucoma, G). The possible combinations of the groups N/G and LQ/HQ contain 30 subjects each. "All" is the mean over all layer boundaries. Values are given in [μm].

for the complete dataset or subgroup of the dataset by the average of $SE_{Group,l}$ among all layer boundaries, were *Group* is a specific subgroup of images of the segmentation evaluation dataset with a certain property. Naturally, the segmentation error measurements can be made for combinations of subgroups and layers. An example equation is the segmentation error of the ONFL on normal (N) eyes $SE_{ONFL,N}$:

$$SE_{ONFL,N} = \frac{1}{\#N} \sum_{i \in N} SE_{ONFL,i} \tag{3.21}$$

where N is the subgroup of images in the segmentation evaluation dataset with diagnosis N (healthy H and ocular hypertension OHT combined), $\#N$ is the number of these images and $SE_{ONFL,i}$ is the segmentation error of the ONFL for the image i. Table 3.9 summarizes the segmentation error SE for the complete segmentation evaluation dataset and the different layers and subgroups of the dataset. The subindex to SE can be added from the context. The ONFL boundary segmentation resulted in the largest segmentation error of $SE_{ONFL} = 2.84\mu$m, closely followed by the IPL/INL boundary $SE_{IPL/INL} = 2.56\mu$m. This is of no surprise, as these two layer boundaries are the greatest challenge in retinal layer segmentation. All other layer segmentations performed better, with the RPE being the layer with the smallest error $SE_{RPE} = 0.51\mu$m.

Before discussing the subgroups of the segmentation evaluation dataset, we will look at the distribution of the segmentation error $SE_{l,i}$ among the whole dataset

Boundary	0-2	2-4	4-6	6-8	8-10	>10
ILM	113	2	2	1	0	2
ONFL	70	31	6	7	2	4
IPL/INL	69	35	7	4	0	5
OPL/ONL	117	1	0	0	0	2
ELM/IP	117	1	0	0	0	2
RPE	117	0	1	0	0	2
Overall	110	6	2	0	0	2

Table 3.10: Number of B-Scans from the evaluation dataset with the segmentation error $SE_{i,l}$ within a certain range. The values in the header are in $[\mu m]$ with the lower boundary included and the higher excluded.

in Table 3.10. It can be seen that the segmentation error is not equally distributed among the scans. A large fraction, more than 100 of the 120 scan in the segmentation evaluation database, has a segmentation error below 4μm for any layer. The overall error is smaller than 2μm for 110 of the images. Errors on the ILM, OPL/ONL, ELM/IP and RPE boundary are very seldom and errors on the ONFL and IPL/INL boundary are more common. Each layer boundary has at least 2 images with a segmentation error above 10μm. With one exception, the ILM, these always include the examples with extreme errors as shown in Figure 3.12. In fact, these two scans contribute, for example, 35.33% to the overall, 25.42% to the ONFL and 81.56% to the RPE error. As mentioned above, the two failure scans belong to the same scan subgroups, high quality and PG scans. With the knowledge of this segmentation error distribution and the two extreme failures, we will look at the results for the subgroups also shown in Table 3.9. It is no surprise that the averaged numbers show a higher segmentation error for high quality and PG eyes. The segmentation errors for the OPL/ONL, ILM/EP and RPE are near to non-existent for the other subgroups. To find out whether there truly is a connection between quality or glaucoma diagnosis and the segmentation error, the correlation of the membership of an image in these groups with its segmentation errors was computed. No correlation yielded any significance ($P < 0.01$). The conclusion is that segmentation error differences in Table 3.9 are due to random effects in the segmentation evaluation database, but not due to algorithm design. The algorithm provides similar segmentation error results on scans of high or low quality, and on scans of normal and glaucomatous eyes.

While the segmentation error SE defined as the mean absolute difference of the automated segmentation to the gold standard gives insight on how much error the automated segmentation makes on the evaluation database, the signed segmentation error SSE defined as the mean signed difference of the automated segmentation to the gold standard shows in which direction the layer boundaries are set wrongly to a larger extent:

$$SSE_{l,i} = \frac{1}{\#R} \sum_{r \in R} L_{autom,i,l}(r) - L_{GS,i,l}(r); \tag{3.22}$$

The results are summed up similarly to the SE results of Table 3.9 in Table 3.11. On the whole dataset, the ONFL and IPL/INL boundaries are more often positioned

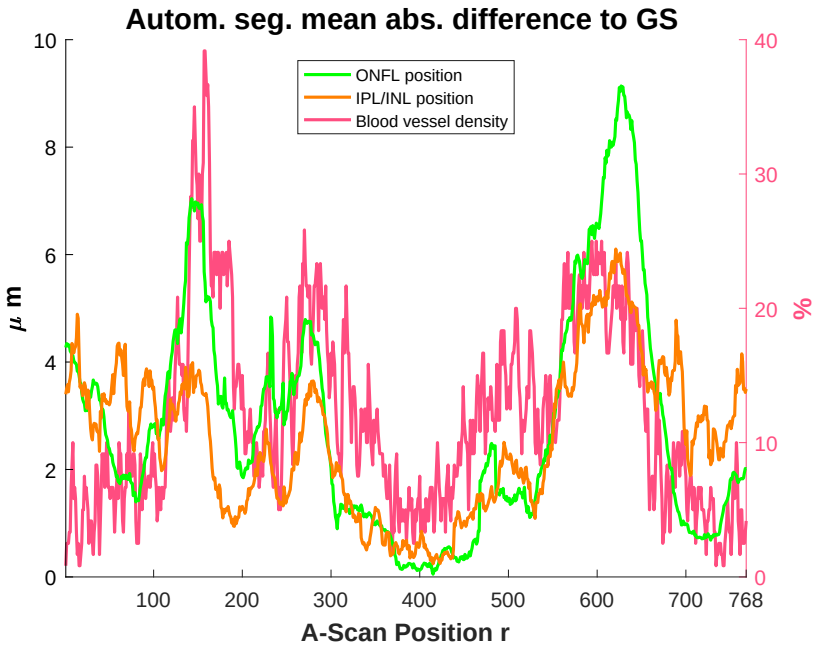
Group	ILM	ONFL	IPL/INL	O/O	ELM/IP	RPE	All
Overall	0.69	-1.23	-1.46	0.30	0.12	0.22	-0.23
HQ	1.12	-0.75	-1.33	0.56	0.23	0.43	0.43
LQ	0.25	-1.71	-1.59	0.04	0.01	0.01	0.01
Normal	0.12	-2.15	-2.15	0.01	0.05	0.04	0.04
Glaucoma	1.25	-0.30	-0.77	0.59	0.19	0.40	0.40
Healthy	0.28	-1.81	-1.92	0.05	0.10	0.10	0.10
OHT	-0.04	-2.50	-2.37	-0.04	-0.01	-0.02	-0.02
PPG	-0.15	-1.03	-1.03	0.06	0.05	0.05	0.05
PG	2.65	0.42	-0.51	1.12	0.33	0.74	0.74
N & HQ	0.06	-1.97	-2.25	0.02	0.12	0.11	0.11
N & LQ	0.18	-2.33	-2.04	-0.00	-0.03	-0.03	-0.03
G & HQ	2.19	0.47	-0.42	1.10	0.34	0.75	0.75
G & LQ	0.31	-1.08	-1.13	0.08	0.04	0.05	0.05

Table 3.11: Mean signed difference (SSE) of the automatically segmented layer boundary positions to the gold standard for all segmentation evaluation B-Scans and subgroups of the dataset. A positive difference indicates that the automatically segemented boundary positions is in average set wrongly to the outer Z-direction (i.e. too low in the B-Scan image), a negative differences indicates the average position wrongly set to the inner Z-direction. Abbreviations and further information see Table 3.9.

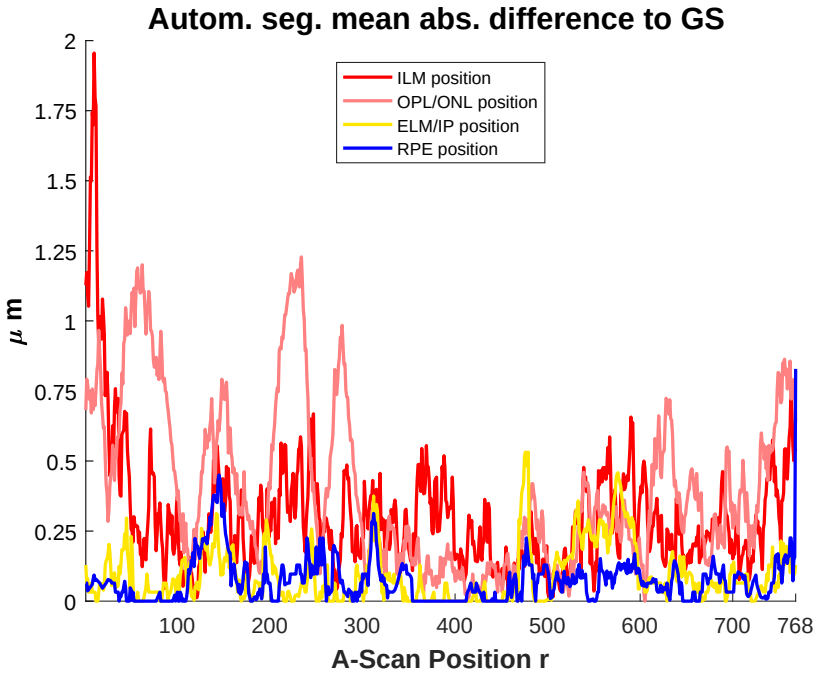
wrongly to the inner side of the scan. The separated evaluation on subgroups gives further insight into this behavior. The wrong positioning of the ONFL and IPL/INL boundaries to the inner side seems to be more pronounced on normals compared to glaucoma patients, i.e. subjects with a larger RNFL. Indeed, the computed correlations of the $SSE_{l,i}$ with glaucoma (PPG and PG) or the single PG diagnosis for the ONFL and IPL/INFL are higher for the SSE (in the range of 0.15 to 0.2) than for the SE (near to 0). However, the correlation again did not yield any real significance ($P < 0.01$) for any layer, with the P value of the SSE to glaucoma correlation for the ONFL and IPL/INFL boundary being just below 0.1. All other numbers of the Table 3.11 are also not considered to add more information. Of course, there is the possibility that the significance of the correlations would increase, i.e. P would decrease, with a larger evaluation database. It has to be pointed out that the number of scans included in our evaluation is in the same magnitude and comparable to almost all other works published about retinal layer segmentation (see literature overview Table B.1 and following). This shows that conclusions drawn out of pure overall segmentation error results on scan groups have to be taken with a grain of salt when no correlation and corresponding significance is given and the database is only in the size of about a hundred, e.g. in [Rath 14].

The segmentation error along the A-Scans gives insight how the errors are distributed locally:

$$SE_l(r) = \frac{1}{\#DB} \sum_{i \in DB} |L_{autom,i,l}(r) - L_{GS,i,l}(r)|; \tag{3.23}$$



(a)



(b)

Figure 3.13: Segmentation error along the A-Scans $SE_l(r)$. (a) The segmentation error along the A-Scans for the ONFL and IPL/INL boundaries plotted with the blood vessel density. (b) The segmentation error along the A-Scans for the other segmented boundaries. Note the different scales of the y-axis of the graphs a) and b).

Boundary	BV Distrib.	Retina thickn.	RNFL thickn.	Obs. diff.
ILM	0.07	0.25	0.12	0.61
ONFL	0.64	0.82	0.83	0.74
IPL/INL	0.31	0.73	0.54	0.42
OPL/ONL	-0.20	0.31	-0.03	0.49
ELM/IP	-0.02	0.40	0.20	-0.06
RPE	0.02	0.41	0.22	0.08

Table 3.12: Correlation of mean absolute position difference to the gold standard of the automatically segmented layer positions $SE_l(r)$ along the A-Scan position r with the blood vessel (BV) distribution, retina thickness (thickn.), retinal nerve fiber layer (RNFL) thickness and the mean absolute observer difference (Obs. diff) to the gold standard ($DOG_l(r)$). All correlations with an absolute value above 0.1 are significant with $P < 0.001$.

The $SE_l(r)$ is plotted for the ONFL and IPL/INL boundary in Figure 3.13 a). In addition, the blood vessel distribution as defined in Section 3.4 is shown as a representative for the correlated measures blood vessel distribution, RNFL and retina thickness. The $SE_l(r)$ for the other layers is plotted in Figure 3.13 b). The correlations between the blood vessel distribution, the RNFL and retina thickness and the segmentation error along the A-Scans are given in Table 3.12. The segmentation errors of the ONFL and IPL/INL are significantly ($P < 0.001$) correlated to all 3 measures along the A-Scan, especially to the retina and RNFL thickness. From that we can conclude that either high RNFL thickness or the blood vessels are the main challenges to the automated segmentation of the ONFL and IPL/INL layers, as expected. The segmentation error of the other layers seems to be more evenly distributed in Figure 3.13 b), but there are significant correlations to the thickness of the complete retina. A high retina thickness seems to impair the segmentation results. One possible explanation is an increased chance that a larger quantitative amount of remaining speckle after denoising inside the retina leads to more false layer boundary detections.

A last point of interest is how the segmentation error of the automated segmentation relates to the observers. We therefore compute the mean absolute observer difference to the gold standard along the A-Scans:

$$DOG_l(r) = \frac{1}{\#O * \#DB} \sum_{o \in O} \sum_{i \in DB} |L_{o,i,l}(r) - L_{GS,i,l}(r)|; \tag{3.24}$$

This measure is similarly constructed as the segmentation error $SE_l(r)$ and tells how far the average observer layer boundary position is from the gold standard. The $SE_l(r)$ and $DOG_l(r)$ values for the ONFL are plotted in Figure 3.14 a) and for the IPL/INL in Figure 3.14 b). These two layers are chosen as examples as they exhibit the highest segmentation error. By no surprise, the measures are correlated, as shown in Table 3.12 in the last column, as observer differences to the gold standard may appear only at positions were at least two observers corrected the automated segmentation. It is a surprise, however, that the average observer difference to the

gold standard is roughly in the same order of magnitude as the segmentation error, as Figure 3.14 shows. For the ONFL, the disagreement of the observers with the gold standard is in some positions even higher than the segmentation error, and for the IPL/INL the segmentation error is in most positions not higher than twice the observer difference to the gold standard. The automated segmentation therefore is not much worse in its results than the average observer, even if the observer should only correct the automated segmentation results.

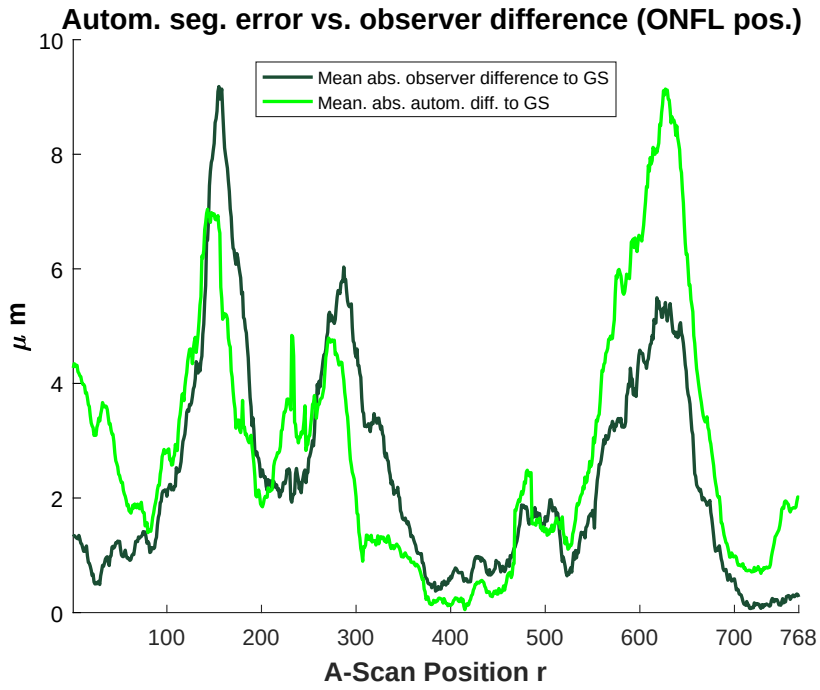
To conclude the evaluation of the automated segmentation, the findings can be summarized as follows: The ONFL and IPL/INL yield the highest segmentation errors. The other layers have errors almost exclusively on 2 scans from the dataset, that let the segmentation fail due unexpected properties of the image content. There is no significant correlation between the segmentation error and either scans of bad quality or glaucomatous eyes. The ONFL and IPL/INL are more wrongly positioned to the inner side of the scan. But even the automated segmentation results of these two layers with the highest segmentation error are evaluated not much worse than the average observer compared to the gold standard.

3.6 Outlook

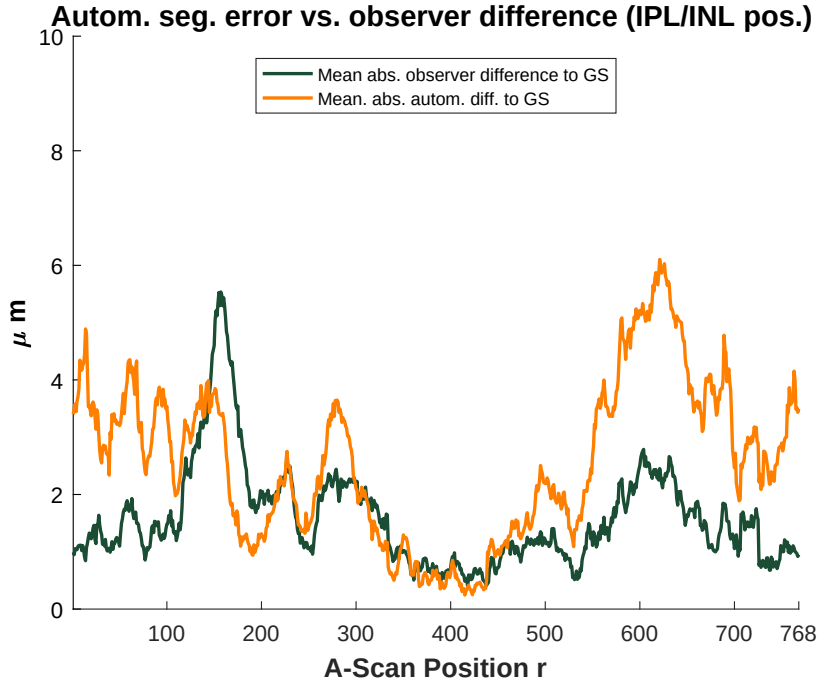
The presented automated retinal layer segmentation algorithm on circular OCT B-Scans has been extended to volume scans, and preliminary results were published at a conference [Maye 11]. The volume segmentation algorithm is intended for the use on volumes around the ONH.

The averaging and some layer smoothing steps were performed in 3D, while all other steps were performed on individual B-Scans. Some of the general algorithm assumptions do not hold in the case of a linear B-Scan out of a volume, i.e. while we still assume the RPE boundary is not disrupted, has a simple shape and this shape can be modeled by a polynomial with a rather low degree, i.e. below or equal to 5, a simple L2 norm polynomial fit will not provide adequate results for outlier detection as the ONH leads to massive outliers that prevent a good L2 norm polynomial fit. The L2 polynomial fit was therefore replaced by a RANSAC fit of a polynomial with an L0 norm as an error measure. The L0 norm is 1 if the distance of a boundary position is within a certain threshold to the polynomial, and 0 otherwise. A similar RANSAC fit was used for outlier detection on the inner layers, as the assumption that they are roughly parallel to the RPE is also not true for volumes. A step to find the ONH position was introduced after the RPE detection. This ONH detection works on the en-face view of a slice of pixels above the RPE and finds the ONH by pre-processing the en-face view with thresholding, followed by morphological operators and computing the center of gravity afterwards. The ONH positions are invalidated for all layer boundaries after the ONH detection. Most of the parameters from the circular B-Scan method had to be adapted for volume scans.

The data available are volume scans from one eye of 3 healthy subjects and 7 glaucoma patients, with varying numbers of B-Scans and A-Scans per B-Scans. These are too few for a resilient evaluation. Preliminary results (taken from [Maye 11]) were that the percentage of the RNFL thickness with no more than a 10μm thickness derivation from a manually corrected result was 11% on the glaucoma volumes and



(a)



(b)

Figure 3.14: Segmentation error $SE_l(r)$ and mean absolute observer difference to gold standard $DOG_l(r)$ along A-Scan positions r. (a) ONFL boundary position. (b) IPL/INL boundary position.

4.4% on volumes of normal subjects. Two example segmentation results, one from a normal and one from a glaucomatous eye are shown with the corresponding RNFL thickness maps in Figure 3.15.

While results of the volume segmentation extension of the proposed algorithm were promising, the computation time, at least of the Matlab implementation, is far too long for a use in daily clinical practice. In addition, graph-based approaches have shown their validity and good applicability on OCT data, be it for 2D or 3D segmentation. In the author's opinion, the graph-cut-based methods proposed in [Anto 13, Dufo 13, Lang 13, Cara 14] provide the best base for future research, as they are algorithmically compact, most easily applicable to volume data and allow for fast computation times. Dynamic programming as used by Chiu et al. [Chiu 15] is as much promising, but lacks the ability for an easy 3D extension.

The use of model-based approaches is suboptimal in the case of segmenting scans of glaucoma patients. Models are, up to now, always built upon data of healthy subjects. This may lead to more segmentation errors on scans of glaucoma patients, as shown by the only publication known to the author that was based on a model but used glaucoma data in the evaluation [Rath 14]. Models could include glaucomatous eyes, but the forms of glaucoma defects in the retina and their progression are manifold [Leun 12] and models would need a huge database to built on to capture the variety of the layer structure change of glaucomatous eyes. However, the possibility to directly draw a glaucoma score or quality measure of the segmentation from the derivation of the segmentation from the model as proposed by Rathke et al. [Rath 14] is exceptionally well thought and this idea deserves future attention.

One of the most interesting ideas for retinal layer segmentation was published in [Chiu 15] and [Srin 14a]. Before the actual segmentation, the content of the scan is preclassified, e.g. in [Chiu 15] regions of fluid-filled edema of DME patients are detected before the actual segmentation and in [Srin 14a] the number of layers to segment on scans of mice eyes is identified. These approaches could be thought further for daily clinical use: A classification step based purely on image content or a simple and robust retina segmentation determines the presence of diseases. A suitable segmentation algorithm is chosen based on the classification result, e.g. a model can be utilized in the case of a normal subject, complete RNFL loss has to be taken into account for glaucoma patients, and fluid filled regions or other structures deviating from normal shape are identified for AMD and DME patients before the actual layer segmentation.

In this chapter, we presented a retinal layer segmentation algorithm to segment 5 layers or layer groups on circular B-Scans around the ONH of normal subjects and glaucoma patients. Its validity was proven in an extensive evaluation. Retinal layer segmentation is only the first step in building an automated glaucoma classification system or glaucoma score. In the next Chapter 4, we make use of the layer segmentation by computing features from it that may discriminate between normal subjects and glaucoma with the help of a classifier.

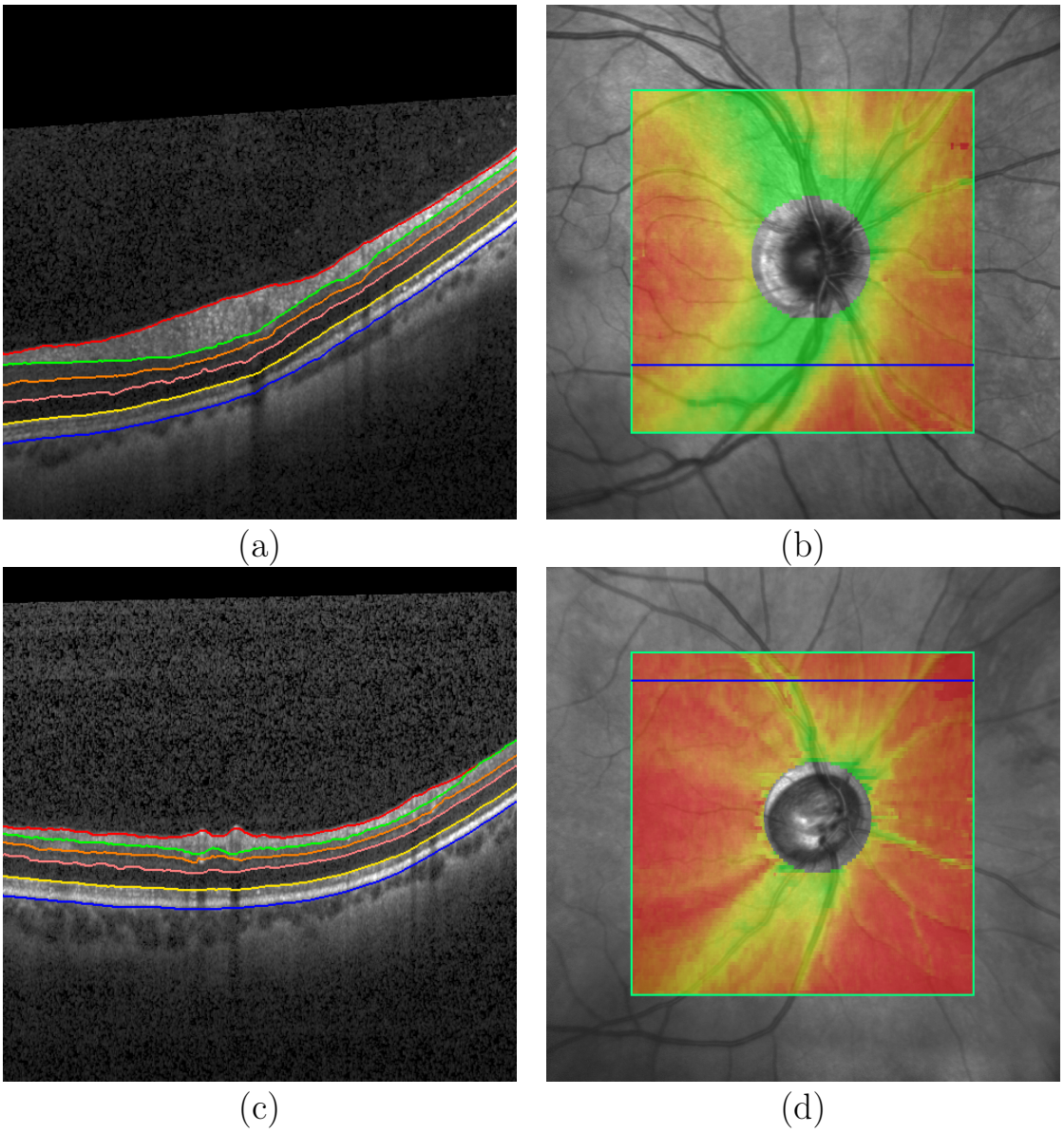


Figure 3.15: Example volume segmentation results: (a) B-Scan from a volume scan of a normal subject. The automated layer segmentations are drawn. (b) Corresponding RNFL thickness map of the normal subject. The position of the B-Scan in a) is marked. (c) B-Scan out of a volume scan of a glaucoma patient. (d) Corresponding thickness map to (c). Color code of the thickness map: Pure green (thick RNFL > 150 μm), pure yellow (RNFL = 75 μm), red (thin RNFL).

Chapter 4

Glaucoma classification

Recalling the standard pattern recognition pipeline presented in the introduction Section 1.4 and in Figure 1.6 we have just taken the first, but most crucial step, in pre-processing of the data. Circular OCT scans are transformed to thickness measurements of retinal layers. What follows in this chapter are all further processing steps required for a complete pattern recognition pipeline that yields a classification result. First, the thickness measurements are further processed by a focus and age normalization. Details are given in Section 4.2. Out of the thickness profiles of the 5 segmented layers and the complete retina, as well as the blood vessel positions, various features are computed. The features are not limited to common mean, minimum and maximum values, but include ratios and principal component analysis (PCA) features. We have a closer look, especially at the PCA features, in Section 4.3. The remainder of the pattern recognition pipeline is built up on basic methods for feature selection and classification. "Forward selection and backward elimination" is the feature selection method, and three widely used classifiers are compared. They are briefly introduced in Section 4.4. This work does not focus on getting the best classification result at any cost, but rather evaluates how the choices made in building up a classification system influence the result, i.e. not only the classification accuracy, but also, for example, the feature selection. The decisions that have to be made for a final classification system are: The dataset is labeled with 4 diagnoses. What is the most interesting classification task that can be constructed out of these diagnoses? Do we utilize thickness normalization? What classifier is chosen? Does a manual correction of retinal layer segmentation results from an observer improve the classification, or can we rely on the automated result? Combining every possible answer to these questions builds up a huge evaluation parameter space. Therefore, in Section 4.5 on the evaluation, first the evaluation parameter space is structured in a meaningful way (Subsection 4.5.1) and the results and necessary discussions are presented step by step in Subsections 4.5.2 to 4.5.5. For all the classification experiments in the evaluation section, a 10-fold cross-validation on the data is used to split it into disjoint training and test samples. The cross-validation rule that training and test data do not influence each other is violated three times during the computation and normalization of the features, i.e. by performing calculations over the complete dataset. These violations of the cross-validation rule are explicitly mentioned and reasoning why classification results are not affected are given in the respective subsections. Finally,

a glaucoma score constructed from the classification results is proposed in Section 4.6. An outlook how a glaucoma classification system can be further improved is given in Section 4.7. But, first of all, before the classification process, its properties, and results are described, the novelty of the approach is shown by an overview of the published work in related research fields in the following Section 4.1.

4.1 State of the art

The diagnostic power of OCT for detecting glaucoma and its progression was proven early with prototype and the first available commercial TD-OCT systems [Bowd 01, Aydi 03, Gued 03]. With the increasing placement of OCT systems in clinics, a great number of publications have investigated the reliability, reproducibility and possible operator errors of measurements, comparisons of different commercial systems, and the relation of measurements to age, ethnic groups and sex. The invention of FD-OCT increased the publication count once more: Various scan protocols and differing segmentation methods built-in by the manufacturers have to be validated for daily clinical use. Publications on the glaucoma diagnostic capability of OCT measurements are just as numerous. The majority of works in this field does focus on the correlation of a single measurement parameter, e.g. a mean layer thickness, with the disease. To name just a few examples: Polo et al. [Polo 08] used a Zeiss Stratus OCT system to measure RNFL thickness and papilla parameters and investigated the glaucoma discrimination ability of measurements computed over quadrants. Vizzeri et al. [Vizz 09] showed with case studies that local RNFL loss is visible on the printout of 3 commercial available OCT systems. Multiple publications compare the diagnostic value of the standard mean RNFL thickness obtained from the 3.46mm scan circle with other methods [Leun 10b, Na 11, Hwan 12, Na 12]. The overall outcome is that measures in the macula region are not as discriminative as the standard circle for glaucoma and volume measures around the ONH increase diagnostic performance.

In this work, we do not look at a single parameter, as for example the mean RNFL thickness, but at a bulk of parameters, the features computed from the layer thickness profiles. Automated algorithms should select diagnostically relevant ones and combine them to a single decision. There are only few publications for the specific task of glaucoma detection from OCT data with a classifier, but glaucoma classification was investigated on other modalities before. Especially visual field tests for the eye function, that do not image structure directly, have a history in being the basis of glaucoma classification systems. The numerical visual sensitivity plot with 53 locations combined with age is usually used as the feature set. Goldbaum et al. [Gold 02] compared several classifiers, e.g. linear discriminant analysis (LDA), support vector machines (SVM) and Bayesian classifiers of which the assumed feature probability distribution is a mixture of Gaussians (MoG). On a PCA-reduced set of features, i.e. the 53 visual field features were reduced to 8 dimensions, the MoG classifier performed best. Sample et al. [Samp 02] not only used the classification approach to classify current VF data to a diagnosis, but to predict glaucomatous change. Again, several classifiers like SVM and MoG were compared. The machine learning classifiers detected abnormalities in VF tests earlier than traditional methods. This result was confirmed by Wroblewski et al. [Wrob 09]. They proved that SVM classifiers have an

accuracy of 75% in detecting glaucoma suspects and pre-perimetric glaucoma. Given that the definition of pre-perimetric glaucoma is that no VF defects are exhibited, this result proves that classifiers can detect subtile changes before a defect is observed on the sensitivity plot by humans. Another remarkable approach of this work is the usage of a 2-stage classification: The VF tests were first classified into 3 diagnoses, ranging from normal, suspect and preperimetric glaucoma to glaucoma. The later two classes were then split into a refined diagnosis in a second classification stage. A large database with more than 2000 samples allowed for this detailed classification. A glaucoma score was proposed. Several other classification works have been based on VF data since then, e.g. Goldbaum et al. proposed a glaucoma progression score for VF data [Gold 12] and Bowd et al. combined SLO and VF data, i.e. structural and functional measurements for predicting glaucoma progression [Bowd 12].

Eye imaging modalities beside OCT that measure or image structure were also used to classify for the glaucoma disease. Swindale et al. proposed a glaucoma probability score (GSP) for the Heidelberg Retina Tomograph (HRT) in [Swin 00]. Features are generated by a parametric model fitted to the shape of the tomographic HRT data. The parameters of the model are used as the features. The ability of the GPS to discriminate normal subjects from glaucoma patients and to predict glaucoma progression was validated in multiple publications, e.g. [Alen 08, Masl 15]. HRT printout parameters were the features for a SVM classifier in [Zang 04]. For feature selection, "Forward selection and backward elimination" was used, as it is also proposed in this work. An 0.99 area under the receiver operating characteristic (ROC) curve was achieved for discriminating healthy eyes from eyes with early to moderate glaucomatous visual field damage. Asakoa et al. [Asao 14] used a random forest classifier on HRT parameters. The possible features from the imaging modalities like SLO, HRT and fundus imaging were expanded by works that were not based on the manufacturer's built-in parameters, which are usually based on segmentations on the data, but derive features from the images directly. Bock et al. [Bock 10] were the first to derive image features directly from fundus images without a segmentation. Non disease-relevant variations like illumination inhomogenities were removed in a pre-processing stage and the preprocessed images were then directly compressed by a PCA transform to gain a 30 dimensional feature vector for the glaucoma classification. In additon also fast Fourier transform and spline coefficient features were used. Dua et al. [Dua 12] and Mookiah et al. [Mook 12] followed a similar approach by using features derived from a wavelet transformation of fundus images. Zhu et al. [Zhu 14] computed shift invariant wavelets and used a kernel PCA with a non-normalized isotropic Gaussian kernel on combined SLO/HRT images to compute features directly out of the image data to detect abnormalities.

Similar to fundus, SLO and HRT images, OCT images show structure. But not only en-face views of the retina (fundus images, SLO) or topology (HRT), but also the depth layer structure is imaged. Layer segmentation algorithms were early introduced in commercial systems to derive diagnostic parameters, e.g. mean layer thickness values. These parameters provided by the commercial OCT systems are taken as features in all the published works on glaucoma detection from OCT data. The first two are from Burgansky et al. [Burg 05] and Huang et al. [Huan 05]. Both used the printout parameters the Zeiss Stratus TD-OCT provided as features and tested

multiple classifiers like decision trees, SVM, Mahalanobis distance to the healthy and patient feature distributions, LDA and artificial neural networks (ANN). Burgansky et al. reported that the SVM machine was the best classifier with 0.92 area under the receiver operating characteristic (ROC) curve. Huang et al. even had a 0.991 area under the ROC with the decisions made according to Mahalanobis distances. These impressive numbers can be explained with the limited data of both studies: All glaucoma patients already had visual field loss, similar to the data in the HRT glaucoma classification study [Zang 04] mentioned earlier. After these initial publications based on TD-OCT data, the next relevant work was based on FD-OCT data, by Baskaran et al. [Bask 12]. An extensive database similar in size to the one utilized in this work with similar diagnoses (healthy, mild and moderate glaucoma) was the basis for the classification with LDA and classification-and-regression-tree classifiers. Again, the features used were parameters provided by a commercial system, the Zeiss Cirrus OCT. The classifiers outperformed the single parameters in their diagnostic capability. A similar method was proposed by Mwanza et al. [Mwan 13] using parameters from the Zeiss Cirrus OCT and a logistic regression classifier. Both publications rejected scans with poor image quality in their database and Mwanza et al. also rejected scans with segmentation errors. While not being aimed at glaucoma classification but multiple sclerosis (MS), the work of Garcia-Martin et al. [Garc 12] should be mentioned. Multiple mean values of the RNFL thickness profile from the HE Spectralis were used as features for LDA classification. Scans with poor quality were rejected, but the authors explicitly mention the problem that in daily clinical practice this is not always feasible. As mentioned in Chapter 2 we use all the data in the database regardless of quality. The scan quality does not impact the probability of a segmentation error for the presented algorithm. Scans with segmentation errors are not excluded from the classification experiments, but it is evaluated later on whether the correction of these errors affects the classification results.

Some works that classify on OCT data have also to be mentioned: Qi et al. [Qi 10] use image features, namely intensity and speckle distribution, as well as stripe orientation, to find dysplasia in Barrett's esophagus on endoscopic OCT images. Liu et al. [Liu 11] also classify on texture and shape features computed directly from OCT image data. The images are grouped to macular pathologies: Macular edema, macular hole and AMD. Zhang et al. [Zhan 13] derive a virtual VF chart from RNFL and GC/IPL thickness features of OCT scans with linear regression and prove that they are as sensitive for glaucoma detection as real VF data. Finally, Belgith et al. [Belg 15] detect glaucomatous change on intensity distribution change maps, i.e. they also use the image data directly. An overview over all the published research on classification tasks from OCT data known to the author is given in Tables B.6 and B.7 in the Appendix.

The classification system presented in this work is similar to the ones in [Burg 05, Huan 05, Bask 12, Mwan 13, Garc 12] with respect to that retinal layer thickness profiles, or features computed from those are the input to the classifier. However, there is a major difference: The former works had to use the parameters that are provided by a commercial OCT system and thus had no access to the complete thickness profiles of layers, with the exception of Garcia-Martin et al. [Garc 12], but they only computed mean features from the RNFL profile. By implementing a segmentation

algorithm, this work is not bound to the restrictions and limitations of commercial systems. Features can be computed for multiple layers, even if their significance for the glaucoma disease is doubtful. Furthermore, not only common mean features are computed, but various types, including ratios and PCA features. No human expert decides on the discriminative value of a feature, but algorithms, the desicion of which is based on training data, i.e. we present a "data mining" method for the glaucoma classification challenge from OCT data.

4.2 Layer thickness normalization

The basis for the classification are retinal layer thickness profiles. The thickness profiles are computed out of the positions of the inner and outer segmented boundaries of the respective layer, as given in Equation 3.16. From the 6 segmented boundaries 5 thickness profiles and in addition the whole retinal thickness are generated:

$$\begin{aligned} LT_{Retina}(r) &= L_{RPE}(r) - L_{ILM}(r); \\ LT_{RNFL}(r) &= L_{ONFL}(r) - L_{ILM}(r); \\ LT_{GCL+IPL}(r) &= L_{IPL/INL}(r) - L_{ONFL}(r); \\ LT_{INL+OPL}(r) &= L_{OPL/ONL}(r) - L_{IPL/INL}(r); \\ LT_{ONL+ELM}(r) &= L_{ELM/IP}(r) - L_{OPL/ONL}(r); \\ LT_{IP+OP+RPE}(r) &= L_{RPE}(r) - L_{ELM/IP}(r); \end{aligned} \tag{4.1}$$

The blood vessel indices are an additional seventh "virtual" layer thickness LT_{BV}, i.e. they do not reflect real thickness values but are treated as one. The thickness profiles can be further processed in two ways. First, Bendschneider et al. [Bend 10] measured the mean RNFL thickness in healthy eyes and found that it decreases with age. This leads to the idea to perform a novel **age normalization** of the thickness profiles. Second, the HE Spectralis scans the circle around the ONH with a fixed opening angle of the laser beam. Given the radius of the curvature of the corneal surface of the eye scanned, which may differ from eye to eye, it computes the assumed ocular magnification factor and the spacing of A-Scans in the R-direction. To compensate for these different assumed spacings of the A-Scans for each eye is a new **magnification normalization**. The two layer thickness normalization methods are detailed in the following:

Age normalization: Bendschneider et al. [Bend 10] used Spectralis RNFL segmentations to relate the overall, quadrant and mean RNFL thickness in 32 segments to age. Given the segmentation data in this work, this approach can be extended: All segmented layer groups can be investigated. In Figure 4.1, the mean layer thickness of the 453 healthy eyes of the classification database is shown for 3 example layer groups in relation to age. The manually corrected data is used. A least-squares line fit through all the data samples is shown in addition to single data samples. It can be observed that not only the RNFL, but also the other layer groups become thinner with age. The fitted lines and thus the thinning correlate with the data samples with $P < 0.01$ for all layer groups and the retina.

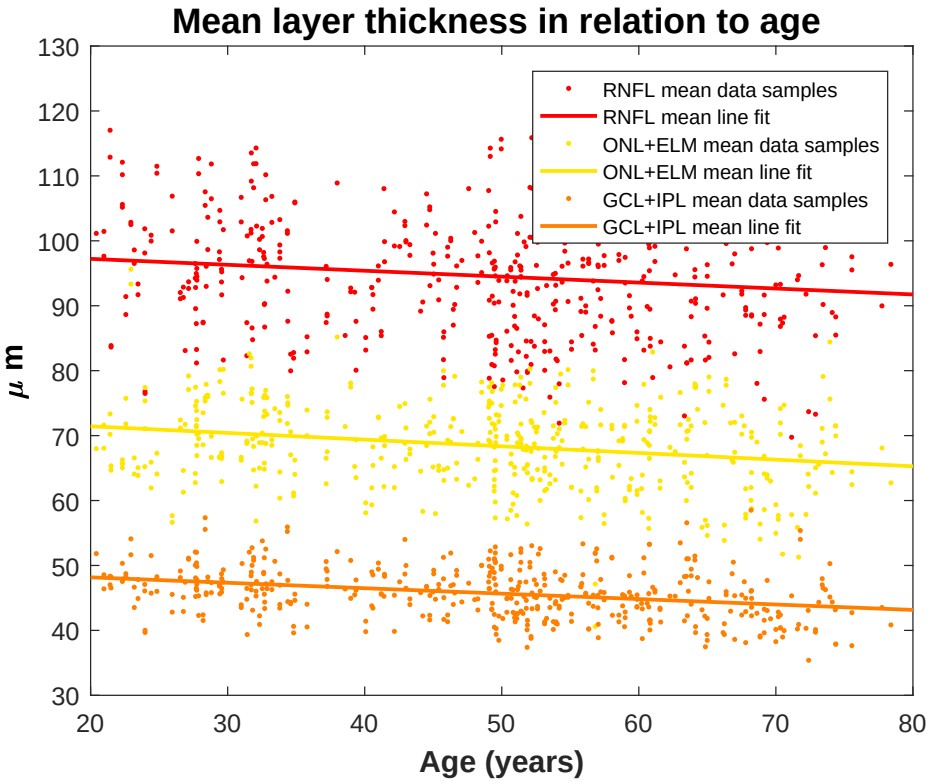


Figure 4.1: Mean thickness of the RNFL, ONL+ELM and GCL+IPL of healthy eyes in relation to subject age. The lines are computed by an least-squares fit through all the data samples of the respective layer on manually corrected data. Mean over all data samples RNFL thickness: $94.60 \pm 9.20\mu$m, mean ONL+ELM thickness: $68.49 \pm 6.70\mu$m, mean GCL+IPL thickness: $45.77 \pm 3.95\mu$m. All line fits correlate with the data samples with $P < 0.01$.

To quantify the age-related thinning further, least-square line fits with respect to age are computed not only on the overall mean thickness as shown in Figure 4.1, but on the mean thickness of 32 segments along the A-Scans, as proposed by Bendschneider et al. [Bend 10]. The segments are numbered from the left side of the scan, i.e. the middle of the temporal quadrant, to the right from 1 to 32. The fitted lines can be represented with an offset and gradient value:

$$\bar{LT}_i(a) = o_i + g_i \cdot a; \tag{4.2}$$

where $\bar{LT}_i(a)$ is the line fit through the mean values of healthy eyes for the respective layer (the layer subindex is omitted in this and following equations for better readability) and segment number i, i.e. it can be interpreted as a representive mean thickness value for age a. The offset value is o_i and the scalar gradient g_i. The gradient g_i, i.e. the expected change with respect to age is of most interest. If it is negative, the layer thins with age. The descriptive statistics (mean, standard deviation, minimum and maximum value) of the thickness gradients g_i in the 32 segments are given in Table 4.1. Except for the RNFL, all 32 gradients of the segments are negative for each layer group, i.e. the layers thin over time. Figure 4.2 plots the g_i values over

Layer group	Mean±Std.	Min	Max
Retina	-0.34 ± 0.06	-0.47	-0.21
RNFL	-0.09 ± 0.08	-0.26	0.05
GCL+IPL	-0.08 ± 0.03	-0.15	-0.02
INL+OPL	-0.03 ± 0.01	-0.05	-0.00
ONL+ELM	-0.10 ± 0.02	-0.13	-0.05
IP+OP+RPE	-0.03 ± 0.01	-0.05	-0.01

Table 4.1: Descriptive statistics of the line fit gradients g_i on scans of healthy subjects in relation to age, computed in 32 segments along the A-Scans. The values are given in [μm/year] units. Mean, Std., Min, and Max refer to the values of the 32 segments, e.g. the Mean column shows the mean gradient value $\frac{1}{32}\sum_{i=1}^{32} g_i$ for the respective layer groups.

the segments i for 3 layer groups as examples. The RPE gradients are very small in their absolute values, the GCL+IPL gradients have larger absolute values and are all negative. The RNFL has a different gradient curve along the segments: While in the temporal, superior and inferior quadrants the RNFL thins over time, the nasal quadrant remains constant or grows slightly. This confirms the results of [Bend 10], who made the same observation on a smaller dataset and with a different segmentation method. The fine scale distribution of the gradients along the 32 segments does not match the results from [Bend 10]. This can be explained with different segmentation methods used, yielding different results especially in BV regions.

A possible age normalization is constructed as follows:

$$LT_{age}(r) = LT_r \cdot \frac{\bar{LT}_i(a_{ref} - (a_s - a_{ref}))}{\bar{LT}_i(a_{ref})}, \text{for } r \in S_i \tag{4.3}$$

where $LT_{age}(r)$ is the age-normalized thickness value. A-Scan position r is within the range of segment S_i. The arbitrary reference age a_{ref} is set to 50 in this work. The age of the subject the scan was taken from is a_s. The normalization scales the thickness with a scalar factor linearly dependent on the distance from the actual age a_s to the reference age a_{ref}, e.g. when the gradient g_i of $\bar{LT}_i(a)$ is negative and the age a_s is older than the reference age, the thickness value $LT(r)$ will be enlarged. If we consider the absolute gradient values of Table 4.1, we expect only a minor altering of the thickness values. For example the mean gradient of the fit to the RNFL mean thickness values is -0.09μm per year. Compared to the mean and standard deviation of the RNFL thickness values of healthy eyes, which is $94.60 \pm 9.20\mu$m, this is only a very small value, i.e. an age difference of over 100 years from the subject age to the reference age would be necessary to equal the standard deviation in healthy eyes. The evaluation in Section 4.5.3 will show wheter the age normalization has any effect on classification results at all.

The thickness normalization and the feature computation described in the next section take place before the cross-validation experiments are constructed, first of all to avoid extensive computation times. Features are computed once for each possible thickness normalization and stored, and not separately in each single cross-validation

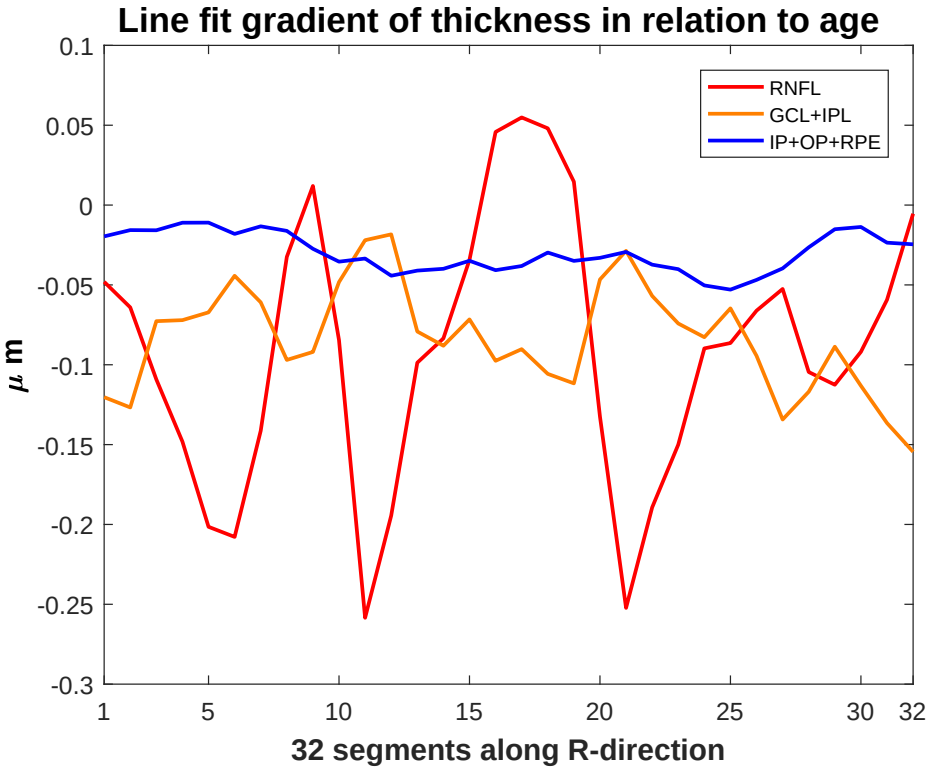


Figure 4.2: Gradients g_i of the line fits $\bar{LT}_i(a)$ to mean thickness values of healthy eyes in 32 segments (denoted by i) for 3 example layer groups, the RNFL, GCL+IPL and IP+OP+RPE.

run. Second, this procedure eases the code structuring and is common practice. But special care has to be taken. Numbers (e.g. line parameters for the age correction or features) computed not on single scans but on groups might violate the cross-validation rule that training and test data are strictly separated. This is the case for the age normalization. By using *all* healthy data of the database for computing an age correction factor, we will violate the strict separation of training and test data in the cross-validation of the classification experiments. This is valid, as Bendschneider et al. [Bend 10] found similar results for the RNFL on a much smaller database (170 scans of healthy subjects instead of 453 in this work). Concrete numbers do not match due to the different databases and segmentation methods used. The mean thickness and absolute gradient values with respect to age computed in this work are smaller. However, general observations, i.e. a thinning of the RNFL except in the nasal quadrant, confirm each other. Therefore, it can be assumed that a reduction of the samples by 10% in strict cross-validation runs would not alter results significantly.

Focus normalization: As written before, the HE Spectralis scans the circle around the ONH with a fixed opening angle of the laser beam. Given the radius of the corneal surface of the eye derived from the focus setting of the Spectralis and the ametropia measured using special instruments, it computes the assumed ocular magnification factor with the Garway-Heath formula [Garw 98] and the spacing of A-Scans in R-direction given this ocular magnification factor. If no radius of the corneal

surface is set by the operator a default value is assumed. The pixel spacing in the R-direction $Scale_R$ is stored in the VOL files exported from the system. The mean and standard deviation of $Scale_R$ are $14.50 \pm 0.56\mu$m among the whole classification dataset. We normalize the thickness profiles by this eye-specific factor $Scale_R$:

$$LT_{mag}(r) = LT(r) \cdot Scale_R; \tag{4.4}$$

The thickness profile $LT_{mag}(r)$ at A-Scan position r thus represents an area measure. Both the age normalization and the magnification normalization can be combined to $LT_{mag,age}(r)$. We compute the magnification normalization first, followed by the age normalization.

4.3 Feature computation

Out of the 7 layer thickness profiles, i.e. 5 layer groups, the complete retina and the virtual thickness of the blood vessel indices, the features for the classifier are generated. The features are **standard measures over the complete profile**, **means in sections**, **ratios in sections** and **PCA features**. The features are defined in the following:

Standard measures over the complete profile: The mean, minimum, maximum and median values are computed over the whole profile:

$$\begin{aligned} f_{mean} &= \frac{1}{\#R} \sum_{r \in R} LT(r), \\ f_{min} &= \min_{r \in R} LT(r), \\ f_{max} &= \max_{r \in R} LT(r), \\ f_{med} &= \operatorname*{median}_{r \in R} LT(r); \end{aligned} \tag{4.5}$$

For a specific layer group, the layer group name will be denoted as an additional subindex to the feature f, e.g. $f_{GCL+IPL,mean}$.

Means in sections: The mean value is not only calculated for the whole thickness profile, but also in quadrants and the 32 segments already used in Section 4.2 for the age normalization:

$$f_{mean,i} = \frac{1}{\#S_i} \sum_{r \in S_i} LT(r) \tag{4.6}$$

where $\#S_i$ is the number of A-Scans in segment S_i. The means in the segments are denoted by an additional subindex in the following, e.g. $f_{RNFL,mean,27}$ for the mean RNFL thickness in segment 27. The means in the quadrants are denoted with the quadrant's first letter, e.g. $f_{RNFL,mean,n}$ for the mean RNFL thickness in the nasal quadrant.

Ratios in sections: The mean in segments from a specific layer is combined with the mean in the same segment from the retina by calculating the proportion the layer occupies inside the retina:

$$f_{ratio,i} = f_{mean,i} / f_{Retina,mean,i} \quad ; \tag{4.7}$$

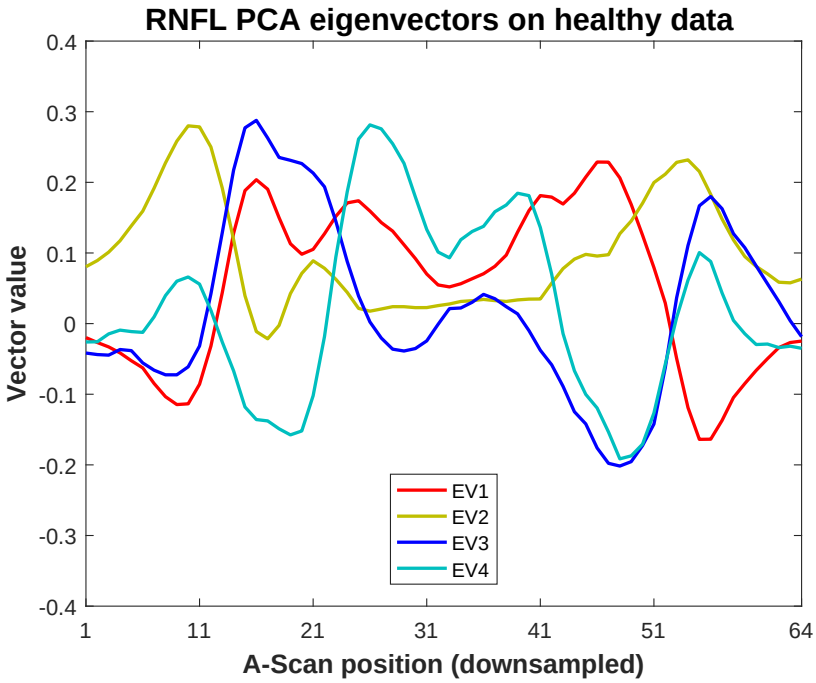
Again, these features are computed for the quadrants and 32 segments. The ratio features do not make sense in case of the retina, as they always yield 1, and the blood vessel indices. However, they are taken into the complete feature set for the sake of structural simplicity.

PCA features: The principal component analysis (PCA) results in a transformation that converts a set of observations, e.g. vectors of data samples, into a set of values with linearly uncorrelated variables. The transformation matrix is composed of the so called PCA eigenvectors (EV). They are ordered by descending variance of the $1D$-projection of the set of observations to them. The PCA EV are the EV of the covariance matrix of the data set vectors. The principal components build an uncorrelated orthogonal basis set of the data set space. PCA is a tool that is often used to reduce the number of data samples to build feature vectors with less entries than the original data samples, but covering most of the information. We use the PCA to reduce the dimension of the layer thickness profiles $LT(r)$ from 768, as there are 768 possible thickness values at the A-Scan positions r, to a lower 10-dimensional space.

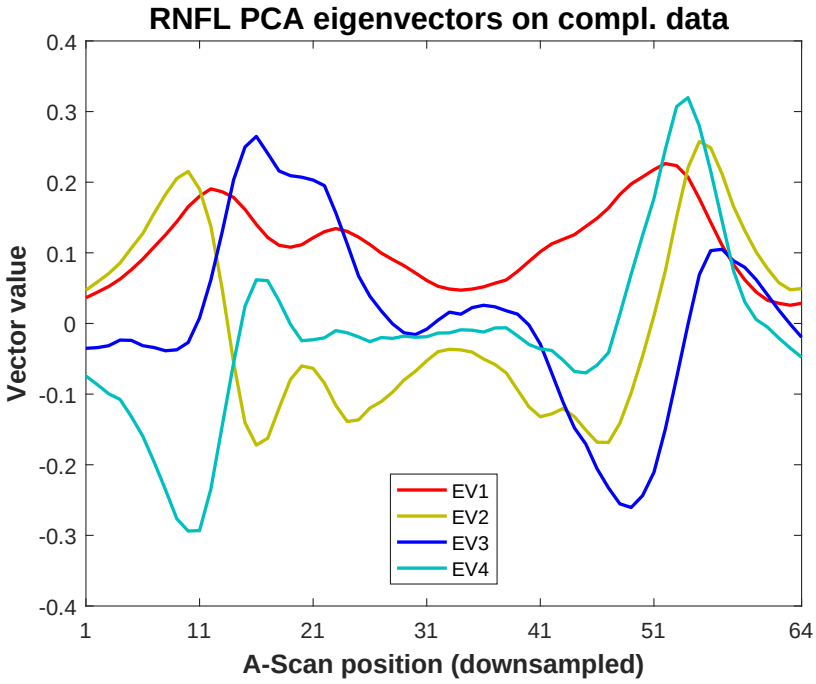
Before computing the PCA, the thickness profiles are downsampled by computing the mean in 64 segments to reduce small scale variance. The PCA EV and therefore the transformation matrix is calculated from two datasets: On the complete classification database and on healthy eyes only. The PCA transformation calculated on the complete classification database therefore captures most of the variance of both healthy as well as glaucomatous eyes, while the PCA transformation calculated on healthy eyes captures only most of the variance of healthy eyes in the first few EV. The first 10 principal components of both PCA transformations for each thickness profile are taken as features $f_{PCAall,i}$ and $f_{PCAhealthy,i}$.

The PCA EV can be plotted to visualize the areas of highest variance. Examples are given in Figures 4.3 and 4.4. Figure 4.3 a) shows the first 4 EV of the PCA transformation for the RNFL thickness profile computed on healthy eyes. Most of the variance of the RNFL thickness on healthy eyes is centered in the middle of the OCT scan, i.e. the inferior, nasal and temporal quadrants. Compared to the first 4 EV of the PCA transformation computed on the complete classification dataset plotted in Figure 4.3 b), some observations can be made: The EV change, but the first 3 EV are similar in shape. The highest variance, as indicated by EV 1 is more local on healthy data than on all the eyes - most likely due to the RNFL thinning as a result of the disease. EV 4 is completely different on the PCA transformations of the two datasets bases. For the EV of the GCL+IPL plotted in Figure 4.4, the first 4 EV have similar shape for both datasets, with the first 2 being near to identical. This is an indicator that the general shape of the GCL+IPL thickness profile is not altered to such an extent by glaucoma than the RNFL thickness profile shape.

Computing the PCA EV on all the data or all healthy eye data is a critical violation of the cross-validation rule of strict separation of training and test data. No

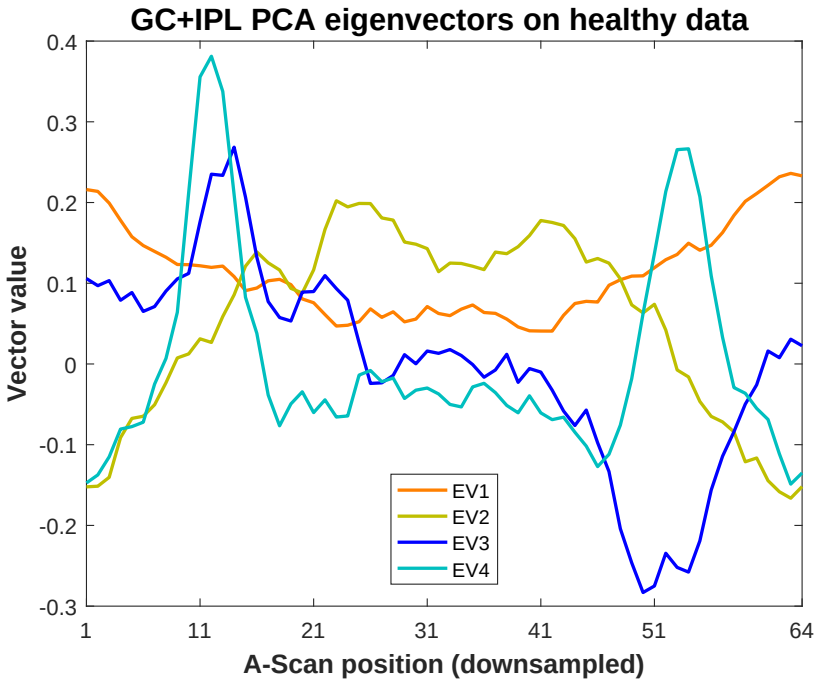


(a)

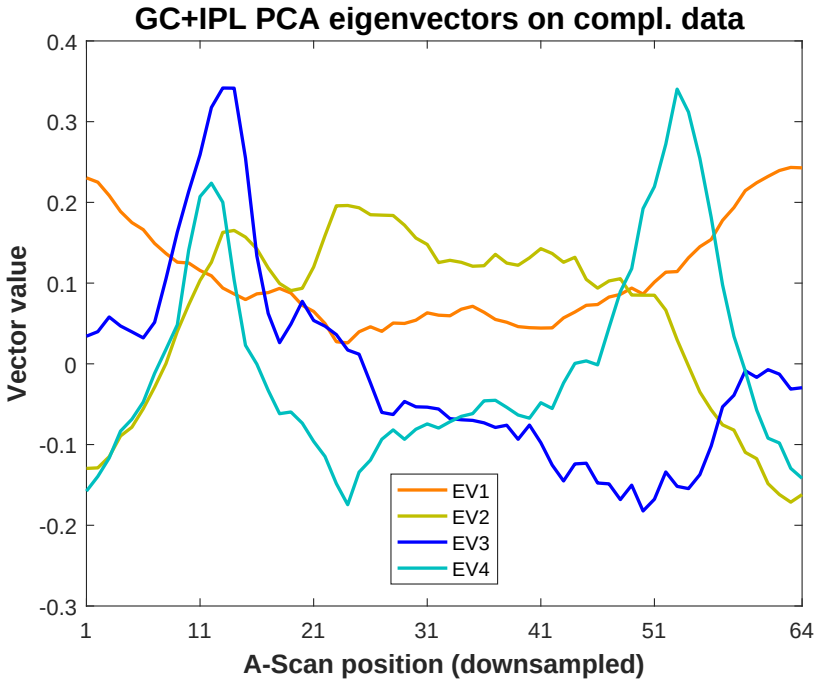


(b)

Figure 4.3: The first 4 PCA eigenvectors of the RNFL thickness profiles. (a) PCA transformation computed on healthy eyes. (b) PCA transformation computed on complete classification dataset.



(a)



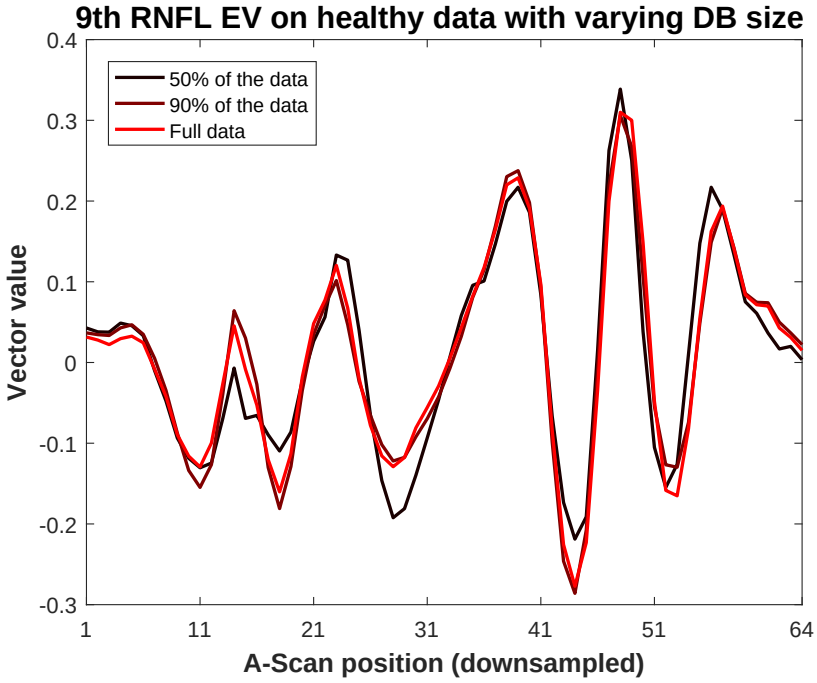
(b)

Figure 4.4: The first 4 PCA eigenvectors of the GC+IPL thickness profiles. (a) PCA transformation computed on healthy eyes. (b) PCA transformation computed on complete classification dataset.

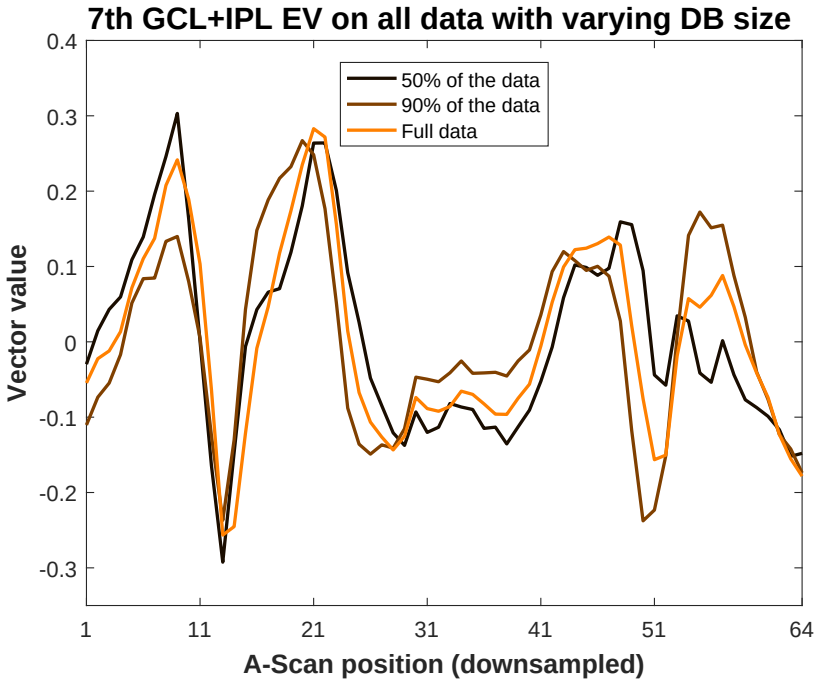
preceding work from literature justifies this and thus it must be further investigated, as a modification of the data basis for the PCA cannot only alter features slightly, but may also change the shape of the PCA EV and yield features with a different information content. To verify the validity of the computation on the complete datasets, the PCA is computed on datasets of reduced size. A random selection of 90% of the eye scans in the original 2 PCA datasets is done. The PCA EV are computed on this reduced datasets, which are of the size of a training data set in a 10-fold cross-validation. As this single random selection is only a single example that does not capture all the possible effects in the databases, also a random reduction to 50% is performed. The Euclidean vector distances of the first 10 PCA EV, i.e. the PCA transformation matrix entries that produce the PCA features, computed on the original full dataset to the PCA EV computed on the randomly reduced datasets are calculated. The mean and standard deviation of these distances are shown in Table 4.2. The numbers are generally very small, e.g. the PCA vectors on full and 50% reduced dataset size differ on average by 0.26 for the RNFL and by 0.42 for the GCL+IPL. However, the pure number does not allow a conclusion that the general shape of the PCA EV did not change. Therefore case examples are given in Figure 4.5. The 9th RNFL EV plotted in Figure 4.5 a) has the largest vector distance of the first 10 RNFL EV from full dataset size to a database size of 90% of the original data. The vector distance from full to reduced database size is 0.14 for 90% and 0.35 for 50% of the data. The shape of even the EV computed on 50% of the data does not differ significantly from the one computed on the complete data. The 7th GCL+IPL EV plotted in Figure 4.5 b) has the largest vector distance of the first 10 GCL+IPL EV from full dataset size to a database size of 90% of the original data and is generally among the EV with the largest vector distances for all layer groups. The vector distance from the full to reduced database size is 1.27 for 90% and 0.40 for 50% of the data. The vector for 50% of the data is more similar to the original one than the one with 90% of the data, which differs in shape in the inferior quadrant. This can only be explained with random effects in the database. However, as said, this is an extreme example. It can therefore be concluded that the computation of the PCA transformation on the complete classification database or all the healthy eyes do not alter the features significantly, compared to a strict cross-validation.

In total, 672 features are computed. For each of the 7 layer groups (retina, 5 retina layer groups, and the BV indices), there are 4 standard features, 4 means in quadrants, 32 means in segments, 4 ratios in quadrants, 32 ratios in segments, 10 PCA features with the PCA being computed on all data of all diagnoses, and 10 PCA features with the PCA being computed on healthy eyes only. 96 features are thus computed for each layer group. $96 \cdot 7$ yields the 672 features.

After the computation of the features, the 672 entries of the feature vector are separately normalized, i.e. linearly scaled such that each entry has zero mean and a standard deviation of 1 on the complete data set. Again, this violates the strict cross-validation. However, this normalization is common practice to avoid numerical problems in the classifier training. It must be noted that the resulting zero mean and norm standard deviation hold only for the complete dataset. Data selections as performed in the cross-validation or by diagnosis groups may not have zero mean and norm distributed features after this normalization.



(a)



(b)

Figure 4.5: Comparison of PCA eigenvectors (EV) computed on datasets with different size. Dataset sizes were reduced by random removal of data. (a) 9th RNFL EV computed on healthy eyes. The 9th RNFL EV has the largest vector distance of the first 10 RNFL EV from full dataset size to a database size of 90% of the original data. The vector distance from full to reduced database size is 0.14 for 90% and 0.35 for 50% of the data. (b) 7th GCL+IPL EV computed on data of all diagnoses. The 7th GCL+IPL EV has the largest vector distance of the first 10 GCL+IPL EV from full dataset size to a database size of 90% of the original data. The vector distance from the full to reduced database size is 1.27 for 90% and 0.40 for 50% of the data.

Boundary	90% All	90% Healthy	50% All	50% Healthy
Retina	0.13 ± 0.07	0.14 ± 0.10	0.20 ± 0.19	0.40 ± 0.23
RNFL	0.07 ± 0.04	0.10 ± 0.06	0.24 ± 0.25	0.31 ± 0.14
GCL+IPL	0.38 ± 0.44	0.17 ± 0.09	0.35 ± 0.19	0.67 ± 0.40
INL+OPL	0.13 ± 0.08	0.19 ± 0.10	0.33 ± 0.14	0.54 ± 0.33
ONL+ILM	0.15 ± 0.10	0.16 ± 0.10	0.74 ± 0.52	0.36 ± 0.20
IP+OP+RPE	0.18 ± 0.18	0.17 ± 0.17	0.79 ± 0.42	0.47 ± 0.39
BV	0.42 ± 0.28	0.41 ± 0.15	0.91 ± 0.34	1.07 ± 0.39

Table 4.2: Mean and standard deviation of the Euclidean vector distance of the first 10 PCA basis vectors obtained from the complete data to the PCA basis vectors obtained from a random selection of 90% or 50% of the data. The PCA is performed on all the diagnoses (All) or only thickness profiles from healthy eyes.

4.4 Classification and feature selection

The classification experiments are carried out with a 10-fold cross-validation. The training and testing of the classifier happens 10 times, each time with a random selection of 10% of the classification data for testing and the rest for training. The random selection is done such that after all the cross-validation runs, each data sample was used once for testing and 9 times for training. If data is obtained from both eyes of a glaucoma patient, the glaucomatous damage and specifically the RNFL measurements are correlated between both eyes [Bert 09]. As the classification dataset contains subjects with both eyes imaged, the selection of data samples for the cross-validation runs is made in a patient-sensitive manner: In a single cross-validation run, the eyes of one subject are either in the test or training data, and not split between both, to avoid the advantage of test data subject knowledge from training.

A classification task is defined by selecting a pair, multiple or a combination of diagnoses as classes. In general, all the data for these diagnoses are then used to construct the cross-validation, i.e. no sex or age information is taken into account. Sex information is not present in the dataset. Performing age matching in each classification experiment would reduce the dataset size in some classification tasks significantly. However, it will be of interest whether differing age distribution in the classes influences the results. Therefore an age matching selection by decades preceding the cross-validation construction is implemented for 2-class classification challenges. The datasets are randomly ordered. One data sample is taken from the first class. If a data sample with the same age decade is found in the second class, both are taken into the classification experiments. Otherwise the data sample from the first class is taken out of the experiment. The selection of a data sample from the first class and searching for a age-matched sample in the second class is repeated until all data samples in the first class have been looked at once. Classes are roughly age-matched after the procedure and equal in size. The age matching is used twice in the results Section 4.5 and specifically noted.

We compare 3 different classifiers: **Linear naïve Bayes** (Bayes), **k-nearest neighbor** (kNN) and **linear support vector machines** (SVM). This selection is

not random: The Bayes classifier is common in classification experiments and yields a baseline result. The kNN classifier is the most widely used representative of parameter free classifiers. The SVM is still a state-of-the-art classifier that has proven to give best results for various tasks. Other and mathematically more advanced classifiers, e.g. neural networks, boosting trees, and random forests might enhance classification results and have to be taken into account if the goal of a work is to reach best classification scores at any cost. In this work, the challenge definition, influence of layer thickness profile normalization and segmentation correction as well as the feature selection are the centers of research. Linear naïve Bayes, kNN and linear SVM have all necessary properties that allow for this study and some advances more, e.g. an easy geometric interpretation of the classification boundary in feature space. In preliminary tests, naïve Bayes with quadratic decision boundary and SVM with polynomial and radial basis function (RBF) kernels have been investigated. They did not improve results compared to the linear classifiers, most likely due to a too small dataset size for a robust parameter estimation. In the case of the SVM with RBF kernel, overfitting to the training data was the reason for worse results compared to the linear SVM. The classifiers will now by explained in a few words. For more detailed insight in classifiers and data mining see [Fayy 96, Duda 00, Hast 03, Niem 03]:

Linear naïve Bayes: The naïve Bayes classifier assumes that the entries of the feature vector $\mathbf{f}$ are independent of each other. The decision rule is the Bayes classifier decision rule. The class with the maximum a posteriori probability is chosen:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|\mathbf{f}) = \underset{y}{\operatorname{argmax}}\, p(y) \prod_{i=1}^{\#f} p(f_i|y) \tag{4.8}$$

where y is the class number out of k possible classes ($y \in \{1, ..., k\}$). The entries of the feature vector $\mathbf{f}$ with dimension $\#f$ are f_i. The class the classifier decided for is $\hat{y}$. The posterior probability that the given feature vector $\mathbf{f}$ belongs to class y is $p(y|\mathbf{f})$. The class prior $p(y)$ is set to be equal for all classes in the classification experiments of this work and therefore omitted. The posterior $p(y|\mathbf{f})$ can be written as the multiplication of the posteriors of the single feature entries $p(f_i|y)$ due to the independence assumption. In addition to the independence of the entries of the feature vector, the linear naïve Bayes classifier assumes Gaussian distributed class conditionals $p(\mathbf{f}|y)$ that share the same covariance matrix for every class. The posteriors are estimated by a maximum-likelihood estimation. This breaks down to estimating the mean and variance of each feature vector entry f_i independently, which then can be inserted into the normal distribution equation to yield the posterior $p(f_i|y)$.

k-Nearest neighbor: The kNN classifier is a non parametric classifier. For a feature vector $\mathbf{f}$ from the test data, the k nearest neighbors in feature space $\mathbf{f}_{n,1...k}$ are searched in the training data. "Nearest" means that a norm distance measure is computed between $\mathbf{f}$ and all feature vectors from the training data. Most of the times the Euclidean distance, i.e. the L2 norm, is utilized. The $\mathbf{f}_{n,1...k}$ are the feature vectors from the training data with smallest distance. A majority vote on the classes y that are given for $\mathbf{f}_{n,1...k}$ determines the assignment of the class $\hat{y}$ to $\mathbf{f}$.

Support vector machine: The support vector machine is a classifier for 2-class problems, but may be extended to multiple classes by various strategies, e.g. multiple one-class-against-all-other-classes classifiers with majority voting. Two parallel hyperplanes sharing the same normal vector (e.g. lines in case the feature space has 2 dimensions) are fitted such through the feature space, that the data samples belonging to each class are separated by the gap in between the hyperplanes. The gap is as wide as possible. If the classes are linearly separable, these hyperplanes are unique. The hyperplanes can be described by the feature vectors that lie nearest to them and that define the gap in between the hyperplanes. These are called "support vectors". The hyperplane halfway trough the gap with the same normal vector as the gap defining hyperplanes is called the maximum margin hyperplane. The optimization problem of making the gap in between separating hyperplanes as wide as possible is equivalent to minimizing the norm of the maximum margin hyperplane:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}} && \|\mathbf{w}\| \\ &\text{subject to} && y_k(\mathbf{w}\cdot\mathbf{f}_k + b) \geq 1 \; \forall \; 1 \leq k \leq \#TD; \end{aligned} \tag{4.9}$$

The maximum margin hyperplane $\mathbf{w}\cdot\mathbf{x} + b = 0$ is defined by its normal vector $\mathbf{w}$, bias b and $\mathbf{w}\cdot\mathbf{x}$ being the dot product of $\mathbf{w}$ and $\mathbf{x}$. The class assignment for a feature vector $\mathbf{f}_k$ of the training data with identifier k is y_k, which can only take the values 1 and -1. As mentioned above, the SVM in its base form is suited only to 2-class problems. The number of feature vectors in the training data is $\#TD$. The constraints in the equation 4.9 ensure that the maximum margin hyperplane separates the two classes perfectly.

To classify data that is not linearly separable in feature space, a loss function ξ_k is introduced that is 0, if the training feature vector $\mathbf{f}_k$ is on the side of the maximum margin hyperplane of the given class, and the Euclidean distance to the maximum margin hyperplane otherwise. The SVM training problem is now to minimize the loss function summed over all training data samples, but keeping the gap in between the separating hyperplanes as wide as possible:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}} && \frac{1}{2}\|\mathbf{w}\| + C\sum_{k=1}^{\#TD}\xi_k \\ &\text{subject to} && y_k(\mathbf{w}\cdot\mathbf{f}_k + b) \geq 1 - \xi_k \; \forall \; 1 \leq k \leq \#TD; \end{aligned} \tag{4.10}$$

The tradeoff between those goals is controlled by the SVM regularization parameter C. Various optimization strategies can be used to solve the problem. The SVM can be extended to non-linear separating functions by applying a kernel transformation, i.e. implicitly lifting the feature vectors into a higher dimensional space by generating new feature vectors that are nonlinear combinations of the entries of the original ones. This lifting is only implicit as only the dot product of two vectors in the higher dimensional space must be computed for the minimization problem, the so called kernel. This kernel replaces the dot product in the original space. As already mentioned, we restrict the usage of SVM in this work to linear SVMs and do not apply kernel transformations.

GS/Res.	y_1	y_2	y_3
y_1	n_{11}	n_{12}	n_{13}
y_2	n_{21}	n_{22}	n_{23}
y_3	n_{31}	n_{32}	n_{33}

(a) Confusion matrix

GS/Res.	y_1	y_2	y_3
y_1	$\frac{n_{11}}{\#y_1}$	$\frac{n_{12}}{\#y_1}$	$\frac{n_{13}}{\#y_1}$
y_2	$\frac{n_{21}}{\#y_2}$	$\frac{n_{22}}{\#y_2}$	$\frac{n_{23}}{\#y_2}$
y_3	$\frac{n_{31}}{\#y_3}$	$\frac{n_{32}}{\#y_3}$	$\frac{n_{33}}{\#y_3}$

(b) Confusion matrix (normed)

Table 4.3: Example of a confusion matrix result of a 3 class classification experiment. y_i are the classes, $\#y_i$ the number of data samples of the respective gold standard (GS) class in the test data set. n_{ij} are the number of data samples with GS class i assigned to class j by the classifier. The classwise averaged classification rate is given by $\frac{1}{3}\sum_{i=1,2,3}\frac{n_{ii}}{\#y_i}$ for 3 classes.

From the possible feature selection methods (see [Guyo 03, Niem 03]), e.g. "Select the best features independent of each other", "Select the next best feature" and "Feature elimination", preliminary tests have shown that one method surpasses the others on our data. The feature selection is thus performed by "Forward selection and backward elimination" only. It has to be noted that the linear naïve Bayes classifier performs slightly better in the experiments when no feature selection is used, i.e. all 672 features are the input to the classifier. The difference between the results are only minor, and for a general comparability of the results, we decided to use feature selection in every experiment.

The algorithm "Forward selection and backward elimination" is written in pseudo code in Figure 4.6. It is described as follows: Classifications are performed by using the training data of the current cross-validation run for training and testing during the feature selection process. Starting with an empty feature vector, the feature with the best classification rate increase with respect to the currently selected features is added. Then, all selected features are tested to see whether an elimination increases the classification rate. If so, the feature that contributes the least, i.e. increases the classification rate most when removed, is eliminated from the feature vector. The elimination of features is repeated until no feature can be removed without decreasing the classification rate. Then, the algorithm iterates again with the addition of a feature to the feature vector. This procedure of adding a feature and then testing if one or more features can be eliminated is repeated until a predefined number of features is reached or the algorithm converges, e.g. the feature set does not change from one iteration to the next. The feature selection process is performed for each cross-validation run separately. The classifier used in the feature selection is the classifier used for the overall classification experiment. In the experiments, the maximum number of selected features was set to 30. However, this number was never reached as the selection process always converged before.

The results of a classification experiment are described with multiple measures. For all experiments the results are shown in a confusion matrix, i.e. a tabular overview of the classification results on the test data. The confusion matrix is the sum of the confusion matrices of the single cross-validation runs. An example confusion matrix for a 3-class classification experiment is shown in Table 4.3. Classes are denoted by y_i.

```
fSelected = empty vector;
fSelectedLastIter = undefined;
cRateBest = undefined;
while size(fSelected) < MAXF and fSelected ≠ fSelectedLastIter do
  fSelectedLastIter = fSelected;
  cRate = vector of size #f filled with 0;
  for all f do
    if f is not in fSelected then
      fSelectedTemp = fSelected and f;
      cRate[f] = trainAndTest(fSelectedTemp);
    end if
  end for
  fBestAddition = positionOfMaximum(cRate);
  if cRate[fBestAddition] > cRateBest then
    fSelected = fSelected and fBestAddition;
    cRateBest = cRate[fBestAddition];
  end if
  removeHappened = true;
  while removeHappened do
    cRate = vector of size #f filled with 0;
    for all f in fSelected do
      fSelectedTemp = fSelected without f;
      cRate[f] = trainAndTest(fSelectedTemp);
    end for
    fBestRemove = positionOfMaximum(cRate);
    if cRate[fBestRemove] > cRateBest then
      fSelected = fSelected without fBestRemove;
      cRateBest = cRate[fBestRemove];
    else
      removeHappened = false;
    end if
  end while
end while
```

Figure 4.6: The feature selection algorithm "Forward selection and backward elimination" written in pseudo code. Variable and function names are chosen such that they describe their content or behaviour. MAXF is 30 in this work. The function *trainAndTest* trains the selected classifier with the training data and computes the classwise averaged classification rate on the training data. In the end, the vector *fSelected* contains the feature selection.

The gold standard classes are the rows of the matrix and the classification result the columns. Each matrix entry n_{ij} is the number of data samples with the respective gold standard class y_i assigned to the result class y_j. For a better view, beside the confusion matrix with the results printed in numbers of data samples always a normed confusion matrix is shown. The rows are divided by the number of data samples in the respective gold standard class $\#y_i$, and the entries therefore show the fraction of data samples with the respective gold standard class assigned to the result class. A standard measure that can be computed from the confusion matrix is the accuracy or overall recognition rate, i.e. the sum of the diagonal of the confusion matrix divided by the total number of data samples in the experiments. The accuracy is the total fraction of data samples classified correctly. However, this measure is inadequate in the present work: We do not know the prior probabilities for the classes. They are therefore always set equal in the Bayes classification experiments, or given by the data samples themself in the kNN and SVM experiments. We do not make classes equal in size, which would be a major reduction in the number of data samples in some experiments, and could lead to less robust classifier training, especially for the naïve Bayes and kNN classifier. Therefore the most prominent number used as classification result is the classwise averaged classification rate, i.e. the mean of the diagonal of the normed confusion matrix:

$$CR = \frac{1}{\#y} \sum_{i=1,\ldots,\#y} \frac{n_{ii}}{\#y_i} \tag{4.11}$$

where n_{ii} is the number of correctly classified samples of class y_i. In the remainder of the work, "classification rate" (CR) always refers to the classwise averaged classification rate, not the overall accuracy.

For 2-class classification experiments additional result measurements are provided: Sensitivity, specificity and, where appropriate, the receiver-operator curve (ROC). The sensitivity is the true positive rate, where "positive" means the presence of a disease. The disease class is always the second class y_2 in the confusion table. The sensitivity (true positive rate) is then:

$$SENS = n_{22}/\#y_2; \tag{4.12}$$

The specificity is the true negative rate:

$$SPEC = n_{11}/\#y_1; \tag{4.13}$$

The ROC curve plots $SENS$ against $(1 - SPEC)$ by altering the prior probabilities of the classes throughout the range from all data samples being classified to the negative class to all data samples being classified to the positive class. The ROC curve is easily computed on the results of the linear naïve Bayes classifier. Altering the prior probabilities of the classes is a translation of the linear decision boundary. The ROC curve can therefore be calculated by moving a threshold along the posterior probabilities of the data samples without recomputing the Bayes classifier for specific priors. The SVM classifier does not take priors into account. However, we can compute a "virtual" ROC similar to the ROC on Bayes results: The threshold is moved along the distance to the decision boundary. For the kNN classifier, no ROC

can be computed. A common measure to quantify a classification result is the area under the ROC (auROC), which would be 1.0 for an ideal classification result. The auROC is always given in conjunction with the ROC plot in a figure.

All classification experiments and the feature selection are written in Matlab (Mathworks, Inc.). The implementation used for the classifiers are the native Matlab function *classify* with the option *'diaglinear'* to perform linear naïve Bayes classification. The kNN and SVM classifier are taken from the "Statistical Pattern Recognition Toolbox for Matlab" by Michal Schlesinger, Vaclav Hlavac and Vojtech Franc, downloadable from `http://cmp.felk.cvut.cz/cmp/software/stprtool/`. Different SVM optimizers are available, the one chosen for this work is *'svmlight'* which performs best on large feature sets. The implementation had to be modified to remove endless loop bugs in the SVM optimizer implementation. The parameters of the classification experiments were always optimized for the best classwise averaged classification rate. For the SVM, this included a grid search on the C parameter with grid refinement in each cross-validation run. In the feature selection process, this grid search leads to unbearable computation times, and therefore the SVM was used with a fixed C of 1 and less allowed optimization iterations during feature selection. For the kNN classifier the number k of neighbors was optimized.

A single classification experiment with the features already loaded into the Matlab workspace, including cross-validation and feature selection takes about 2 minutes for the linear naïve Bayes classifier, 8 minutes for the kNN classifier and 50 minutes for the SVM classifier with linear kernel on a MacBook Pro, Intel Core 2 Duo, 2,66 GHz with 4GB main memory. The SVM optimization and the kNN neighbor search are performed utilizing compiled Mex-Files for computational efficiency.

4.5 Results and discussion

In the following, the classification experiments are presented. First of all, the approach how to reduce the number of experiments to a feasible count is presented in Subsection 4.5.1. In Subsections 4.5.2 to 4.5.5, the results are given and discussed. The confusion matrix is shown for most of the experiments carried out. ROC plots and the features selected are only shown and discussed if appropriate.

4.5.1 Parameter matrix

The number of differing classification experiments that can be constructed with the methods given in Sections 4.2 to 4.4 is large:

- There are many ways to formulate classification challenges out of the 4 diagnoses H, OHT, PPG and PG and their combinations. For example, the diagnoses can directly be used as classes, which would give one 4-class challenge (H vs. OHT vs. PPG vs. PG), four 3-class challenges (e.g. H vs. PPG vs. PG) and twelve 2-class challenges (e.g. H vs. PG). The classification challenges are abbreviated by "first class versus (vs.) second class vs. third class ...". Out of the possible combinations of diagnoses two are meaningful: H and OHT are combined to a "Not glaucoma" class, called Normal (N), and PPG and PG are combined to a "glaucoma" (G) class. With these two combined classes additional classification challenges can be formulated (e.g. N vs. G, N vs. PPG vs PG, H vs. G).

- The thickness normalizations yield 4 possible normalization combinations: Without normalization, age normalization, magnification normalization, and both normalizations together.

- 3 classifiers can be tested.

- Taking purely automated or manually corrected segmentation results are 2 possibilities to run a classification experiment. But training on manually corrected and testing with automated segmentations and vice versa has also to be considered, i.e. all 4 combinations of the two segmentation types are of interest.

Thus for each classification challenge $4 \times 3 \times 4 = 48$ parameter combinations can be tested, yielding a number of experiments that have a high overall computation time and, more importantly: Not all of the experiments provide substantial information content. Therefore, the experiments have been structured in a meaningful way:

1. First, the classification challenge that is of most interest is searched for in Subsection 4.5.2. All further experiments use this classification challenge. The experiments are based on manually corrected segmentations and no layer thickness normalization is applied. Only the baseline linear naïve Bayes classifier is used.

2. After the classification challenge is fixed, the influence of the layer thickness normalizations is investigated in Subsection 4.5.3. We assume that the effects of

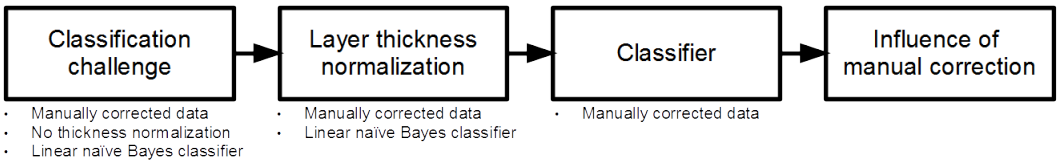


Figure 4.7: Classification evaluation structure. First, the classification challenge of most interest is searched for. The possibilities to perform a thickness normalization are looked into afterwards, followed by the comparison of the 3 classifiers. Finally, the influence of the manual correction of the segmentations is investigated. Below the step boxes, the parameters fixed by intention in the respective experiments are noted.

the normalizations, if there are any, show with manually corrected segmentation data and the Bayes classifier. The best normalization method is applied in all the remaining experiments.

3. The 3 classifiers are compared in Subsection 4.5.4 with manually corrected segmentation data.

4. Finally, the selected classifier on manually corrected segmentation data is used to study the influence of the manual corrections in Subsection 4.5.5.

The structure of the classification evaluation is shown schematically in Figure 4.7.

4.5.2 Challenge definition

Only manually corrected segmentation data and the linear naïve Bayes classifier are used to find the most interesting classification challenge. No thickness normalization is applied. The obvious classification challenge we will have to look at first is the 4-class problem H vs. OHT vs. PPG vs. PG. Each diagnosis is directly taken as a class. Results are given in Table 4.4:

GS/Res.	H	OHT	PPG	PG
H	256	122	71	1
OHT	75	48	47	7
PPG	25	28	88	27
PG	3	4	41	174

(a) Confusion matrix

GS/Res.	H	OHT	PPG	PG
H	0.57	0.27	0.16	0.00
OHT	0.42	0.27	0.27	0.04
PPG	0.15	0.17	0.52	0.16
PG	0.01	0.02	0.18	0.78

(b) Confusion matrix (normed)

Table 4.4: Classification result for the 4-class challenge H vs. OHT vs. PPG vs. PG. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.537.

The CR is 0.537. While the PG class has been detected with a classification rate of 0.78, the other classes and especially the OHT class, with a classification rate of only 0.27, fall behind. The confusion is distributed around the diagonal of the confusion matrix: There is almost no H eye classified as PG and vise versa. H eyes may be classified as OHT and PPG. OHT are most of the times classified as H, and as OHT and PPG with the same rate. The wrongly classified PPG eyes are almost evenly distributed over the other classes. PG eyes are sometimes classified as PPG, but most of the times correct. This structure of the confusion matrix resembles the fact that the classes are ordered with increasing severity of the diagnosis. As the overall CR is very low, other classification challenges are formulated.

GS/Res.	H	OHT
H	245	210
OHT	91	88

(a) Confusion matrix

GS/Res.	H	OHT
H	0.54	0.46
OHT	0.51	0.49

(b) Confusion matrix (normed)

Table 4.5: Classification result for H vs. OHT. No layer thickness normalization applied. Linear naïve classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.515. Sensitivity: 0.492. Specificity: 0.538.

The classes with the highest confusion in the 4-class problem, H and OHT, built the next classification challenge H vs. OHT. It is tested whether they are separable at all. The classification result in Table 4.5 nearly resembles a random choice, which is reasonable given the diagnosis definitions. The OHT eyes do not, except the elevated eye pressure, show signs of structural or functional damage. To enlarge class sizes, a combination of the two diagnoses into one "normal" class (N) suggests itself.

Before investigating the N class, another two challenges with the original diagnoses are looked at. The two extremes of the diagnoses are tested against each other, i.e. H vs. PG. Results are given in Table 4.6. The classifier separates the two classes with

GS/Res.	H	PG
H	443	4
PG	19	204

(a) Confusion matrix

GS/Res.	H	PG
H	0.99	0.01
PG	0.09	0.91

(b) Confusion matrix (normed)

Table 4.6: Classification result for H vs. PG. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.953. Sensitivity: 0.915. Specificity: 0.991.

a CR of 0.953. The auROC is 0.985 (ROC plot not shown). The CR and auROC numbers are in the same range as in other publications that investigated this problem [Burg 05, Huan 05, Bask 12] and as in the preceding work by the author [Maye 09]. All these publications only included glaucoma patients with visual field defects in their database. Advanced glaucoma cases can thus be easily separated from healthy eyes in an automated process. Even the simple linear naïve Bayes classifier is able to

reach a high differentiation. However, more interesting from a clinical point of view is the detection of PPG cases. As these eyes do not show visual field defects, it is more likely that the disease has been unnoticed by the subject. Dedicated screening centers that keep examination times low for both the subject and the examiner by using automated scores may help to detect the disease early. Shorter examination times may improve the chance that a subject decides to have his or her eyes checked on a regular basis. Table 4.7 shows the results for the H vs. PGG challenge. The CR drops compared to the H vs. PG challenge and is 0.73. The task of differentiating PPG eyes from healthy ones is not as simple as in the advanced PG cases. The confusion with the respective other class is roughly similar for both classes, H and PPG.

GS/Res.	**H**	**PPG**
H	338	104
PPG	50	118

(a) Confusion matrix

GS/Res.	**H**	**PPG**
H	0.76	0.24
PPG	0.30	0.70

(b) Confusion matrix (normed)

Table 4.7: Classification result for H vs. PPG. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.734. Sensitivity: 0.702. Specificity: 0.765.

GS/Res.	**N**	**PG**
N	617	18
PG	22	202

(a) Confusion matrix

GS/Res.	**N**	**PG**
N	0.97	0.03
PG	0.10	0.90

(b) Confusion matrix (normed)

Table 4.8: Classification result for N vs. PG. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.937. Sensitivity: 0.902. Specificity: 0.972.

GS/Res.	**N**	**PPG**
N	457	178
PPG	51	117

(a) Confusion matrix

GS/Res.	**N**	**PPG**
N	0.72	0.28
PPG	0.30	0.70

(b) Confusion matrix (normed)

Table 4.9: Classification result for N vs. PPG. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.708. Sensitivity: 0.696. Specificity: 0.720.

Finally, we investigate the combined N class. Therefore three classification challenges are defined: N vs. PG, N vs. PPG, and N vs. the combination of PPG and PG, i.e. all glaucomatous eyes G. The results are given in Tables 4.8, 4.9 and 4.10.

GS/Res.	N	G
N	578	49
G	95	295

(a) Confusion matrix

GS/Res.	N	G
N	0.92	0.08
G	0.24	0.76

(b) Confusion matrix (normed)

Table 4.10: Classification result for N vs. G. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.839. Sensitivity: 0.756. Specificity: 0.922.

The results of the N vs. PG and N vs. PPG challenges are, while slightly worse, very similar to the H vs. PG and H vs. PPG results. This is another indicator that the combination of H and OHT to N is valid. The N vs. G result ($CR = 0.839$) is in the middle between N vs. PPG and N vs. PG. It is better than the average of these two results, which might be due to the enlarged training data. For N vs. G, less N eyes are wrongly assigned than for N vs. PPG. Figure 4.8 shows the ROC plots of the 3 challenges with the N class. The auROC for the N vs. PG challenge is 0.975, the auROC for the N vs. PPG challenge 0.775, and the auROC for the N vs. G challenge 0.903. The N vs. G challenge is selected to be the most interesting one for two reasons. First, it can be medically reasoned: The aim is to differentiate eyes without glaucoma from eyes with glaucoma, leaving out the other factors, namely elevated eye pressure or the severity of the disease. Second, the combination of the diagnoses enlarges the data samples available for training and testing and yields more resilient results.

Before investigating the effects of the thickness normalizations, we have a closer look on which features are chosen in the N vs.G classification experiment. There were 35 different features chosen during the cross-validation runs. On average, the feature selection in each cross-validation run picked 5.10 features. Table 4.11 a) shows the features selected most often, i.e. more than once. Only the mean RNFL thickness in the inferior quadrant was used in every single cross-validation run. It was expected that RNFL features lead the feature ranking as this is the layer with highest diagnostic relevance for glaucoma, proven by all publications on glaucoma diagnosis from OCT. The other most selected features are mean values of the INL+OPL in two segments and, interestingly, blood vessel indices ratio and PCA features. The blood vessel indices ratio and PCA features have no straightforward relation to a distance or area measure on the OCT image. However, they seem to add to the discriminative value of the RNFL measure in case of the linear naïve Bayes classifier. Also the INL+OPL layer is not a common glaucoma indicator. Summing up the times a feature was used from a specific layer group yields Table 4.11 b). The table does reflect the feature ranking, with RNFL, BV and INL+OPL features chosen most often. The overall retina features do not seem to contribute much to a glaucoma discrimination when features of single layer groups are available. Table 4.11 c) sums up the times a feature with a specific type was used. The means in segments are the most relevant features. We will again have a look at the features in Section 4.5.5 in the final classification experiments. Up to now, the presence of the INL+OPL and BV features in the feature vectors for classification are a surprise.

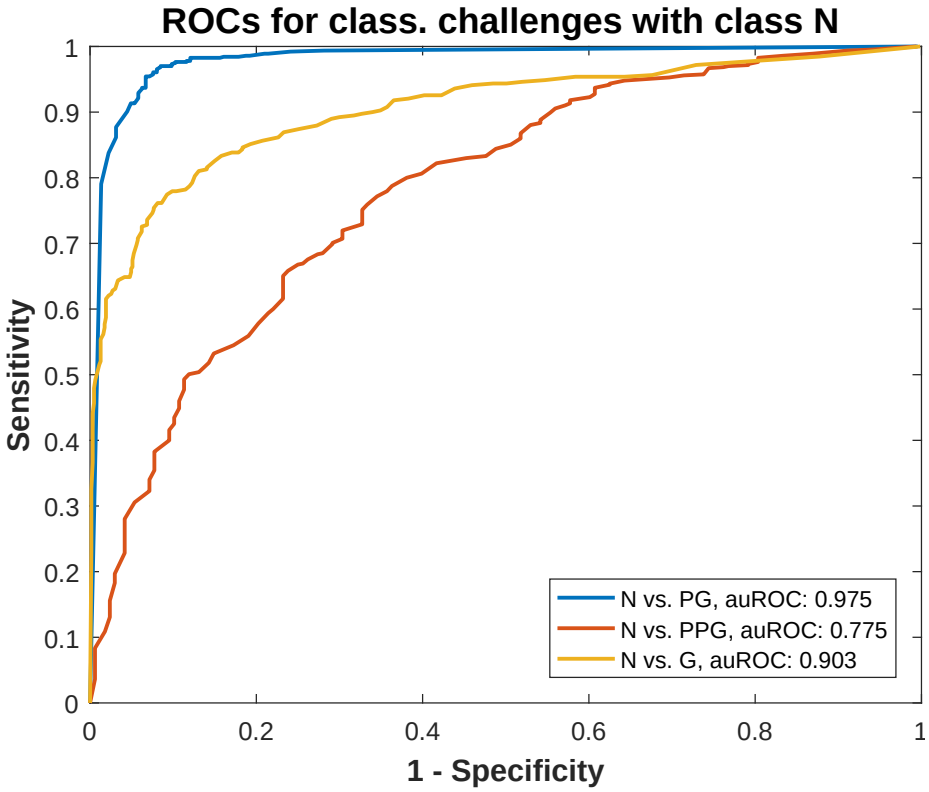


Figure 4.8: ROC plots of the classification challenges including the "normal" (N) class, i.e. the combination of H and OHT. No thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data.

4.5.3 Influence of thickness normalization

In Section 4.2, the method for an age normalization of the thickness profiles was presented. However, it was already shown that the effects of the normalization are small compared to the normal thickness distribution of the layer groups. And indeed, as Table 4.12 shows, we were not able to obtain better results. The CR decreases to 0.826. Performing the N vs. G experiments on decade age-matched data gives a similar result. Without thickness normalization, the CR is 0.822, and with age normalization 0.821, i.e. the CR stays almost the same on age-matched data (confusion matrices are not shown for the age-matched data experiments). A proper test on the effect of the age normalization would be not only using age-matched data, but data that is evenly distributed among an age range in both classes. However, this is not possible with the database of this work as the resulting data sample reduction would prevent robust classifier training. The CR results together with the analysis of the method in Section 4.2 lead to the conclusion that the age normalization has no pronounced effect on classification results, especially no positive effect.

The same is true for the magnification normalization, as can been seem from Table 4.13. The CR is 0.826, i.e. also lower than without normalization ($CR = 0.839$). If both normalization methods are combined, the CR is 0.830 (confusion matrix not shown) and does again not surpass the result without normalization. The conclu-

Feature	#Incl.
$f_{RNFL,mean,i}$	10
$f_{BV,ratio,5}$	3
$f_{INL+OPL,mean,10}$	2
$f_{INL+OPL,mean,30}$	2
$f_{BV,ratio,1}$	2
$f_{BV,ratio,29}$	2
$f_{BV,PCAhealthy,2}$	2

(a) Feature ranking

Layer	#Feat.
Retina	2
RNFL	13
GCL+IPL	5
INL+OPL	8
ONL+ELM	7
RPE grp.	4
BV	12

(b) Layer groups

Type	#Feat.
Std.	2
4 Quad.	11
32 Seg.	17
4 Ratios	1
32 Ratios	10
PCA All	3
PCA H.	7

(c) Feature types

Table 4.11: Selected features for N vs. G. No layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Total number of different features chosen during cross validation: 35. Average number of features chosen during cross validation runs: 5.10. a) Features that were chosen more than once. #Incl.: Number of cross validation runs this feature was included. b) Summarized feature inclusions from a specific layer group during the cross validation. RPE grp.: IP+OP+RPE. #Feat.: summarized #Incl. of features from this layer group during the cross validation. c) Summarized feature inclusions from a specific feature type during the cross validation. #Feat.: summarized #Incl. of features from this feature type during the cross validation.

GS/Res.	N	G
N	568	59
G	100	290

(a) Confusion matrix

GS/Res.	N	G
N	0.91	0.09
G	0.26	0.74

(b) Confusion matrix (normed)

Table 4.12: Classification result for N vs. G. Age layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.825. Sensitivity: 0.744. Specificity: 0.906.

sion is that, while the layer thickness normalization methods seem reasonable from a theoretical point of view, they do not improve results in real classification experiments. Therefore, the upcoming experiments are performed without any thickness normalization.

GS/Res.	N	G
N	573	54
G	102	288

(a) Confusion matrix

GS/Res.	N	G
N	0.91	0.09
G	0.26	0.74

(b) Confusion matrix (normed)

Table 4.13: Classification result for N vs. G. Magnification layer thickness normalization applied. Linear naïve Bayes classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.826. Sensitivity: 0.738. Specificity: 0.914.

4.5.4 Classifier selection

In the previous results sections the N vs. G classification challenge was decided to be the most interesting, and it turned out that layer thickness normalization has no positive effect. We will now compare the 3 classifiers linear naïve Bayes, kNN and linear SVM. Still manually corrected segmentation data is utilized. The classification result of the linear naïve Bayes was already shown in Table 4.10. A CR of 0.839 is achieved. The kNN results are given in Table 4.14. Using $k = 7$ neighbors yielded the best CR of 0.823, worse than the linear naïve Bayes classifier. As the kNN classifier is sensitive to an uneven class distribution in the training data, the experiment has also been carried out on age-matched data with even class sizes, yielding a even lower CR of 0.810 (confusion matrix not shown).

GS/Res.	**N**	**G**
N	580	47
G	109	281

(a) Confusion matrix

GS/Res.	**N**	**G**
N	0.93	0.07
G	0.28	0.72

(b) Confusion matrix (normed)

Table 4.14: Classification result for N vs. G. No thickness normalization applied. kNN classifier with $k = 7$. Manually corrected segmentation data. Classwise averaged classification rate: 0.823. Sensitivity: 0.721. Specificity: 0.925.

The results of the linear SVM surpass the linear naïve Bayes classifier and kNN results, as can be seen from Table 4.15. A CR of 0.859 is achieved. The number of wrongly assigned samples from both the N and G class are lower than for the linear naïve Bayes classifier. Therefore the SVM is chosen as the classifier for the final experiments in the next section. The features chosen are also discussed there.

GS/Res.	**N**	**G**
N	587	40
G	85	305

(a) Confusion matrix

GS/Res.	**N**	**G**
N	0.94	0.06
G	0.22	0.78

(b) Confusion matrix (normed)

Table 4.15: Classification result for N vs. G. No thickness normalization applied. Linear SVM classifier. Manually corrected segmentation data. Classwise averaged classification rate: 0.859. Sensitivity: 0.782. Specificity: 0.936.

4.5.5 Manually corrected and automated results

Up to now, we only used segmentation data that was manually corrected. However, as stated before, this is not common practice. Most publications use the built-in segmentation methods from the OCT manufacturer without correcting the results. Excluding scans of low quality should guarantee the quality of the segmentations. But it cannot be denied that each segmentation method fails sometimes, especially if the image content does not match the assumptions the segmentation algorithm developer made. In the extensive evaluation of the segmentation algorithm presented in this work in Chapter 3 it was shown that, in general, good segmentation results can be expected, with minor errors on the ONFL and IPL/INL boundary. But there were also major failures, with 2 scans standing out prominently from the segmentation evaluation dataset. The segmentation failure on these two scans could be explained by wrong retina positioning and the possibility of the presence of another disease besides glaucoma. The explanation of the failures does not alter the fact that in daily clinical practice such scans will be recorded and sometimes must be used, in case the patient's condition does not allow for a better scan. It will now be investigated whether the correction of all the minor and major segmentation errors influences the glaucoma classification. The automatically generated segmentations of the complete classification dataset were corrected for segmentation errors by the author. The manual corrections are a valid representation of the corrections by a single observer, as we have shown in the observer evaluation in Section 3.4. If we do not classify on this manually corrected data but on the purely automated segmentation results, the classification rate gets worse, as Table 4.16 shows. On manually corrected data, a CR of 0.859 is achieved and on the purely automated data 0.842.

GS/Res.	N	G
N	591	36
G	101	289

(a) Confusion matrix

GS/Res.	N	G
N	0.94	0.06
G	0.26	0.74

(b) Confusion matrix (normed)

Table 4.16: Classification result for N vs. G. No layer thickness normalization applied. Linear SVM classifier. Automated segmentation data. Classwise averaged classification rate: 0.842. Sensitivity: 0.741. Specificity: 0.943.

But looking only at confusion matrices and classification rates tells only part of the story. Depending on the data used, different features are selected during the training process, as Tables 4.17 and 4.18 show. The features selected when using manually corrected segmentation data are summarized in Table 4.17. Compared to the selection with the linear naïve Bayes classifier, we see the similarity that again the mean RNFL thickness in the inferior quadrant is the most often used feature. However, the BV features, the presence of which cannot be fully explained in the selected features by the Bayes classifier, have vanished. The features used most often by the linear SVM are dominated by RNFL features. Also the features of the complete retina are selected often - again, contrary to the selection of the linear naïve Bayes classifier. Features from the other layers play only a minor role, but they are

Feature	#Incl.
$f_{RNFL,mean,i}$	9
$f_{RNFL,mean,s}$	4
$f_{RNFL,PCAhealthy,9}$	2
$f_{Retina,PCAhealthy,4}$	2
$f_{INL+OPL,ratio,6}$	2

(a) Feature ranking

Layer	#Feat.
Retina	10
RNFL	25
GCL+IPL	5
INL+OPL	6
ONL+ELM	5
RPE grp.	4
BV	0

(b) Layer groups

Type	#Feat.
Std.	2
4 Quad.	16
32 Seg.	12
4 Ratios	1
32 Ratios	19
PCA All	0
PCA H.	5

(c) Feature types

Table 4.17: Selected features for N vs. G. No layer thickness normalization applied. Linear SVM classifier. Manually corrected segmentation data. Total number of different features chosen during cross validation: 41. Average number of features chosen during cross validation runs: 5.50. For a description of the tables and abbreviations see 4.11.

also selected. However, there is no layer group standing out beside the RNFL. As the GCL+ICL is now also in the focus of glaucoma research, one could expect more features selected from this layer, but this is not the case. Perhaps the scan pattern is the reason: GCL+ICL measures are usually taken from the macula region, not on the circular scan pattern around the ONH. Regarding the feature types selected most often, means or ratios in segments are prominent, sometimes the PCA features with the PCA transformation computed on healthy data. Standard measures from the complete layer thickness profiles are only selected twice in a cross-validation run.

Switching from manually corrected segmentation data to purely automated segmentations results in the feature selection summarized in Table 4.18. The classifier seems to adapt to the data that is less reliable. Still, the RNFL is the layer most features used originate from, and the distribution of features from the different layers in Table 4.18 b) is very similar to the one from 4.17 b). But the top-ranking feature is now the overall mean RNFL thickness - a feature that was only selected once when using manually corrected data. The feature type ranking of automated data in Table 4.18 c) compared to the feature type ranking on manually corrected data in Table 4.17 c) adds to the image: Overall measures, i.e. standard measures and PCA features) are selected more often when automated data containing segmentation errors is used, local measures, i.e. means and ratios in sections, less often. The feature selection adapts to less reliable data by choosing measures of more global scale.

What happens when the training of the classifier is made on one data type (purely automated segmentations or manually corrected ones), but the testing is made on the other? One could imagine the scenario that extensive care is taken when building a classification system, and all errors are removed from the segmentations the system is trained with. But in daily clinical practice, with tight time schedules, the correction of the segmentation results is omitted. Results for this scenario are given in Table 4.19. The classification system after feature selection and training is the same as for the results in Table 4.15, with the feature selection summarized in Table 4.17, i.e. the

Feature	#Incl.
$f_{RNFL,mean}$	6
$f_{RNFL,mean,i}$	3
$f_{RNFL,ratio,i}$	2
$f_{Retina,mean,25}$	2
$f_{RNFL,PCAall,1}$	2
$f_{RNFL,ratio,s}$	2
$f_{RNFL,mean,16}$	2

(a) Feature ranking

Layer	#Feat.
Retina	7
RNFL	27
GCL+IPL	3
INL+OPL	5
ONL+ELM	6
RPE grp.	2
BV	0

(b) Layer groups

Type	#Feat.
Std.	9
4 Quad.	3
32 Seg.	14
4 Ratios	6
32 Ratios	9
PCA All	6
PCA H.	3

(c) Feature types

Table 4.18: Selected features for N vs. G. No layer thickness normalization applied. Linear SVM classifier. Automated segmentation data. Total number of different features chosen during cross validation: 38. Average number of features chosen during cross validation runs: 5.00. For a description of the tables and abbreviations see 4.11.

more local features. But in the cross-validation runs, the features generated from the automated segmentations have been used for the test data samples. Interestingly, the CR does barely change and is 0.861. It does not get worse. It seems that although the segmentation errors influence the feature selection and training of the classifier, they are not critical for the single classification results.

GS/Res.	N	G
N	581	46
G	80	310

(a) Confusion matrix

GS/Res.	N	G
N	0.93	0.07
G	0.21	0.79

(b) Confusion matrix (normed)

Table 4.19: Classification result for N vs. G. No thickness normalization applied. Linear SVM classifier. Classifier training with manually corrected segmentation data and testing on automated data. Classwise averaged classification rate: 0.861. Sensitivity: 0.795. Specificity: 0.927.

When performing this experiment the other way round, i.e. using the classification system trained on purely automated data that selected more global features summarized in Table 4.18, and testing with the manually corrected data, there is a change in the results (shown in Table 4.20): The CR drops to 0.824. To have a closer look, the ROCs of the 4 relevant classification experiments of this section are plotted in Figure 4.9. The classification systems trained with manually corrected data perform best with auROC values of of 0.910 and 0.915. Training on automated data yields auROC values of 0.898 and 0.884.

While every segmentation method will most likely give slightly different results and other segmentation errors, depending on the assumptions made, it is most likely that the following statement holds in general: When training a glaucoma classification system on OCT data, manually corrected data should be used, no matter if the classification system is later used with manually corrected or purely automated data.

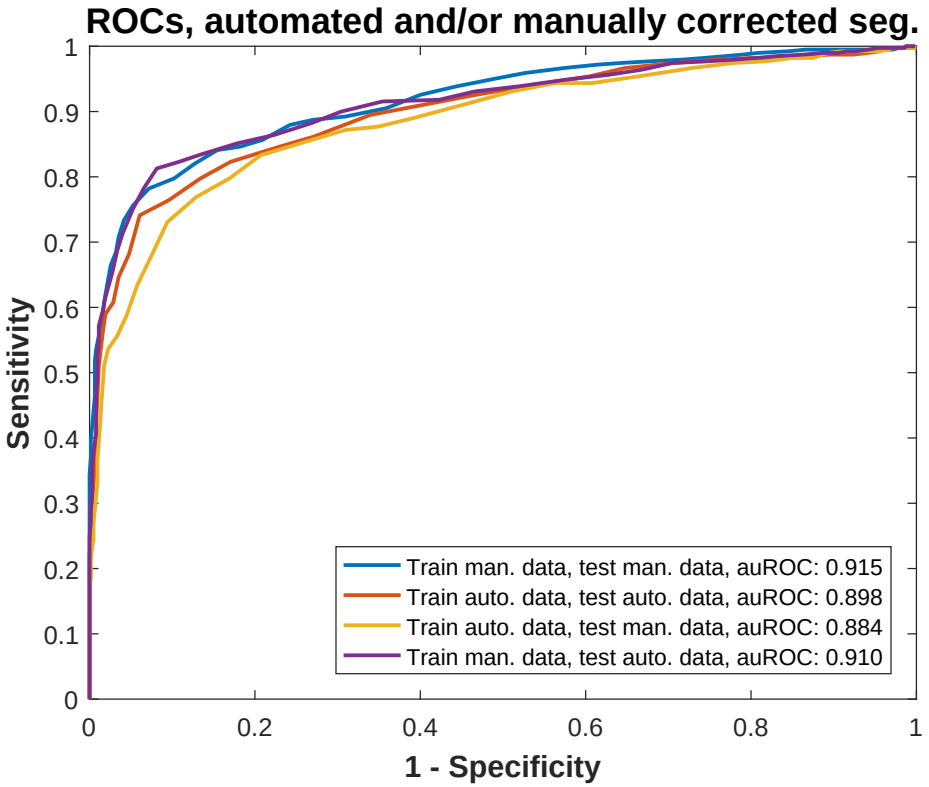


Figure 4.9: ROC plots of classification experiments on automated and manually corrected segmentation data. No thickness normalization applied. Linear SVN classifier.

GS/Res.	N	G
N	559	68
G	95	295

(a) Confusion matrix

GS/Res.	N	G
N	0.89	0.11
G	0.24	0.76

(b) Confusion matrix (normed)

Table 4.20: Classification result for N vs. G. No thickness normalization applied. Linear SVM classifier. Classifier training with automated segmentation data and testing on manually corrected data. Classwise averaged classification rate: 0.824. Sensitivity: 0.756. Specificity: 0.892.

4.6 Proposal of an OCT glaucoma probability score

The linear SVM classifier has a property we can utilize to construct a glaucoma probability score for OCT (OCT-GPS). Its decision boundary can be geometrically interpreted. It is a hyperplane through the feature space. While this hyperplane leads to binary decisions in the classifier, i.e. the feature vector is assigned to the class of the side of the hyperplane it lays, we can also derive a continuous measure: The distance of the feature vector from the decision boundary. This distance is the base for the proposed OCT-GPS.

A remark has to be made on the decision boundary and the class distribution of samples in the data set. Standard SVMs train best when the data is balanced, as the cost of misclassification in both classes is the same, no matter what the priors of the classes are. In [Fran 11] two common heuristic methods for dealing with unbalanced classes are mentioned: First, the introduction of an additional cost factor during training. Second, the tuning of the bias after an unaltered training to achieve better error rates. In this work, the training process and the data distribution are also taken as they are and no weighting is introduced due to the uneven class sizes. Using the distance to the decision boundary as a continuous measure is exactly the same heuristic as tuning the bias for differing priors.

All the classification experiments performed in Section 4.5 performed the feature selection inside the single cross-validation runs, i.e. the number of features chosen might differ from one run to the next. To make the distance to the decision boundary comparable between cross-validation runs, again a classification experiment is performed. This time a fixed feature set is used. All 41 features that were chosen at least once when using a linear SVM classifier and manually corrected segmentation data built this feature set. The result is given in Table 4.21. With the fixed feature set, the CR improved compared to the forward selection and backward elimination feature selection to 0.882.

GS/Res.	**N**	**G**
N	590	37
G	69	321

(a) Confusion matrix

GS/Res.	**N**	**G**
N	0.94	0.06
G	0.18	0.82

(b) Confusion matrix (normed)

Table 4.21: Classification result for N vs. G. No thickness normalization applied. Linear SVM classifier. Manually corrected segmentation data. No automated feature selection, but fixed feature set used. Classwise averaged classification rate: 0.882. Sensitivity: 0.823. Specificity: 0.941.

The OCT-GPS is defined by:

$$OCT-GPS(d) = \begin{cases} 0\%, d \leq D_{min} \\ 100\%, d \geq D_{max} \\ \frac{d-D_{min}}{D_{max}-D_{min}} \cdot 100\%, \text{otherwise} \end{cases} \tag{4.14}$$

where d is the distance of a feature vector to the SVM decision boundary. D_{min} is the left and D_{max} the right boundary of the distance range that yield OCT-GPS number between 0% and 100%. The distances to the decision boundary on the classification evaluation dataset range in $[-7.51; 5.31]$. D_{min} and D_{max} are set to -3 and 3, respectively. These numbers were chosen such that the resulting histogram distribution of the OCT-GPS on the classification evaluation dataset is as even as possible. The classification result can also be directly derived from the OCT-GPS. Below 50%, the data sample is assigned to the N class, above to the G class. The correlation of the OCT-GPS to the glaucoma diagnosis G on the classification evaluation dataset

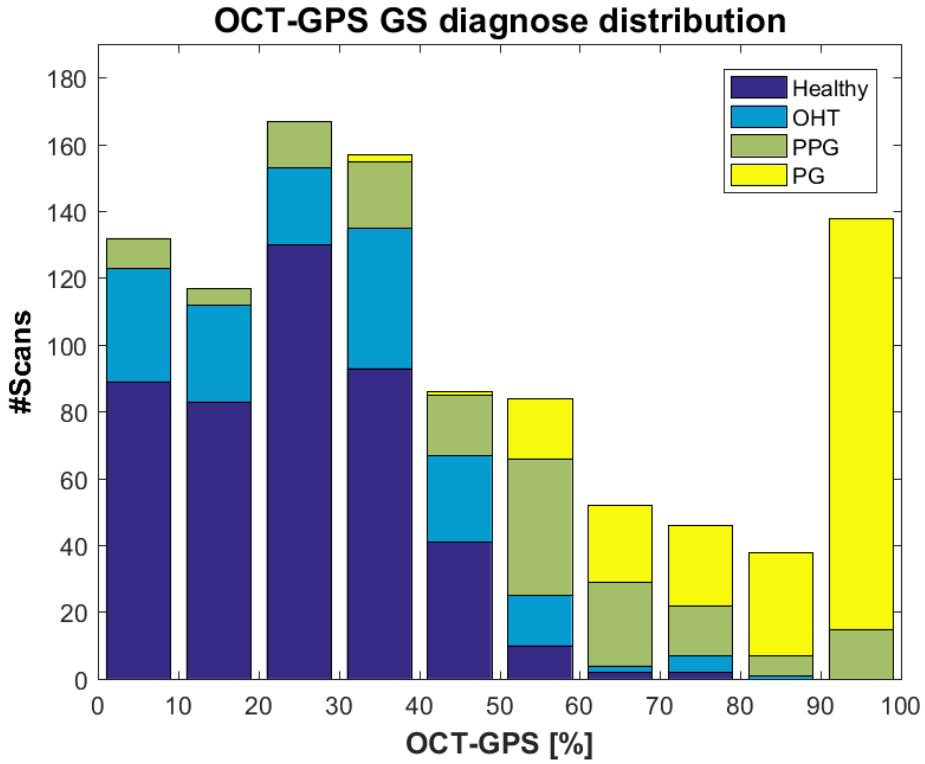


Figure 4.10: OCT-GPS distribution histogram on the classification evaluation dataset. The gold standard (GS) diagnoses falling into a OCT-GPS tenth are counted.

is 0.747, which is much higher than the best correlation a single feature yields (0.698 for $F_{RNFL,mean,i}$).

The histogram distribution is shown Figure 4.10. Besides the general distribution of the OCT-GPS, the distribution of the original GS diagnoses along the OCT-GPS can be read from the figure. The histogram bins count the GS diagnoses. The properties of this GS distribution support the idea of the OCT-GPS. Below 30% OCT-GPS in the histogram, no PG eye can be found. In general, there are only very few PG datasets misclassified and therefore below 50% OCT-GPS. Above 80% OCT-GPS, there is no healthy eye. OHT eyes are clearly positioned to the lower side of the OCT-GPS, which again validates the combination of them with H eyes to the N class. The PPG eyes are the most difficult diagnosis to detect automatically by OCT. They are the ones that are assigned to the wrong class most often, but still a majority (60.7%) is above 50% OCT-GPS. The reasonable GS diagnosis distribution among the OCT-GPS together with its straightforward and easy-to-explain construction suggest its use as a glaucoma diagnosis parameter.

4.7 Outlook

In nearly every step of the pattern recognition pipeline, there are possibilities to further enhance the glaucoma classification system presented in the future. First of all, while the database size of the classification evaluation dataset is huge compared to even recent works [Garc 12, Mwan 13, Yiu 14], the number of samples still only allows for a robust estimation of linear decision boundaries, as preliminary tests have shown. As database sizes in the clinics increase, the presented experiments should be repeated, also using more advanced classifiers.

There are already numerous publications on segmentation methods for OCT volume data (see Section 3.1 and Tables B.1 and following) that sometimes segment up to 10 layers. Parameters derived from volume segmentations may surpass the ones from circular scans around the ONH in their diagnostic capability for glaucoma [Leun 10b, Hwan 12, Seo 12]. But there is a tradeoff: For a proper data mining and for automated feature selection methods as presented in this work, database sizes must be as large as possible, as mentioned in the previous paragraph. As volume scans were only recently introduced to daily clinical practice, it will take years to build up proper data bases. Publications with automated classification systems using OCT volume data underline this: Zhang et al. [Zhan 13] used a database of 232 volume scans. Srinivasan et al. [Srin 14b] used 45 volume scans. No publication with OCT volume scans had database of around a thousand data samples available as in this work, with the exception of [Bask 12] with 794 low resolution cube scans from the Zeiss Cirrus OCT.

Using only circular scans may also be a benefit: Simple and cheap OCT systems built with off-the-shelf components can be set up in eye disease mass screening centers. The speed of cheap systems may not be on par with high-end or research OCT systems, but enough for a high quality circular scan that is sufficient to automatically detect glaucoma suspects, e.g. with the OCT-GPS score presented in this work.

Finally, the features in this work were restricted to features derived from retinal layer segmentations. Features taken directly from image data could also be tested for their capability to detect glaucoma. Especially concepts like the flat space introduced by Carass et al. [Cara 14] may yield features that do not rely on a segmentation algorithm.

To conclude this chapter, the results can be briefly summarized as follows: Among the possible classification challenges for the diagnoses in the database, trying to separate non-glaucomatous (healthy and ocular hypertension) from glaucomatous eyes was of most interest. Normalizations of the layer thickness profiles did not have a positive effect on the classification results. The linear SVM classifier was the classifier with the best results. When using purely automated segmentation data, the feature selection process tends to choose features of more global scale compared scenarios where manually corrected data is used. Performing the feature selection and classifier training on manually corrected data gives better classification rates, also in the case that it is tested on purely automated segmentation results without corrections. A glaucoma probability score derived from the distance of the feature vector to the SVM decision boundary was proposed. This score exhibits a favorable distribution of the gold standard diagnoses and may thus be used as a glaucoma parameter.

Chapter 5

Summary

Glaucoma is a chronic disease which may cause blindness. The structural damage done by glaucoma is irreversible. It is only possible to slow down its progression. Therefore the early detection of the disease is of significant importance for the patient. There exists no single measurement on which the diagnosis can be based on. Instead, the ophthalmologist relies on multiple modalities, for example visual field tests, fundus photography, the Heidelberg retina tomograph (HRT) and optical coherence tomography (OCT).

OCT is a modality that utilizes the properties of short coherent light to generate depth profiles of tissue. Since its invention [Huan 91], its main application was ophthalmology, as OCT allows to visualize the retinal structure in vivo [Ferc 93]. Modern systems can acquire 2D scans consisting of multiple depth profiles in a fraction of a second, making eye motion artifacts negligible. While 3D imaging has been introduced into clinical practice in recent years, the most common scan pattern is still a circular scan around the optic nerve head (ONH) with a standardized diameter of 3.46mm. The retinal nerve fiber layer (RNFL) thickness measured on this circular scan is an important glaucoma indicator [Gued 03].

As the number of glaucoma patients is expected to increase in the future [Quig 06], supporting the diagnosis by automizing parts of the routine is necessary to keep examination times low. For example, the HRT has a built-in glaucoma probability score (GPS) [Swin 00] that is calculated by machine learning methods and breaks the imaged topography of the retina down to a single number. In this work, a GPS for OCT (OCT-GPS) is proposed. It is based on the development, implementation, and evaluation of a complete pattern recognition pipeline, which for the first time allowed to break a restriction of the built-in methods from OCT manufacturers, i.e. to use complete layer segmentation results and not only summarized parameters. First, the retinal layers are segmented on a circular OCT scan around the ONH to generate thickness profiles of 6 retinal layer groups and to find the blood vessel positions. The thickness profiles are optionally normalized and then multiple features are generated from them, among others novel principal component analysis features. A feature selection process, i.e. a data mining method, chooses relevant features. These are assigned to a diagnosis class by a classifier. In the following, the properties of the data collection this work is based on, the methods and evaluation results for the segmentation algorithm, and the classification system are summarized.

Circular scans around the ONH were acquired with a Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany) from subjects included in the "Erlangen Glaucoma Registry" (www.clinicaltrials.gov, NCT00494923). They reflect data from daily clinical practice. Scans were only excluded when a human observer is not able to differentiate retinal layers on the scan, i.e. if the retina is not completely in the scan area or severe averaging artifacts are present. Contrary to other publications in the field [Bask 12, Garc 12] scans were explicitly not excluded when they had low image quality or when other diseases beside glaucoma are present. For 1024 scans age information and a diagnosis were available. These form the classification data set. The diagnosis was carried out by medical experts based on an extensive ophthalmic examination. There are 453 healthy (H) eyes, 179 eyes with ocular hypertension (OHT), 168 preperimetric glaucoma (PPG) eyes, and 224 perimetric glaucoma (PG) eyes from 575 subjects in total. From this classification dataset a segmentation evaluation dataset was created by random selection that fulfilled these properties: 30 scans from each diagnosis group, half of them from a left and half from a right eye. Only one eye from a subject. The built-in quality index from the Spectralis (HE quality) is available. With the quality index, the scans from the segmentation evaluation dataset can be separated into 60 scans of low and 60 scans of high quality. The separation into low and high HE quality correlates with the glaucoma (PPG and PG) diagnosis with a correlation coefficient of -0.03 and is therefore nearly independent of the diagnosis. The segmentation evaluation dataset is comparable in size to datasets used in other publications on retinal layer segmentation. The full classification dataset is the largest dataset used up to now for a classification task on OCT data.

The presented segmentation algorithm is an extension of [Maye 10]. The goal of the algorithm development was that it is applicable on healthy and glaucomatous eyes without a parameter change, even in the presence of a RNFL hole. Bad quality of the scan should not influence the segmentation results. The algorithm is built around a few general assumptions, e.g. the most reflecting layers are the RNFL and retinal pigment epithelium (RPE), the shape of the RPE is not disrupted and the inner layer boundaries of the retina are to a large extent parallel to the RPE boundary. 6 layer boundaries are segmented and the blood vessel positions are found. The layer boundaries are segmented step-by-step. Each processing step contains a proper pre-processing, i.e. denoising, and post-processing, i.e. smoothing of the segmented layer boundary. The inner limiting membrane (ILM), the topmost layer boundary on an OCT scan of the retina and the RPE are easily found: First, the image is heavily blurred. The minimum inside two maximum peaks in each the A-Scan, corresponding to the RNFL and RPE, splits the retina into an inner segment (IS) and an outer segment (OS). The ILM is the greatest contrast rise (seen from the top of the image) in the IS and the RPE the greatest contrast drop in the OS. Blood vessel (BV) positions are detected by an adaptive thresholding along the sum of the pixels just above the RPE: An average value is computed in a window around an A-Scan position. If the local value at the window center lies below a fixed percentage of the window average, the A-Scan is marked as a BV. On the further segmented boundaries, the BV positions are always invalidated in the smoothing step and interpolated over. For the segmentation of the inner layer boundaries an average filter on the original image is a sufficient denoiser. The inner layer boundaries are found by edge detection taking

the derivative along the A-Scan into account. Denoising by median filtering or simple averaging does not give good results to detect the outer nerve fiber layer boundary (ONFL). Therefore complex diffusion as proposed in [Fern 05] is used. An initially distorted segmentation is found by edge detection and some heuristics, i.e. if there is not exactly one edge found between the inner plexiform (IPL)/inner nuclear layer (INL) boundary, the initial ONFL segmentation is set to the ILM. A discrete energy minimization followed by a final smoothing determines the ONFL segmentation result. The energy term consists of a gradient measure, a neighborhood smoothness measure and a smoothness measure between blood vessels.

The automated segmentation results on the segmentation evaluation dataset were manually corrected with the OCTSEG software by 5 observers with experience in the field of ophthalmic imaging. The author corrected the automated results of the complete classification dataset. A gold standard (GS) was constructed from the observers' corrections: If at least two observers had corrected the respective layer at an A-Scan position, the GS is the layer position nearer to the automated segmentation in case of exactly 2 observer corrections and the median observer position otherwise. It turned out that the observers differ most at blood vessel positions, i.e. the observer standard deviation along the A-Scans and the blood vessel density computed over the segmentation evaluation dataset linearly correlated for almost all layer boundaries. The author's corrections can be seen as representative for the corrections a single OCT operator might carry out, as its mean absolute difference to the observers is within the mean inter-observer difference $\pm$ standard deviation of the inter-observer difference range, except for one layer boundary. The comparison of the GS with the automated results yielded that the ONFL and IPL/INL boundaries exhibit the highest segmentation error, with a mean absolute difference to the GS of $2.84\mu m$ and $2.56\mu m$. The algorithm development goals are fulfilled, because no significant correlation between the segmentation error and scans of bad quality or glaucomatous eyes could be found. The automated segmentation differs not much worse to the GS than the single observer's manual corrections.

The segmentation algorithm was already modified for volume data and first preliminary results presented [Maye 11]. However, in the author's opinion, the most promising segmentation methods for future volume segmentation algorithm development are graph-cut methods, as they are algorithmically compact and allow for fast computation times. The properties of glaucomatous eyes, e.g. the possibility of a complete RNFL loss, have to be taken into account when designing the algorithm. Retinal layer segmentation algorithms may incorporate a content classification stage, e.g. the presence of a disease is detected before the actual segmentation and the algorithm parameters or even the algorithm itself is adapted to the disease.

The retinal layer segmentations on the circular scans are the basis for a classification system to discriminate between glaucoma patients and healthy subjects. The classification process includes the following steps: Thickness profile normalization, feature computation, feature selection and classification. In all the experiments carried out, a 10-fold cross-validation was used. The cross-validation is patient-sensitive, e.g. both eyes of a person are either in the test or training dataset. The thickness profile normalization can use two methods: First an age normalization can be performed, i.e. we try to remove the effects of declining retinal layers thickness with

age from the data. Second, the magnification normalization transforms the thickness measures to area measures by multiplying the layer thickness values with the pixel spacing in R-direction. After the thickness normalization, the features are computed out of the 7 layer thickness profiles, i.e. 5 layer groups, the complete retina and a virtual thickness profile of blood vessel indices. The features are standard measures over the complete profile, means in sections, ratios in sections and PCA features, yielding 762 features for each B-Scan in total. To the author's knowledge, this is the first work that utilizes the full retinal layer segmentation data for OCT glaucoma detection and also applies a data mining method, i.e. feature selection, to identify the features with highest relevance. The feature selection process takes place on the training data in every cross-validation run, with the same classifier as used for the test. "Forward selection and backward elimination" is the feature selection method chosen. The linear naïve Bayes, k-nearest neighbor (kNN) and linear support vector machine (SVM) are the classifiers compared.

The first experiments are carried out with the linear naïve Bayes classifier on manually corrected segmentation data without any thickness normalization. While the classification challenge to separate H from PG eyes yielded the highest classwise averaged classification rate (CR) of 0.953 and an area under the receiver-operator curve (auROC) of 0.985, discriminating the normal (N, combined H and OHT) from the glaucomatous (G, PPG and PG) eyes is the most interesting classification challenge. The CR for the N vs. G challenge is 0.839. The thickness normalization methods did not improve this result, therefore they were omitted in the following experiments. The SVM classifier topped the result of the linear naïve Bayes with an CR of 0.859. Not using manually corrected segmentation data but purely automated results let the CR drop to 0.842. More interesting than the pure CR number is the feature selection: The classification system adapted to the segmentations containing errors by choosing features of more global scale. Training with manually corrected and testing with pure automated data and vice versa showed that it is of advance to use manually corrected data for training, no matter what the type of test data is.

A glaucoma probability score for OCT (OCT-GPS) can be constructed with the decision boundary of the SVM: The classification experiment was again carried out with a fixed feature set, containing all the features selected at least once during the cross-validation runs by the feature selection with manually corrected data. The resulting CR is 0.882. The OCT-GPS is defined as a mapping of the distance from the feature vector to the decision boundary in feature space to a percent number. It has a correlation with the glaucoma diagnosis on the classification dataset of 0.747, which is higher than any single feature. Furthermore, the original H, OHT, PPG and PG diagnoses are distributed among the OCT-GPS in a meaningful way.

In the future, the classification system may be further enhanced by utilizing volume data and image features. However, a focus on circular scans may also be beneficial: Large data collections for classifier training are more likely to be established. The circular scan around the ONH may be sufficient for a cheap, fast, and easy-to-use OCT system that can automatically detect glaucoma suspects in dedicated screening centers by using the OCT-GPS.

Appendix A

Abbreviations and symbols

Abbreviation	Explanation
ACG	Angle closure glaucoma
auROC	Area under the ROC
BV	Blood vessels
ELM	Extermal limiting membrane
EV	Eigen vector
FD-OCT	Frequency domain OCT
G	Glaucoma diagnosis group, i.e. PPG and PG combined
GPS	Glaucoma probability score
GS	Gold standard
H	Healthy group
HE	Heidelberg engineering
HE quality	Built-in B-Scan quality measure of the HE Spectralis
HRT	Heidelberg retina tomograph
ILM	Inner limiting membrane
INL	Inner nuclear layer
IPL	Inner plexiform layer
IPR	Inner photoreceptors
kNN	k-nearest neighbor classifier
LDA	Linear discriminant analysis
MoG	Mixture of Gaussians
MS	Multiple sclerosis
N	Normal diagnosis group, i.e. H and OHT combined
OAG	Open angle glaucoma
OCT	Optical coherence tomography
OD	Right eye
OHT	Ocular hypertension group
ONH	Optic nerve head
ONL	Outer nuclear layer
OPL	Outer plexiform layer
OPR	Outer photoreceptors
ONFL	Outer retinal nerve fiber layer boundary
OS	Left eye
PCA	Principal component analysis
PCV	Polypoidal choroidal vasculopathy
PG	Perimetric glaucoma group
PPG	Preperimetric glaucoma group
ROC	Receiver-operator curve
RGC	Retinal ganglion cells
RNFL	Retinal nerve fiber layer
RPE	Retinal pigment epithelium
SLO	Scanning laser ophthalmoscope
SVM	Support vector machine
TD-OCT	Time domain OCT
VF	Visual field (test)
VH	Vitreous humor
Zero quality	B-Scan quality measure as proposed in [Maye 10]

Table A.1: Table of abbreviations in the text in alphabetical order.

Symbol	Explanation
BVR	Region in between two BV
CF	Corrected fraction of A-Scans
CR	Classwise averaged classification rate
$D(r)$	Regional smoothness term at A-Scan r
DB	Image database
DOG	Mean absolute observer difference to GS
$E(r)$	Energy function
f	Single feature, the subindices detail the type
$\mathbf{f}$	Feature vector
$G(z, r)$	Gradient at position z, r
I	Image intensities
$I(z, r)$	Image intensity at position z, r
IOD	Mean absolute inter-observer difference
$L(r)$	Layer boundary position at A-Scan r
$LT(t)$	Layer thickness at A-Scan r
$\bar{LT}(a)$	Mean layer thickness for age a
$N(r)$	Local smoothness term at A-Scan r
O	Set of observers
ODA	Mean absolute difference of the observer correction to the automated segmentation
$ONFL(r)$	Outer nerve fiber layer position at A-Scan r
P	Significance of Pearson's correlation coefficient
R	Possible A-Scan positions
S_i	Segment i of 32 of the circular scan
$Scale_R$	Pixel Spacing in R-direction in $\mu m/pixel$
$Scale_Z$	Pixel Spacing in Z-direction in $\mu m/pixel$
$SDOG$	Mean signed observer difference to GS
SE	(Absolute) segmentation error
$SENS$	Sensitivity
$SPEC$	Specificity
SSE	Signed segmentation error
$std(...)$	Standard deviation
$STDO(r)$	Standard deviation of the observer corrections at A-Scan r
y	Class label

Table A.2: Table of abbreviations and symbols used in equations in alphabetical order (first part).

Symbol	Explanation
$\chi(...)$	Indicator function. 1 if the expression inside the brackes is true, 0 otherwise
α	Weighting factor
β	Weighting factor
σ	Standard deviation
σ_{CD}	Noise standard deviation estimate for complex diffusion
$\#DB$	Number of images in the database
$\#f$	Dimension of the feature vector $\mathbf{f}$
$\#N$	Number of pixels in a scan
$\#O$	Number of observers
$\#R$	Number of A-Scans in an image
$\#S_i$	Number of A-Scans in segment S_i
$\#TD$	Number of samples in the training data

Table A.3: Table of abbreviations and symbols used in equations in alphabetical order (second part).

Appendix B

Published research overview

Author	Objective	Method	Data	Evaluation
Koozekanani et al. [Kooz 01]	Retina seg.	Edge detection. Regularization by a Markov model.	1450 TD-OCT B-Scans scans from normal eyes.	Quantitative evaluation with man. corrected seg.
Ishikawa et al. [Ishi 02]	RNFL seg.	Edge detection with integritiy check.	TD-OCT circular B-Scans: 86 scans from 21 NS, 131 scans from 32 OHP, 184 scans from 45 GP.	Quantitative evaluation by marking errors.
Fernandez et al. [Fern 05]	7 layers seg.	Complex diffusion and coherence enhanced diffusion followed by edge detection.	TD-OCT B-Scans: 72 scans from NS, scans of 4 different pathologic cases.	Visual inspection.
Ishikava et al. [Ishi 05]	5 layers seg.	Edge detection with integritiy check.	TD-OCT circular B-Scans: 144 scans from 24 NS, 144 from 24 GS included.	Quantitative evaluation by marking errors. Exclusion of bad quality images.
Mujat et al. [Muja 05]	RNFL seg.	Anisotropic noise suppression and deformable splines.	SD-OCT volumes of NS.	Visual inspection.
Shahidi et al. [Shah 05]	3 layer groups seg.	Averaging A-Scans and egde detection.	TD-OCT B-Scans of 10 NS.	Reproducibility.
Haecker et al. [Haek 06]	ILM, RPE seg.	3D geometric graph cut and a priori contraints.	TD-OCT radial scan sets: 9 scan sets from NS, 9 from PP.	Qualitative evaluation by marking errors.
Baroni et al. [Baro 07]	2 layer groups seg.	Maximization of a likelihood function consisting of a gradient and local smoothness term.	TD-OCT B-Scans: Scans of 18 NS, scans of 16 CCMP.	Parameter adaption and error judging by 2 reviewers.
Fuller et al. [Full 07]	Multiple or single layer seg.	SVM classifier training for each volume out of manually drawn regions. Semi-automated method.	SD-OCT volumes of NS and patients.	Segmentation time evaluation. Comparison to man. seg.

Table B.1: Overview (first part) over published research in the field of retinal layer segmentation on OCT data. Taken over from [Maye 10] and complemented with recent works. Abbreviations see caption of Table B.5.

Author	Objective	Method	Data	Evaluation
Joeres et al. [Joer 07]	Retina, OPL and subretinal tissue seg.	Man. seg. with OCTOR software.	TD-OCT B-Scans of 60 AMD patients.	Repeatablity and agreement of two operators.
Sadda et al. [Sadd 07]	Retina seg.	Man. seg. with OCTOR software.	TD-OCT B-Scans of patients with macular diseases.	Repeatablity and agreement of two operators.
Somfai et al. [Somf 07]	Effect of operator error on seg.	Analysis with custom [Fern 05] and commercial software.	TD-OCT B-Scans of 8 NS and 1 DME patient. 4 scans with different operator errors per person.	Comparison of optimal automatic seg. with seg. on images with worse quality.
Szulmowski et al. [Szul 07]	Group of posteriour layers seg.	Classifier training out of manually drawn regions. Semi-automated method.	SD-OCT volume data of NS and patients.	Visual inspection.
Garvin et al. [Garv 08]	5 layers seg.	3D geometric graph cut and a priori contraints.	TD-OCT radial scan sets from 12 ONP. 1 diseased eye and 1 normal from each patient.	Qualitative evaluation using man. seg. of 3 observers.
Götzinger et al. [Gotz 08]	RPE seg.	Two algorithms with different complexity.	SD-PS-OCT volumes of NS and patients.	Visual inspection.
Tolliver et al. [Toll 08]	RNFL, RPE seg.	Boundary detection by spectral rounding.	SD-OCT volumes of 2 NS and 9 patients.	Quantitative evaluation with man. seg.
Tan et al. [Tan 08]	5 layers seg.	Progressive edge detection, each step less A-Scan averaging.	TD-OCT B-Scans of 44 NS, 73 PGP and 29 PPGP.	Exclusion of scans with seg. errors in the study.
Fabritius et al. [Fabr 09]	Retina seg.	Maximum intensity search with iterative refinement by outlier removal.	3 OCT volumes: Healthy, AMD, PCV.	A-Scan error rate.
Mishra et al. [Mish 09]	10 intraretinal layer seg.	Approximation and refinement of layer positions with dyn. programming.	SD-OCT B-Scans of healthy and diseased rat retinas.	Visual inspection.

Table B.2: Overview (second part) over published research in the field of retinal layer segmentation on OCT data. Abbreviations see caption of Table B.5.

Author	Objective	Method	Data	Evaluation
Tan et al. [Tan 09]	2 layer groups seg.	Edge detection with 3D neighbor constraints and knowledge model.	SD-OCT volume scans of 65 NS, 78 PGP and 52 PPGG.	Exclusion of scans with seg. errors in the study.
Yazdanpanah et al. [Yazd 09]	5 layers seg.	Active contours: Minimization of an energy functional with a shape prior. Man. initialization.	20 SD-OCT B-Scans of rat eyes.	Quantitative evaluation with man. seg.
Chiu et al. [Chiu 10]	7 layers seg.	Graph theory and dynamic programming.	Volume OCT Scans of 10 NS.	Quantitative evaluation with man. seg. of 2 observers.
Kajić et al. [Kaji 10]	9 layers seg.	Model based segmentation with shape and texture features,	SD-OCT volumes of 17 normal eyes.	Quantitative evaluation with man. seg. of 2 observers.
Lu et al. [Lu 10]	5 layers seg.	BV detection, denoising, egde detection.	4 NS scanned with circular scan pattern 4 times with one month intervalls	Comparison to manual RNFL segmentation.
Quellec et al. [Quel 10]	10 layers seg., abnormality detection	Seg. see [Garv 08], texture and thickness features for abnormality detection.	SD-OCT volumes of 13 NS.	Quantitative evaluation with man. seg. of 2 observers.
Vermeer et al. [Verm 10]	5 layers seg.	Pixelwise classification with SVM, level set regularization.	SD-OCT volumes of 10 NS and 8 GP	Quantitative evaluation with man. seg.
Yang et al. [Yang 10]	8 layers seg.	Dual scale gradient information as graph weights with dynamic programming.	19 GP volumes, 19 control volumes.	Quantitative evaluation with man. seg. of 4 observers.

Table B.3: Overview (third part) over published research in the field of retinal layer segmentation on OCT data. From 2010 on all publications used SD-OCT data, therefore this is not explicitely mentioned in the data column anymore. Abbreviations see caption of Table B.5.

Author	Objective	Method	Data	Evaluation
Yazdanpanah et al. [Yazd 11]	5 layers seg.	Active contour approach with manual initialization.	80 B-Scans from glaucoma model and control rat eyes, synthetic images.	Quantitative evaluation with man. seg. Comparison to other algorithms.
Golzan et al. [Golz 11]	RNFL, RPE seg.	BV detection to improve level set approach.	Circular scans of 20 NS and 8 GP.	Comparison of two automated methods without man. reference.
Yang et al. [Yang 11]	8 layers seg., focus on outer layers.	Adaption of [Yang 10] to RP patients.	Volume scans of 7 RP patients.	Quantitative evaluation with man. seg.
Ghorbel et al. [Ghor 11]	8 layers seg.	Active contour approach with initialization derived from k-means clustering.	700 B-Scans of 100 NS.	Visual inspection and quantitative evaluation with man. seg. of 5 observers.
Antony et al. [Anto 13]	6 layers seg.	First pixelwise classification, graph based refinement.	10 volumes from GP and volumes of mice and basset hounds.	Cross validation. Quantitative evaluation with man. seg. of 2 observers.
Dufour et al. [Dufo 13]	5 layers seg.	Minimal cut graph based method with shape constraints.	28 volumes for shape model, 30/20 volumes of NS/AMD subjects for evaluation.	Quantitative evaluation with man. seg. of 2 observers.
Kafieh et al. [Kafi 13]	10 layers seg.	Diffusion maps based on intensities and textural similarities.	23 2D and 13 3D data sets from NS and GP.	Quantitative evaluation with man. seg. of 2 observers.
Lang et al. et al. [Lang 13]	8 layers seg.	Two step approach: Boundary classification, then graph based refinement.	35 volumes of NS and MS subjects.	Quantitative evaluation with man. seg.
Carass et al. [Cara 14]	8 layers seg.	Builds on [Lang 13]. Boundary classification, then level set refinement.	37 volumes of NS and MS subjects	Quantitative evaluation with man. seg.

Table B.4: Overview (fourth part) over published research in the field of retinal layer segmentation on OCT data. Abbreviations see caption of Table B.5.

Author	Objective	Method	Data	Evaluation
Ehnes et al. [Ehne 14]	11 layer seg.	Dynamic programming approach based on gradients.	114 volumes from 3 OCT devices from 21 NS	Quantitative evaluation with man. seg. of 3 observers.
Niu et al. [Niu 14]	5 layer seg.	Edge detection with refinement.	25 volumes from NS and AMD eyes.	Quantitative evaluation with man. seg. of 2 observers.
Rathke et al. [Rath 14]	6 layer seg.	Probabilistic graphical model with global shape regularization.	80 NS circular B-Scans, 66 GP B-Scans, 35 volumes from NS.	Quantitative evaluation with man. seg. of 2 observers.
Srinivasan et al. [Srin 14a]	6 to 9 layers seg.	Extension of [Chiu 10] with global and local missing layer detection.	2000 B-Scans from mice.	Quantitative evaluation with man. seg. of 2 observers.
Chiu et al. [Chiu 15]	7 layers seg. in presence of DME	Fluid an layer classification as additional weight in dynamic programming approach.	6 DME volumes for training, 10 for validation.	Quantitative evaluation with man. seg. of 2 observers.
Gonzalez-Lopez et al. [Gonz 15]	3 layer segmentation (ILM, RPE/C, M/E, I/RPE)	Active contour approach with initialization from thresholding.	40 volumes from NS and DR subjects.	Quantitative evaluation with man. seg.
Kaba et al. [Kaba 15]	Retina and RNFL seg.	Graph cut approach.	120 scans of various patients.	Quantitative evaluation with man. seg.

Table B.5: Overview (fifth part) over published research in the field of retinal layer segmentation on OCT data. Abbreviations: age related macula degenration (AMD), blood vessels (BV), diabetic macula edema (DME), diabetic retinopathy (DR), dynamic (dyn.), glaucoma patient (GP), , manual (man.), multiple sclerosis (MS), normal subject (NS), papilledema patient (PP), ocular hypertension patient (OHP), outer photoreceptor layer (OPL), optic neuropathy patient (ONP), polypoidal choroidal vasculopathy (PCV), perimetric glaucoma patient (PGP), preperimetric glaucoma patient (PPGP), retinitis pigmentosa (RP), segmentation (seg.). The table does not claim to be complete.

Author	Objective	Data	Features	Classifier
Burgansky et al. [Burg 05]	Glaucoma detection	47 OCT scans of glaucomatous eyes, 42 OCT scans of healthy subjects, good quality.	OCT printout parameters (RNFL thickness, ONH, macula) from Zeiss Stratus OCT.	5 classifiers compared (SVM, RC, RT, GLM, GAM)
Huang et al. [Huan 05]	Glaucoma detection	89 OCT scans of glaucomatous eyes, 100 OCT scans of healthy subjects, good quality.	OCT printout parameters (RNFL thickness, ONH) from Zeiss Stratus OCT.	3 classifiers compared (LDA, MD, ANN)
Qi et al. [Qi 10]	Dysplasia in Barret's esophagus	96 endoscopic OCT images.	Intensity, gradient and texture features.	4 classifiers compared (LDA, kNN, ANN, LVQ, CT)
Liu et al. [Liu 11]	Macular pathologies (ME, MH, AMD)	81 normal, 74 MD, 203 ME, 74 AMD OCT scans. 136 subjects.	Multi-scale texture and shape features from aligned macula scans.	SVM classifier with RBF kernel
Baskaran et al. [Bask 12]	Glaucoma detection	286 scans of glaucomatous eyes/subjects, 508 scans of normals subjects/eyes, good quality.	RNFL and ONH parameters from Zeiss Cirrus OCT.	LDA and CART
Garcia-Martin et al. [Garc 12]	MS detection	OCT scans of 115 MS patients, 115 healthy subjects, good quality scans.	RNFL thickness measurements from HE Spectralis.	LDF
Mwanza et al. [Mwan 13]	Early glaucoma detection	OCT scans (one eye each) of 149 healty subjects and 104 patients with early glaucoma.	Macula and circumpapillary parameters from Zeiss Cirrus OCT, including RNFL and GC/IPL thickness.	Exploratory factor analysis with LR

Table B.6: Overview (first part) over published research in the field of automated disease detection from OCT data. Only the research where more than one parameter or feature and a classifier from the pattern recognition research field is utilized for disease detection is considered. Abbreviations see caption of Table B.7

Author	**Objective**	**Data**	**Features**	**Classifier**
Zhang et al. [Zhan 13]	Derive virtual VF from OCT	84 OCT scans (one eye each) of 84 glaucoma patients and 128 scans of normal subjects.	Segmentation of [Yang 10], without manual correction, gross segmentation errors excluded, RNFL and GC/IPL thickness features.	Multiple linear regression
Srinivasan et al. [Srin 14b]	ME and AMD detection	OCT volumes of 15 normal subjects, 15 AMD and 15 ME patients.	Multiscale gradient features.	SVM
Yiu [Yiu 14]	Glaucoma detection	OCT scans of 133 normal and 239 glaucomatous eyes.	Clock hour RNFL thicknesses from Zeiss Stratus OCT.	ANN
Belgith et al. [Belg 15]	Detect glaucoma progression	117 volume scans. 27 volumes of 27 progressive POA patients, 40 volumes of 23 healthy subjects, 50 volume scans of 26 stable POA patients. Good quality.	MRF based change detection, change map used as features.	SVDD

Table B.7: Overview (second part) over published research in the field of automated disease detection from OCT data. Only the research where more than one parameter or feature and a classifier from the pattern recognition research field is utilized for disease detection is considered. Abbreviations: age-related macula degeneration (AMD), artificial neural network (ANN), classification and regression tree (CART), classification tree (CT), generalized additive model (GAM), ganglion cell and inner plexiform layer (GC/IPL), generalized linear model (GLM), Heidelberg Engineering (HE), k-neareset neighbor classifier (kNN), learning vector quantization (LVQ), linear discriminant analysis (LDA), linear discriminant function (LDF), logistic regression (LR), Mahalanobis distance (MD), macular edema (ME), macular hole (MH), Markov random field (MRF), multiple sclerosis (MS), primary open angle glaucoma (POA), recursive partitioning (RC), retinal nerve fiber layer (RNFL), regression tree (RT), support vector data description classifier (SCDD), support vector machine (SVM), visual field (VF). Classifiers are named as by the respective authors, e.g. LDA and LDF appear as separate classifiers in the table despite being the same. The table does not claim to be complete.

List of Figures

List of Tables

Bibliography

[Alen 08] L. M. Alencar, C. Bowd, R. N. Weinreb, L. M. Zangwill, P. A. Sample, , and F. A. Medeiros. "Comparison of HRT-3 Glaucoma Probability Score and Subjective Stereophotograph Assessment for Prediction of Progression in Glaucoma". *Investigative Ophthalmology & Visual Science*, Vol. 49, No. 5, pp. 1898–1906, May 2008.

[Alon 13] D. Alonso-Caneiro, S. A. Read, and M. J. Collins. "Automatic segmentation of choroidal thickness in optical coherence tomography". *Biomedical Optics Express*, Vol. 4, No. 12, pp. 2795–2812, December 2013.

[Anto 13] B. J. Antony, M. D. Abràmoff, M. M. Harper, W. Jeong, E. H. Sohn, Y. H. Kwon, R. Kardon, and M. K. Garvin. "A combined machine-learning and graph-based framework for the segmentation of retinal surfaces in SD-OCT volumes". *Biomedical Optics Express*, Vol. 4, No. 12, pp. 2712–2728, December 2013.

[Asao 14] R. Asaoka, A. Iwase, T. Tsutsumi, H. Saito, S. Otani, K. Miyata, H. Murata, C. Mayama, and M. Araie. "Combining Multiple HRT Parameters Using the "Random Forests" Method Improves the Diagnostic Accuracy of Glaucoma in Emmetropic and Highly Myopic Eyes". *Investigative Ophthalmology & Visual Science*, Vol. 55, No. 4, pp. 2482–2490, April 2014.

[Aydi 03] A. Aydin, G. Wollstein, L. L. Price, J. G. Fujimoto, and J. S. Schuman. "Optical coherence tomography assessment of retinal nerve fiber layer thickness changes after glaucoma surgery". *Ophthalmology*, Vol. 110, No. 8, pp. 1506–1511, August 2003.

[Bale 09] D. Baleanu, R. P. Tornow, F. K. Horn, R. Laemmer, C. W. Roessler, A. G. Juenemann, F. E. Kruse, and C. Y. Mardin. "Retinal Nerve Fiber Layer Thickness in Normals Measured by Spectral Domain OCT". *Journal of Glaucoma*, Vol. 19, No. 7, pp. 475–482, December 2009.

[Balk 13] L. Balk, M. Mayer, B. M. Uitdehaag, and A. Petzold. "Physiological variation of segmented OCT retinal layer thicknesses is short-lasting". *Journal of Neurology*, Vol. 260, No. 12, pp. 3109–3114, December 2013.

[Balk 14] L. Balk, M. Mayer, B. M. Uitdehaag, and A. Petzold. "Retinal hyperaemia-related blood vessel artifacts are relevant to automated OCT layer segmentation". *Journal of Neurology*, Vol. 261, No. 3, pp. 511–517, March 2014.

[Baro 07] M. Baroni, P. Fortunato, and A. L. Torre. "Towards quantitative analysis of retinal features in optical coherence tomography". *Medical Engineering & Physics*, Vol. 29, No. 4, pp. 432–441, May 2007.

[Bask 12] M. Baskaran, E.-L. Ong, J.-L. Li, C. Y. Cheung, D. Chen, S. A. Perera, C. L. Ho, Y.-F. Zheng, and T. Aung. "Classification Algorithms Enhance the Discrimination of Glaucoma from Normal Eyes Using High-Definition Optical Coherence TomographyClassification Algorithms and Glaucoma Diagnosis". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 4, pp. 2314–2320, April 2012.

[Begu 14] V. U. Begum, U. K. Addepalli, R. K. Yadav, K. Shankar, S. Senthil, C. S. Garudadri, and H. L. Rao. "Ganglion Cell-Inner Plexiform Layer Thickness of High Definition Optical Coherence Tomography in Perimetric and Preperimetric Glaucoma". *Investigative Ophthalmology & Visual Science*, Vol. 55, No. 8, p. 4768, July 2014.

[Belg 15] A. Belghith, C. Bowd, F. A. Medeiros, M. Balasubramanian, R. N. Weinreb, and L. M. Zangwill. "Learning from healthy and stable eyes: A new approach for detection of glaucomatous progression". *Artificial Intelligence in Medicine*, Vol. 64, No. 2, pp. 105–115, June 2015.

[Bend 10] D. Bendschneider, R. P. Tornow, F. K. Horn, R. Laemmer, C. W. Roessler, A. G. Juenemann, F. E. Kruse, and C. Y. Mardin. "Retinal Nerve Fiber Layer Thickness in Normals Measured by Spectral Domain OCT". *Journal of Glaucoma*, Vol. 19, No. 7, pp. 475–482, September 2010.

[Bert 09] F. Bertuzzi, D. C. Hoffman, A. M. D. Fonseka, C. Souza, and J. Caprioli. "Concordance of Retinal Nerve Fiber Layer Defects between Fellow Eyes of Glaucoma Patients Measured by Optical Coherence Tomography". *American Journal of Ophthalmology*, Vol. 148, No. 1, pp. 148 – 154, July 2009.

[Bock 10] R. Bock, J. Meier, L. G. Nyúl, J. Hornegger, and G. Michelson. "Glaucoma Risk Index: Automated glaucoma detection from color fundus images". *Medical Image Analysis*, Vol. 14, No. 3, pp. 471–481, June 2010.

[Bowd 01] C. Bowd, L. M. Zangwill, C. C. Berry, E. Z. Blumenthal, C. Vasile, C. Sanchez-Galeana, C. F. Bosworth, P. A. Sample, and R. N. Weinreb. "Detecting Early Glaucoma by Assessment of Retinal Nerve Fiber Layer Thickness and Visual Function". *Investigative Ophthalmology & Visual Science*, Vol. 42, No. 9, pp. 1993–2003, August 2001.

[Bowd 12] C. Bowd, I. Lee, M. H. Goldbaum, M. Balasubramanian, F. A. Medeiros, L. M. Zangwill, C. A. Girkin, J. M. Liebmann, and R. N. Weinreb. "Predicting Glaucomatous Progression in Glaucoma Suspect Eyes Using Relevance Vector Machine Classifiers for Combined Structural and Functional Measurements". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 4, p. 2382, April 2012.

[Boyk 00] Y. Boykov and M.-P. Jolly. "Interactive organ segmentation using graph cuts". In: *Proceedings of the 3rd international conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 276–286, October 2000.

[Boyk 01] Y. Boykov and M.-P. Jolly. "Interactive graph cuts for optimal boundary amp; region segmentation of objects in N-D images". In: *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, pp. 105 –112, July 2001.

[Boyk 06] Y. Boykov and G. Funka-Lea. "Graph Cuts and Efficient N-D Image Segmentation". *International Journal of Computer Vision*, Vol. 70, No. 2, pp. 109–131, April 2006.

[Burg 05] Z. Burgansky-Eliash, G. Wollstein, T. Chu, J. D. Ramsey, C. Glymour, R. J. Noecker, H. Ishikawa, and J. S. Schuman. "Optical Coherence Tomography Machine Learning Classifiers for Glaucoma Detection: A Preliminary Study". *Investigative Ophthalmology & Visual Science*, Vol. 46, No. 11, pp. 4147–4152, November 2005.

[Cara 14] A. Carass, A. Lang, M. Hauser, P. A. Calabresi, H. S. Ying, and J. L. Prince. "Multiple-object geometric deformable model for segmentation of macular OCT". *Biomedical Optics Express*, Vol. 5, No. 4, pp. 1062–1074, April 2014.

[Chan 99] T. Chan and P. Mulet. "On the convergence of the lagged diffusivity fixed point method in total variation image restoration". *SIAM Journal on Numerical Analysis*, Vol. 36, No. 2, pp. 354–367, July 1999.

[Chen 15] Q. Chen, W. Fan, S. Niu, J. Shi, H. Shen, and S. Yuan. "Automated choroid segmentation based on gradual intensity distance in HD-OCT images". *Optics Express*, Vol. 23, No. 7, pp. 8974–8994, April 2015.

[Chiu 10] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu. "Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation". *Optics Express*, Vol. 18, No. 18, pp. 19413–19428, August 2010.

[Chiu 15] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu. "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema". *Biomedical Optics Express*, Vol. 6, No. 4, pp. 1172–1194, April 2015.

[Dua 12] S. Dua, U. R. Acharya, P. Chowriappa, and S. V. Sree. "Wavelet-Based Energy Features for Glaucomatous Image Classification". *IEEE Transactions on Information Technology in Biomedicine*, Vol. 16, No. 1, pp. 80–87, January 2012.

[Duda 00] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[Dufo 13] P. Dufour, L. Ceklic, H. Abdillahi, S. Schroder, S. De Dzanet, U. Wolf-Schnurrbusch, and J. Kowal. "Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints.". *IEEE Transactions on Medical Imaging*, Vol. 32, No. 3, pp. 531–543, October 2013.

[Ehne 14] A. Ehnes, Y. Wenner, C. Friedburg, M. N. Preising, W. Bowl, W. Sekundo, E. M. z. Bexten, K. Stieger, and B. Lorenz. "Optical Coherence Tomography (OCT) Device Independent Intraretinal Layer Segmentation". *Translational Vision Science & Technology*, Vol. 3, No. 1, p. 1, February 2014.

[Fabr 09] T. Fabritius, S. Makita, M. Miura, R. Myllylä, and Y. Yasuno. "Automated segmentation of the macula by optical coherence tomography". *Optics Express*, Vol. 17, No. 18, pp. 15659–15669, August 2009.

[Fayy 96] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth. "From Data Mining to Knowledge Discovery in Databases". *AI Magazine*, Vol. 17, pp. 37–54, September 1996.

[Feno 13] J.-R. Fenolland, C. Boucher, M. Mayer, W. Rostene, C. Baudouin, and A. Denoyer. "Developments in Optical Coherence Tomography Imaging in a Rat Model of Glaucoma". In: *Investigative Ophthalmology & Visual Science, ARVO Anual Meeting 2013*, p. 4833, June 2013.

[Ferc 03] A. Fercher, W. Drexler, C. Hitzenberger, and T. Lasser. "Optical coherence tomography - principles and applications". *Reports on Progress in Physics*, Vol. 66, No. 2, pp. 239–303, January 2003.

[Ferc 10] A. F. Fercher. "Optical coherence tomography - development, principles, applications.". *Zeitschrift für medizinische Physik*, Vol. 20, No. 4, pp. 251–276, November 2010.

[Ferc 93] A. F. Fercher, C. K. Hitzenberger, W. Drexler, G. Kamp, and H. Sattmann. "In vivo optical coherence tomography". *American Journal of Ophthalmology*, Vol. 116, No. 1, pp. 113–114, July 1993.

[Ferc 95] A. Fercher, C. Hitzenberger, G. Kamp, and S. El-Zaiat. "Measurement of intraocular distances by backscattering spectral interferometry". *Optics Communications*, Vol. 117, No. 1-2, pp. 43 – 48, May 1995.

[Fern 05] D. C. Fernández, H. M. Salinas, and C. A. Puliafito. "Automated detection of retinal layer structures on optical coherence tomography images". *Optics Express*, Vol. 13, No. 25, pp. 10200–10216, December 2005.

[Fran 11] V. Franc, A. Zien, and B. Sch´olkopf. "Support Vector Machines as Probabilistic Models". In: *Proceedings of the 28th International Conference on Machine Learning*, pp. 665–672, June 2011.

[Fuji 03] J. G. Fujimoto and M. E. Brezinski. *Biomedical Photonics Handbook, Optical Coherence Tomography Imaging*, Chap. 13. CRC Press, 2003.

[Full 07] A. R. Fuller, R. J. Zawadzki, S. Choi, D. F. Wiley, J. S. Werner, and B. Hamann. "Segmentation of Three-dimensional Retinal Image Data". *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, No. 6, pp. 1719–1726, November 2007.

[Garc 12] E. Garcia-Martin, L. E. Pablo, R. Herrero, M. Satue, V. Polo, J. M. Larrosa, J. Martin, and J. Fernandez. "Diagnostic ability of a linear discriminant function for spectral-domain optical coherence tomography in patients with multiple sclerosis.". *Ophthalmology*, Vol. 119, No. 8, pp. 1705–1711, August 2012.

[Garv 08] M. K. Garvin, M. D. Abramoff, R. Kardon, S. R. Russell, X. Wu, and M. Sonka. "Intraretinal Layer Segmentation of Macular Optical Coherence Tomography Images Using Optimal 3-D Graph Search". *IEEE Transactions on Medical Imaging*, Vol. 27, No. 10, pp. 1495 – 1505, October 2008.

[Garw 98] D. Garway-Heath, A. Rudnicka, T. Lowe, P. Foster, F. Fitzke, and R. Hitchings. "Measurement of optic disc size: equivalence of methods to correct for ocular magnification". *British Journal of Ophthalmology*, Vol. 82, No. 6, pp. 643–649, June 1998.

[Ghor 11] I. Ghorbel, F. Rossant, I. Bloch, S. Tick, and M. Paques. "Automated segmentation of macular layers in OCT images and quantitative evaluation of performances". *Pattern Recognition*, Vol. 44, No. 8, pp. 1590–1603, August 2011.

[Gilb 04] G. Gilboa, N. A. Sochen, and Y. Y. Zeevi. "Image Enhancement and Denoising by Complex Diffusion Processes". *IEEE Transaction on Pattern Analysis & Machine Intelligence*, Vol. 26, No. 8, pp. 1020–1036, August 2004.

[Gold 02] M. H. Goldbaum, P. A. Sample, K. Chan, J. Williams, T.-W. Lee, E. Blumenthal, C. A. Girkin, L. M. Zangwill, C. Bowd, T. Sejnowski, and R. N. Weinreb. "Comparing Machine Learning Classifiers for Diagnosing Glaucoma from Standard Automated Perimetry". *Investigative Ophthalmology & Visual Science*, Vol. 43, No. 1, p. 162, January 2002.

[Gold 05] M. H. Goldbaum, P. A. Sample, Z. Zhang, K. Chan, J. Hao, T.-W. Lee, C. Boden, C. Bowd, R. Bourne, L. Zangwill, T. Sejnowski, D. Spinak, and R. N. Weinreb. "Using Unsupervised Learning with Independent Component Analysis to Identify Patterns of Glaucomatous Visual Field Defects". *Investigative Ophthalmology & Visual Science*, Vol. 46, No. 10, pp. 3676–3683, October 2005.

[Gold 12] M. H. Goldbaum, I. Lee, G. Jang, M. Balasubramanian, P. A. Sample, R. N. Weinreb, J. M. Liebmann, C. A. Girkin, D. R. Anderson, L. M. Zangwill, M.-J. Fredette, T.-P. Jung, F. A. Medeiros, and C. Bowd. "Progression of Patterns (POP): A Machine Classifier Algorithm to Identify Glaucoma Progression in Visual Fields". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 10, pp. 6557–6567, September 2012.

[Golz 11] S. M. Golzan, A. Avolio, and S. L. Graham. "Minimising Retinal Vessel Artefacts in Optical Coherence Tomography Images". *Computer Methods and Programs in Biomedicine*, Vol. 104, No. 2, pp. 206–211, 2011.

[Gonz 15] A. Gonzalez-Lopez, M. Ortega, M. G. Penedo, and P. Charlon. "A web-based framework for anatomical assessment of the retina using OCT". *Biosystems Engineering*, Vol. 138, pp. 44–58, October 2015.

[Gotz 08] E. Götzinger, M. Pircher, W. Geitzenauer, C. Ahlers, B. Baumann, S. Michels, U. Schmidt-Erfurth, and C. K. Hitzenberger. "Retinal pigment epithelium segmentation bypolarization sensitive optical coherence-tomography". *Optics Express*, Vol. 16, No. 21, pp. 16410–16422, October 2008.

[Grzy 10] N. M. Grzywacz, J. de Juan, C. Ferrone, D. Giannini, D. Huang, G. Koch, V. Russo, O. Tan, and C. Bruni. "Statistics of Optical Coherence Tomography Data From Human Retina". *IEEE Transactions on Medical Imaging*, Vol. 29, No. 6, pp. 1224 –1237, June 2010.

[Gued 03] V. Guedes, J. S. Schuman, E. Hertzmark, G. Wollstein, A. Correnti, R. Mancini, D. Lederer, S. Voskanian, L. Velazquez, H. M. Pakter, T. Pedut-Kloizman, J. G. Fujimoto, and C. Mattox. "Optical coherence tomography measurement of macular and nerve fiber layer thickness in normal and glaucomatous human eyes". *Ophthalmology*, Vol. 110, No. 1, pp. 177–189, January 2003.

[Guyo 03] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182, March 2003.

[Haek 06] M. Haeker, M. Abramoff, R. Kardon, and M. Sonka. "Segmentation of the Surfaces of the Retinal Layer from OCT Images". In: *Proceedings of the 9th international conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 800–807, Oktober 2006.

[Hast 03] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Corrected Edition)*. Springer, August 2003.

[Haus 98] G. Hausler and M. W. Lindner. ""Coherence Radar" and "Spectral Radar"—New Tools for Dermatological Diagnosis". *Journal of Biomedical Optics*, Vol. 3, No. 1, pp. 21–31, January 1998.

[Heij 02] A. Heijl, C. M. Leske, B. Bengtsson, L. Hyman, B. Bengtsson, and M. Hussein. "Reduction of Intraocular Pressure and Glaucoma Progression". *Archives of Ophthalmology*, Vol. 120, No. 10, pp. 1268–1279, October 2002.

[Hood 07] D. C. Hood and R. H. Kardon. "A framework for comparing structural and functional measures of glaucomatous damage". *Progress in Retinal and Eye Research*, Vol. 26, No. 6, pp. 688–710, November 2007.

[Horn 09] F. K. Horn, C. Y. Mardin, R. Laemmer, D. Baleanu, A. Juenemann, F. E. Kruse, and R. P. Tornow. "Correlation between Local Glaucomatous Visual Field Defects and Loss of Nerve Fiber Layer Thickness measured with Scanning Laser Polarimetry and Spectral Domain Optical Coherence Tomograpy". *Investigative Ophthalmology & Visual Science*, Vol. 50, No. 5, pp. 1971–1977, January 2009.

[Horn 11] F. K. Horn, C. Y. Mardin, D. Bendschneider, A. G. Juenemann, W. Adler, and R. P. Tornow. "Frequency doubling technique perimetry and spectral domain optical coherence tomography in patients with early glaucoma". *Eye*, Vol. 25, No. 1, pp. 17–29, January 2011.

[Huan 05] M.-L. Huang and H.-Y. Chen. "Development and Comparison of Automated Classifiers for Glaucoma Diagnosis Using Stratus Optical Coherence Tomography". *Investigative Ophthalmology & Visual Science*, Vol. 46, No. 11, pp. 4121–4129, November 2005.

[Huan 12] D. Huang, V. Chopra, A. T.-H. Lu, O. Tan, B. Francis, R. Varma, and A. I. for Glaucoma Study (AIGS) Group. "Does Optic Nerve Head Size Variation Affect Circumpapillary Retinal Nerve Fiber Layer Thickness Measurement by Optical Coherence Tomography?". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 8, pp. 4990–4997, July 2012.

[Huan 91] D. Huang, E. A. Swanson, C. P. Lin, J. Schuman, W. Stinson, W. Chang, M. Hee, T. Flotte, K. Gregory, C. Puliafito, *et al.* "Optical coherence tomography". *Science*, Vol. 254, No. 5035, pp. 1178–1181, November 1991.

[Hwan 12] Y. H. Hwang and Y. Y. Kim. "Glaucoma Diagnostic Ability of Quadrant and Clock-Hour Neuroretinal Rim Assessment Using Cirrus HD Optical Coherence TomographyOCT Rim Assessment in Glaucoma". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 4, pp. 2226–2234, April 2012.

[Ishi 02] H. Ishikawa, S. Piette, J. M. Liebmann, and R. Ritch. "Detecting the inner and outer borders of the retinal nerve fiber layer using optical coherence tomography". *Graefe's Archive for Clinical and Experimental Ophthalmology*, Vol. 240, No. 5, pp. 362–371, May 2002.

[Ishi 05] H. Ishikawa, D. M. Stein, G. Wollstein, S. Beaton, J. G. Fujimoto, and J. S. Schuman. "Macular Segmentation with Optical Coherence Tomography". *Investigative Ophthalmology & Visual Science*, Vol. 46, No. 6, pp. 2012–2017, June 2005.

[Izat 94] J. A. Izatt, M. R. Hee, E. A. Swanson, C. P. Lin, D. Huang, J. S. Schuman, C. A. Puliafito, and J. G. Fujimoto. "Micrometer-scale resolution imaging of the anterior eye in vivo with optical coherence tomography". *Archives of Ophthalmology*, Vol. 112, No. 12, pp. 1584–1589, December 1994.

[Joer 07] S. Joeres, J. W. Tsong, P. G. Updike, A. T. Collins, L. Dustin, A. C. Walsh, P. W. Romano, , and S. R. Sadda. "Reproducibility of Quantitative Optical Coherence Tomography Subanalysis in Neovascular Age-Related Macular Degeneration". *Investigative Ophthalmology & Visual Science*, Vol. 48, No. 9, pp. 4300–4307, September 2007.

[Kaba 15] D. Kaba, Y. Wang, C. Wang, X. Liu, H. Zhu, A. G. Salazar-Gonzalez, and Y. Li. "Retina layer segmentation using kernel graph cuts and continuous max-flow". *Opt. Express*, Vol. 23, No. 6, pp. 7366–7384, March 2015.

[Kafi 13] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka. "Intra-retinal layer segmentation of 3D optical coherence tomography using coarse grained diffusion map". *Medical Image Analysis*, Vol. 17, No. 8, pp. 907–928, December 2013.

[Kaji 10] V. Kajić, B. Považay, B. Hermann, B. Hofer, D. Marshall, P. L. Rosin, and W. Drexler. "Robust segmentation of intraretinal layers in the normal human fovea using a novel statistical model based on texture and shape analysis". *Optics Express*, Vol. 18, No. 14, pp. 14730–14744, July 2010.

[Kaji 12] V. Kajić, M. Esmaeelpour, B. Považay, D. Marshall, P. L. Rosin, and W. Drexler. "Automated choroidal segmentation of 1060 nm OCT in healthy and pathologic eyes using a statistical model". *Biomedical Optics Express*, Vol. 3, No. 1, pp. 86–103, January 2012.

[Kale 07] M. Kalev-Landoy, A. C. Day, M. F. Cordeiro, and C. Migdal. "Optical coherence tomography in anterior segment imaging". *Acta Ophthalmologica Scandinavica*, Vol. 85, No. 4, pp. 427–430, June 2007.

[Kana 13] A. Kanamori, M. Naka, A. Akashi, M. Fujihara, Y. Yamada, and M. Nakamura. "Cluster Analyses of Grid-Pattern Display in Macular Parameters Using Optical Coherence Tomography for Glaucoma DiagnosisOptical Coherence Tomography in Glaucoma Diagnosis". *Investigative Ophthalmology & Visual Science*, Vol. 54, No. 9, p. 6401, September 2013.

[Kiri 14] M. Y. Kirillin, G. Farhat, E. A. Sergeeva, M. C. Kolios, and A. Vitkin. "Speckle statistics in OCT images: Monte Carlo simulations and experimental studies". *Optics Letters*, Vol. 39, No. 12, pp. 3472–3475, June 2014.

[Kola 13] R. Kolar, R. P. Tornow, R. Laemmer, J. Odstrcilik, M. A. Mayer, J. Gazarek, J. Jan, T. Kubena, and P. Cernosek. "Analysis of Visual Appearance of Retinal Nerve Fibers in High Resolution Fundus Images: A Study on Normal Subjects". In: *Computational and Mathematical Methods in Medicine*, Hindawi Publishing Corporation, December 2013.

[Kooz 01] D. Koozekanani, K. Boyer, and C. Roberts. "Retinal thickness measurements from optical coherence tomography using a Markov boundary model". *IEEE Transactions on Medical Imaging*, Vol. 20, No. 9, pp. 900–916, September 2001.

[Koto 14] J. Kotowski, G. Wollstein, H. Ishikawa, and J. S. Schuman. "Imaging of the optic nerve and retinal nerve fiber layer: An essential part of glaucoma diagnosis and monitoring". *Survey of Ophthalmology*, Vol. 59, No. 4, pp. 458–467, July 2014.

[Laem 11] R. Laemmer, R. P. Tornow, M. A. Mayer, F. K. Horn, F. E. Kruse, and C. Y. Mardin. "Enhanced Depth Imaging Optical Coherence Tomography of the Choroid-Influence of Age and Glaucomatous Damage". In: *Investigative Ophthalmology & Visual Science, ARVO Anual Meeting 2011*, p. 3500, April 2011.

[Lang 13] A. Lang, A. Carass, M. Hauser, E. S. Sotirchos, P. A. Calabresi, H. S. Ying, and J. L. Prince. "Retinal layer segmentation of macular OCT images using boundary classification". *Biomedical Optics Express*, Vol. 4, No. 7, pp. 1133–1152, July 2013.

[Lee 05] D. A. Lee and E. J. Higginbotham. "Glaucoma and its treatment: A review". *American Journal of Health-System Pharmacy*, Vol. 62, No. 7, pp. 691–699, April 2005.

[Lee 11] P. Lee, W. Gao, and X. Zhang. "Speckle properties of the logarithmically transformed signal in optical coherence tomography". *Journal of the Optical Society of America A*, Vol. 28, No. 4, pp. 517–522, April 2011.

[Lesk 03] C. M. Leske, A. Heijl, M. Hussein, B. Bengtsson, L. Hyman, and E. Komaroff. "Factors for glaucoma progression and the effect of treatment: The early manifest glaucoma trial". *Archives of Ophthalmology*, Vol. 121, No. 1, pp. 48–56, January 2003.

[Leun 05] C. K. Leung, W.-M. Chan, C. Y. Ko, S. I. Chui, J. Woo, M.-K. Tsang, and R. K. K. Tse. "Visualization of Anterior Chamber Angle Dynamics Using Optical Coherence Tomography". *Ophthalmology*, Vol. 112, No. 6, pp. 980–984, June 2005.

[Leun 09] C. K. Leung, C. Y. Cheung, R. N. Weinreb, Q. Qiu, S. Liu, H. Li, G. Xu, N. Fan, L. Huang, C. P. Pang, and D. S. C. Lam. "Retinal Nerve Fiber Layer Imaging with Spectral-Domain Optical Coherence Tomography: A Variability and Diagnostic Performance Study". *Ophthalmology*, July 2009.

[Leun 10a] C. K. Leung, N. Choi, R. N. Weinreb, S. Liu, C. Ye, L. Liu, G. W. Lai, J. Lau, and D. S. Lam. "Retinal Nerve Fiber Layer Imaging with Spectral-Domain Optical Coherence Tomography: Pattern of RNFL Defects in Glaucoma". *Ophthalmology*, August 2010.

[Leun 10b] C. K. Leung, S. Lam, R. N. Weinreb, S. Liu, C. Ye, L. Liu, J. He, G. W. Lai, T. Li, and D. S. Lam. "Retinal Nerve Fiber Layer Imaging with Spectral-Domain Optical Coherence Tomography: Analysis of the Retinal Nerve Fiber Layer Map for Glaucoma Detection". *Ophthalmology*, Vol. 117, No. 9, September 2010.

[Leun 12] C. K. Leung, M. Yu, R. N. Weinreb, C. Ye, S. Liu, G. Lai, and D. S. Lam. "Retinal Nerve Fiber Layer Imaging with Spectral-Domain Optical Coherence Tomography: A Prospective Analysis of Age-Related Loss". *Ophthalmology*, Vol. 119, No. 4, pp. 731–737, April 2012.

[Liu 11] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, , and J. M. Rehg. "Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding.". *Medical Image Analysis*, Vol. 15, No. 5, pp. 748–759, June 2011.

[Lu 10] S. Lu, J. Liu, J. H. Lim, C. Cheung, and T. Y. Wong. "Automated layer segmentation of optical coherence tomography images". In: *Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on*, pp. 2035–2038, June 2010.

[Masl 15] J. S. Maslin, K. Mansouri, and S. K. Dorairaj. "HRT for the Diagnosis and Detection of Glaucoma Progression". In: *The Open Ophthalmology Journal*, pp. 58–67, May 2015.

[Maya 13] C. Mayama, H. Saito, H. Hirasawa, S. Konno, A. Tomidokoro, M. Araie, A. Iwase, S. Ohkubo, K. Sugiyama, T. Otani, S. Kishi, K. Matsushita, N. Maeda, M. Hangai, and N. Yoshimura. "Circle- and Grid-Wise Analyses of Peripapillary Nerve Fiber Layers by Spectral Domain Optical Coherence Tomography in Early-Stage GlaucomaAnalyses of NFLs by SD-OCT in Early-Stage Glaucoma". *Investigative Ophthalmology & Visual Science*, Vol. 54, No. 7, p. 4519, July 2013.

[Maye 09] M. A. Mayer, J. Hornegger, C. Y. Mardin, F. E. Kruse, and R. P. Tornow. "Automated Glaucoma Classification Using Nerve Fiber Layer Segmentations On Circular Spectral Domain OCT B-Scans". In: *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting 2009*, p. 1011, April 2009.

[Maye 10] M. A. Mayer, J. Hornegger, C. Y. Mardin, and R. P. Tornow. "Retinal Nerve Fiber Layer Segmentation on FD-OCT Scans of Normal Subjects and Glaucoma Patients". *Biomedical Optics Express*, Vol. 1, No. 5, pp. 1358–1383, December 2010.

[Maye 11] M. A. Mayer, J. Hornegger, C. Y. Mardin, and R. P. Tornow. "Retinal Layer Segmentation on OCT-Volume Scans of Normal and Glaucomatous Eyes". In: *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, p. 3669, April 2011.

[Mish 09] A. Mishra, A. Wong, K. Bizheva, and D. A. Clausi. "Intra-retinal layer segmentation in optical coherence tomography images". *Optics Express*, Vol. 17, No. 26, pp. 23719–23728, December 2009.

[Mook 12] M. R. K. Mookiah, U. Rajendra Acharya, C. M. Lim, A. Petznick, and J. S. Suri. "Data Mining Technique for Automated Diagnosis of Glaucoma Using Higher Order Spectra and Wavelet Energy Features". *Knowledge-Based Systems*, Vol. 33, pp. 73–82, September 2012.

[Muja 05] M. Mujat, R. Chan, B. Cense, B. Park, C. Joo, T. Akkin, T. Chen, and J. de Boer. "Retinal nerve fiber layer thickness map determined from optical coherence tomography images". *Optics Express*, Vol. 13, No. 23, pp. 9480–9491, November 2005.

[Mwan 13] J.-C. Mwanza, J. L. Warren, and D. L. Budenz. "Combining Spectral Domain Optical Coherence Tomography Structural Parameters for the Diagnosis of Glaucoma With Early Visual Field Loss". *Investigative Ophthalmology & Visual Science*, Vol. 54, No. 13, pp. 8393–8400, December 2013.

[Mwan 14] J.-C. Mwanza, D. L. Budenz, D. G. Godfrey, A. Neelakantan, F. E. Sayyad, R. T. Chang, and R. K. Lee. "Diagnostic Performance of Optical Coherence Tomography Ganglion Cell–Inner Plexiform Layer Thickness Measurements in Early Glaucoma". *Ophthalmology*, Vol. 121, No. 4, pp. 849–854, April 2014.

[Na 11] J. H. Na, K. R. Sung, S. Baek, J. H. Sun, and Y. Lee. "Macular and Retinal Nerve Fiber Layer Thickness: Which Is More Helpful in the Diagnosis of Glaucoma?". *Investigative Ophthalmology & Visual Science*, Vol. 52, No. 11, pp. 8094–8101, October 2011.

[Na 12] J. H. Na, K. R. Sung, S. Baek, Y. J. Kim, M. K. Durbin, H. J. Lee, H. K. Kim, and Y. H. Sohn. "Detection of Glaucoma Progression by Assessment of Segmented Macular Thickness Data Obtained Using Spectral Domain Optical Coherence Tomography". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 7, p. 3817, June 2012.

[Niem 03] H. Niemann. *Klassifikation von Mustern, 2. Auflage*. Internet Publication, 2003.

[Niu 14] S. Niu, Q. Chen, L. de Sisternes, D. L. Rubin, W. Zhang, and Q. Liu. "Automated retinal layers segmentation in SD-OCT images using dual-gradient and spatial correlation smoothness constraint". *Computers in Biology and Medicine*, Vol. 54, pp. 116–128, November 2014.

[Oddo 11] F. Oddone, M. Centofanti, L. Tanga, M. Parravano, M. Michelessi, M. Schiavone, C. M. Villani, P. Fogagnolo, and G. Manni. "Influence of disc size on optic nerve head versus retinal nerve fiber layer assessment for diagnosing glaucoma.". *Ophthalmology*, Vol. 118, No. 7, pp. 1340–1347, July 2011.

[Odst 13] J. Odstrcilik, R. Kolar, A. Budai, J. Hornegger, J. Jan, J. Gazarek, T. Kubena, P. Cernosek, O. Svoboda, and E. Angelopoulou. "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database.". *IET Image Processing*, Vol. 7, No. 4, pp. 373–383, June 2013.

[Odst 14] J. Odstrcilik, R. Kolar, R.-P. Tornow, J. Jan, A. Budai, M. Mayer, M. Vodakova, R. Laemmer, M. Lamos, Z. Kuna, J. Gazarek, T. Kubena, P. Cernosek, and M. Ronzhina. "Thickness related textural properties of retinal nerve fiber layer in color fundus images". *Computerized Medical Imaging and Graphics*, Vol. 38, No. 6, pp. 508–516, September 2014.

[Polo 08] V. Polo, J. M. Larrosa, A. Ferreras, F. Mayoral, V. Pueyo, and F. M. Honrubia. "Retinal nerve fiber layer evaluation in open-angle glaucoma. Optimum criteria for optical coherence tomography.". *Ophthalmologica*, Vol. 223, No. 1, pp. 2–6, October 2008.

[Pots 08] B. Potsaid, I. Gorczynska, V. J. Srinivasan, Y. Chen, J. Jiang, A. Cable, and J. G. Fujimoto. "Ultrahigh speed Spectral / Fourierdomain OCT ophthalmic imaging at70,000 to 312,500 axial scans per second". *Optics Express*, Vol. 16, No. 19, pp. 15149–15169, September 2008.

[Qi 10] X. Qi, Y. Pan, M. V. Sivak, J. E. Willis, G. Isenberg, and A. M. Rollins. "Image analysis for classification of dysplasia in Barrett's esophagus using endoscopic optical coherence tomography". *Biomedical Optics Express*, Vol. 1, No. 3, pp. 825–847, October 2010.

[Quel 10] G. Quellec, K. Lee, M. Dolejsi, M. K. Garvin, M. D. Abramoff, and M. Sonka. "Three-Dimensional Analysis of Retinal Layer Texture: Identification of Fluid-Filled Regions in SD-OCT of the Macula". *IEEE Transactions on Medical Imaging*, Vol. 29, No. 6, pp. 1321 –1330, June 2010.

[Quig 06] H. A. Quigley and A. T. Broman. "The number of people with glaucoma worldwide in 2010 and 2020". *British Journal of Ophthalmology*, Vol. 90, pp. 262–267, March 2006.

[Rath 14] F. Rathke, S. Schmidt, and C. Schn´orr. "Probabilistic intra-retinal layer segmentation in 3-D OCT images using global shape regularization". *Medical Image Analysis*, Vol. 18, No. 5, pp. 781–794, July 2014.

[Sadd 07] S. R. Sadda, S. Joeres, Z. Wu, P. Updike, P. Romano, A. T. Collins, and A. C. Walsh. "Error Correction and Quantitative Subanalysis of Optical Coherence Tomography Data Using Computer-Assisted Grading". *Investigative Ophthalmology & Visual Science*, Vol. 48, No. 2, pp. 839–848, February 2007.

[Samp 02] P. A. Sample, M. H. Goldbaum, K. Chan, C. Boden, T.-W. Lee, C. Vasile, A. G. Boehm, T. Sejnowski, C. A. Johnson, and R. N. Weinreb. "Using Machine Learning Classifiers to Identify Glaucomatous Change Earlier in Standard Visual Fields". *Investigative Ophthalmology & Visual Science*, Vol. 43, No. 8, p. 2660, August 2002.

[Schm 99] J. M. Schmitt, S. H. Xiang, and K. M. Yung. "Speckle in Optical Coherence Tomography". *Journal of Biomedical Optics*, Vol. 4, No. 1, pp. 95–105, January 1999.

[Seo 12] J. H. Seo, T.-W. Kim, R. N. Weinreb, K. H. Park, S. H. Kim, and D. M. Kim. "Detection of Localized Retinal Nerve Fiber Layer Defects with Posterior Pole Asymmetry Analysis of Spectral Domain Optical Coherence TomographyPosterior Pole Asymmetry Analysis for Glaucoma". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 8, p. 4347, July 2012.

[Shah 05] M. Shahidi, Z. Wang, and R. Zelkha. "Quantitative Thickness Measurement of Retinal Layers Imaged by Optical Coherence Tomography". *American Journal of Ophthalmology*, Vol. 139, No. 6, pp. 1056 – 1061, June 2005.

[Shar 08] P. Sharma, P. A. Sample, L. M. Zangwill, and J. S. Schuman. "Diagnostic Tools for Glaucoma Detection and Management". *Survey of Ophthalmology*, Vol. 53, No. 6, pp. S17–S32, November 2008.

[Shin 15] J. W. Shin, K. B. Uhm, and M. Seong. "Retinal Nerve Fiber Layer Defect Volume Deviation Analysis Using Spectral-Domain Optical Coherence TomographyRNFL Defect Volume Deviation Analysis". *Investigative Ophthalmology & Visual Science*, Vol. 56, No. 1, p. 21, January 2015.

[Somf 07] G. M. Somfai, H. M. Salinas, C. A. Puliafito, and D. C. Fernández. "Evaluation of potential image acquisition pitfalls during optical coherence tomography and their influence on retinal image segmentation". *Journal of Biomedical Optics*, Vol. 12, No. 4, p. 041209, July 2007.

[Somm 91] A. Sommer, J. Katz, H. A. Quigley, N. R. Miller, A. L. Robin, R. C. Richter, and K. A. Witt. "Clinically detectable nerve fiber atrophy precedes the onset of glaucomatous field loss". *Archives of Ophthalmology*, Vol. 109, No. 1, pp. 77–83, January 1991.

[Srin 14a] P. P. Srinivasan, S. J. Heflin, J. A. Izatt, V. Y. Arshavsky, and S. Farsiu. "Automatic segmentation of up to ten layer boundaries in SD-OCT images of the mouse retina with and without missing layers due to pathology". *Biomedical Optics Express*, Vol. 5, No. 2, pp. 348–365, February 2014.

[Srin 14b] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Farsiu. "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images". *Biomedical Optics Express*, Vol. 5, No. 10, pp. 3568–3577, October 2014.

[Stif 07] D. Stifter. "Beyond biomedicine: a review of alternative applications and developments for optical coherence tomography". *Applied Physics B*, Vol. 88, No. 3, pp. 337–357, August 2007.

[Swin 00] N. V. Swindale, G. Stjepanovic, A. Chin, and F. S. Mikelberg. "Automated analysis of normal and glaucomatous optic nerve head topography images". *Investigative Ophthalmology & Visual Science*, Vol. 41, No. 7, pp. 1730–1742, June 2000.

[Szul 07] M. Szulmowski, M. Wojtkowski, B. Sikorski, T. Bajraszewski, V. J. Srinivasan, A. Szkulmowska, J. J. Kaluzny, J. G. Fujimoto, and A. Kowalczyk. "Analysis of posterior retinal layers in spectral optical coherence tomography images of the normal retina and retinal pathologies". *Journal of Biomedical Optics*, Vol. 12, No. 4, August 2007.

[Taka 12] K. Takayama, M. Hangai, M. Durbin, N. Nakano, S. Morooka, T. Akagi, H. O. Ikeda, and N. Yoshimura. "A Novel Method to Detect Local Ganglion Cell Loss in Early Glaucoma Using Spectral-Domain Optical Coherence TomographyDetection of Local RGC Loss in Glaucoma". *Investigative Ophthalmology & Visual Science*, Vol. 53, No. 11, p. 6904, October 2012.

[Tan 08] O. Tan, G. Li, A. T.-H. Lu, R. Varma, D. Huang, and A. I. for Glaucoma Study Group. "Mapping of Macular Substructures with Optical Coherence Tomography for Glaucoma Diagnosis". *Ophthalmology*, Vol. 115, No. 6, pp. 949–956, June 2008.

[Tan 09] O. Tan, V. Chopra, A. T.-H. Lu, J. S. Schuman, H. Ishikawa, G. Wollstein, R. Varma, and D. Huang. "Detection of Macular Ganglion Cell Loss in Glaucoma by Fourier-Domain Optical Coherence Tomography". *Ophthalmology*, Vol. 116, No. 12, pp. 2305–2314, December 2009.

[Tian 13] J. Tian, P. Marziliano, M. Baskaran, T. A. Tun, and T. Aung. "Automatic segmentation of the choroid in enhanced depth imaging optical coherence tomography images". *Biomedical Optics Express*, Vol. 4, No. 3, pp. 397–411, March 2013.

[Toll 08] D. Tolliver, I. Koutis, H. Ishikawa, J. S. Schuman, and G. L. Miller. "Unassisted Segmentation of Multiple Retinal Layers via Spectral Rounding". In: *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting 2008*, Association for Research in Vision and Ophthalmology, Fort Lauderdale, April 2008. Poster.

[Torn 11] R. P. Tornow, W. A. Schrems, D. Bendschneider, F. K. Horn, M. Mayer, C. Y. Mardin, and R. Laemmer. "Atypical Retardation Patterns in Scanning Laser Polarimetry Are Associated with Low Peripapillary Choroidal Thickness". *Investigative Ophthalmology & Visual Science*, Vol. 52, No. 10, pp. 7523–7528, September 2011.

[Tuul 93] A. Tuulonen, J. Lehtola, and P. J. Airaksinen. "Nerve fiber layer defects with normal visual fields. Do normal optic disc and normal visual field indicate absence of glaucomatous abnormality?". *Ophthalmology*, Vol. 100, No. 5, pp. 587–597, May 1993.

[Verm 10] K. Vermeer, J. van der Schoot, J. de Boer, and H. Lemij. "Automated Retinal and NFL Segmentation in OCT Volume Scans by Pixel Classification". In: *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting 2010*, p. 219, Association for Research in Vision and Ophthalmology, Fort Lauderdale, May 2010.

[Vizz 09] G. Vizzeri, M. Balasubramanian, C. Bowd, R. N. Weinreb, F. A. Medeiros, and L. M. Zangwill. "Spectral domain-optical coherence tomography to detect localized retinal nerve fiber layer defects in glaucomatous eyes". *Optics Express*, Vol. 17, No. 5, pp. 4004–4018, March 2009.

[Voge 96] C. R. Vogel and M. E. Oman. "Iterative Methods for Total Variation Denoising". *SIAM Journal on Scientific Computing*, Vol. 17, No. 1, pp. 227–238, January 1996.

[Wojt 10] M. Wojtkowski. "High-speed optical coherence tomography: basics and applications". *Applied Optics*, Vol. 49, No. 16, pp. D30–D61, June 2010.

[Woll 05] G. Wollstein, J. S. Schuman, L. L. Price, A. Aydin, P. C. Stark, E. Hertzmark, E. Lai, H. Ishikawa, C. Mattox, J. G. Fujimoto, and L. A. Paunescu. "Optical Coherence Tomography Longitudinal Evaluation of Retinal Nerve Fiber Layer Thickness in Glaucoma". *Archives of Ophthalmology*, Vol. 123, No. 4, pp. 464–470, April 2005.

[Wrob 09] D. Wroblewski, B. Francis, V. Chopra, A. Kawji, P. Quiros, L. Dustin, and R. Massengill. "Glaucoma detection and evaluation through pattern recognition in standard automated perimetry data". *Graefe's Archive for Clinical and Experimental Ophthalmology*, Vol. 247, No. 11, pp. 1517–1530, November 2009.

[Yang 10] Q. Yang, C. A. Reisman, Z. Wang, Y. Fukuma, M. Hangai, N. Yoshimura, A. Tomidokoro, M. Araie, A. S. Raza, D. C. Hood, and K. Chan. "Automated layer segmentation of macular OCT images using dual-scale gradient information". *Opt. Express*, Vol. 18, No. 20, pp. 21293–21307, September 2010.

[Yang 11] Q. Yang, C. A. Reisman, K. Chan, R. Ramachandran, A. Raza, and D. C. Hood. "Automated segmentation of outer retinal layers in macular OCT images of patients with retinitis pigmentosa". *Biomedical Optics Express*, Vol. 2, No. 9, pp. 2493–2503, September 2011.

[Yazd 09] A. Yazdanpanah, G. Hamarneh, B. Smith, and M. Sarunic. "Intra-retinal Layer Segmentation in Optical Coherence Tomography Using an Active Contour Approach". In: *Proceedings of the 12th international conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 649–656, Springer-Verlag, Berlin, Heidelberg, September 2009.

[Yazd 11] A. Yazdanpanah, G. Hamarneh, B. R. Smith, and M. V. Sarunic. "Segmentation of Intra-Retinal Layers From Optical Coherence Tomography Images Using an Active Contour Approach". *IEEE Transactions on Medical Imaging*, Vol. 30, No. 2, pp. 484 –496, February 2011.

[Yiu 14] K. C. Yiu. "Neural Network Analysis for the detection of glaucomatous damage". *Applied Soft Computing*, Vol. 20, No. , pp. 66–69, July 2014.

[Zang 04] L. M. Zangwill, K. Chan, C. Bowd, J. Hao, T.-W. Lee, R. N. Weinreb, T. J. Sejnowski, and M. H. Goldbaum. "Heidelberg Retina Tomograph Measurements of the Optic Disc and Parapapillary Retina for Detecting Glaucoma Analyzed by Machine Learning Classifiers". *Investigative Ophthalmology & Visual Science*, Vol. 45, No. 9, pp. 3144–3151, September 2004.

[Zhan 13] X. Zhang, A. S. Raza, and D. C. Hood. "Detecting Glaucoma With Visual Fields Derived From Frequency-Domain Optical Coherence Tomography". *Investigative Ophthalmology & Visual Science*, Vol. 54, No. 5, pp. 3289–3296, May 2013.

[Zhu 14] H. Zhu, A. Poostchi, S. A. Vernon, and D. P. Crabb. "Detecting abnormality in optic nerve head images using a feature extraction analysis". *Biomedical Optics Express*, Vol. 5, No. 7, pp. 2215–2230, July 2014.